15915 U.S.PTO
042909

U.S. PTO
12/453124
04/29/2009

PATENTS

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Ron M. Redlich

Serial No.:

Filed: Herewith

For: Digital Information Infrastructure and Method

TRANSMITTAL AND CERTIFICATE OF EXPRESS MAILING

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

The below listed documents are being hand delivered to the United States Patent and Trademark Office on April 29, 2009.

X Patent application (1003 pgs. Spec.; 2 pgs. Claims; 1 pg. Abstract.)

X Fifty-Nine (59) sheets drawings - formal

X Unsigned Declaration and Power of Attorney for Patent Application (2 pgs.)

X Return receipt postcard.

Respectfully submitted,

By _____
Robert C. Kain, Jr.
Reg. No. 30,648
Kain & Associates, Attorneys at Law, P.A.
750 S.E. Third Avenue, Suite 100
Ft. Lauderdale, Florida 33316-1153
Telephone: (954) 768-9002
Facsimile: (954) 768-0158

G:\RCK\CLIENTS\Redlich\6851-08-cip-EXPMAIL-PAT.wpd

Best Available Copy

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled DIGITAL INFORMATION INFRASTRUCTURE AND METHOD
the specification of which

(check one) ☒ is attached hereto.

☐ was filed on _____

as Application Serial No. _____

and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 (a-d) or §365(b) of any foreign application(s) for patent or inventor's certificate, or §365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below any foreign application for patent or inventor's certificate, or any PCT international application having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)			Priority Claimed	
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	☐ Yes	☐ No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	☐ Yes	☐ No

I hereby claim the benefit under Title 35, United States Code, §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

SN 11/746,440 (Application Serial No.)	May 9, 2007 (Filing Date)	pending (Status) (patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status) (patented, pending, abandoned)

I hereby claim the benefit under Title 35, United States Code, §119(e) of any United States provisional application(s) listed below:

60/883,522	January 5, 2007
(Provisional Application No.)	(Filing Date)
(Provisional Application No.)	(Filing Date)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. 1001, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.
POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agents to prosecute this application and transact all business in the Patent and Trademark Office connected therewith.
Robert C. Kain, Jr., Reg. No. 30,648.

Send Correspondence to: Robert C. Kain, Jr., **Customer Number 004897**
Kain & Associates
750 Southeast Third Avenue, Suite 100
Ft. Lauderdale, Florida 33316-1153
Direct Telephone Calls to: Robert C. Kain, Jr., (954) 768-9002

Ron M. Redlich
Full name of sole or first inventor

Inventor's signature Date

Miami Beach, FL
Residence

U. S.
Citizenship

4201 Collins Avenue, Apt. 2103
Post Office Address

Miami Beach, FL 33140

Martin A. Nemzow
Full name of second joint inventor, if any

Inventor's signature Date

Miami Beach, FL
Residence

U.S.
Citizenship

2915 Flamingo Drive
Post Office Address

Miami Beach, FL 33140

(Supply similar information and signature for third and subsequent joint inventors.)

G:\RCK\CLIENTS\Redlich\6851-08-cip-pa-dec3.wpd

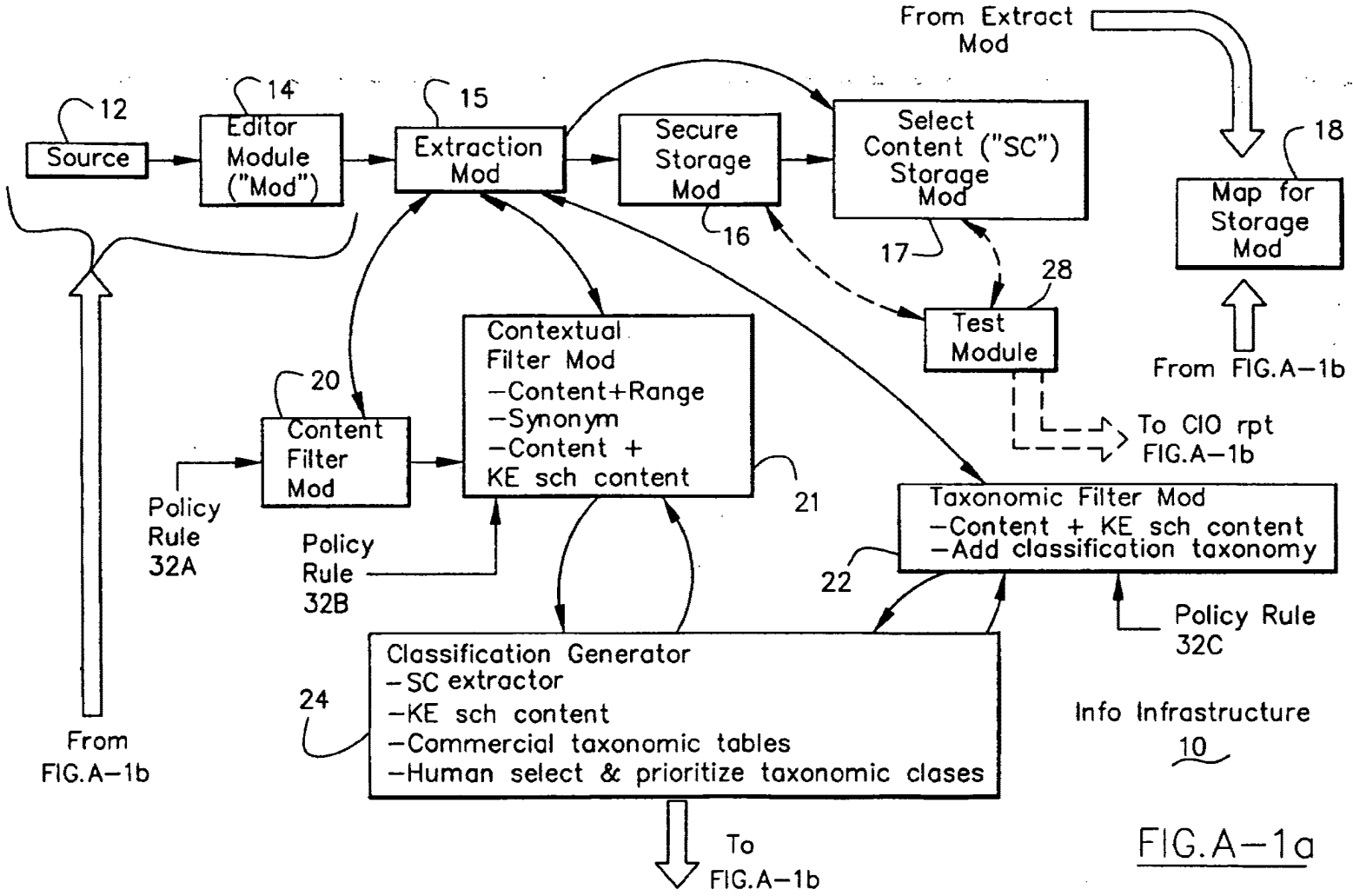

From Extract Mod
12
Source
14
Editor Module ("Mod")
15
Extraction Mod
Secure Storage Mod
16
Select Content ("SC") Storage Mod
17
18
Map for Storage Mod
28
Test Module
From FIG.A–1b
20
Content Filter Mod
Contextual Filter Mod
–Content+Range
–Synonym
–Content + KE sch content
21
To CIO rpt FIG.A–1b
Policy Rule 32A
Policy Rule 32B
Taxonomic Filter Mod
–Content + KE sch content
–Add classification taxonomy
22
Policy Rule 32C
Classification Generator
–SC extractor
–KE sch content
–Commercial taxonomic tables
–Human select & prioritize taxonomic clases
24
From FIG.A–1b
To FIG.A–1b
Info Infrastructure
10

FIG.A–1a

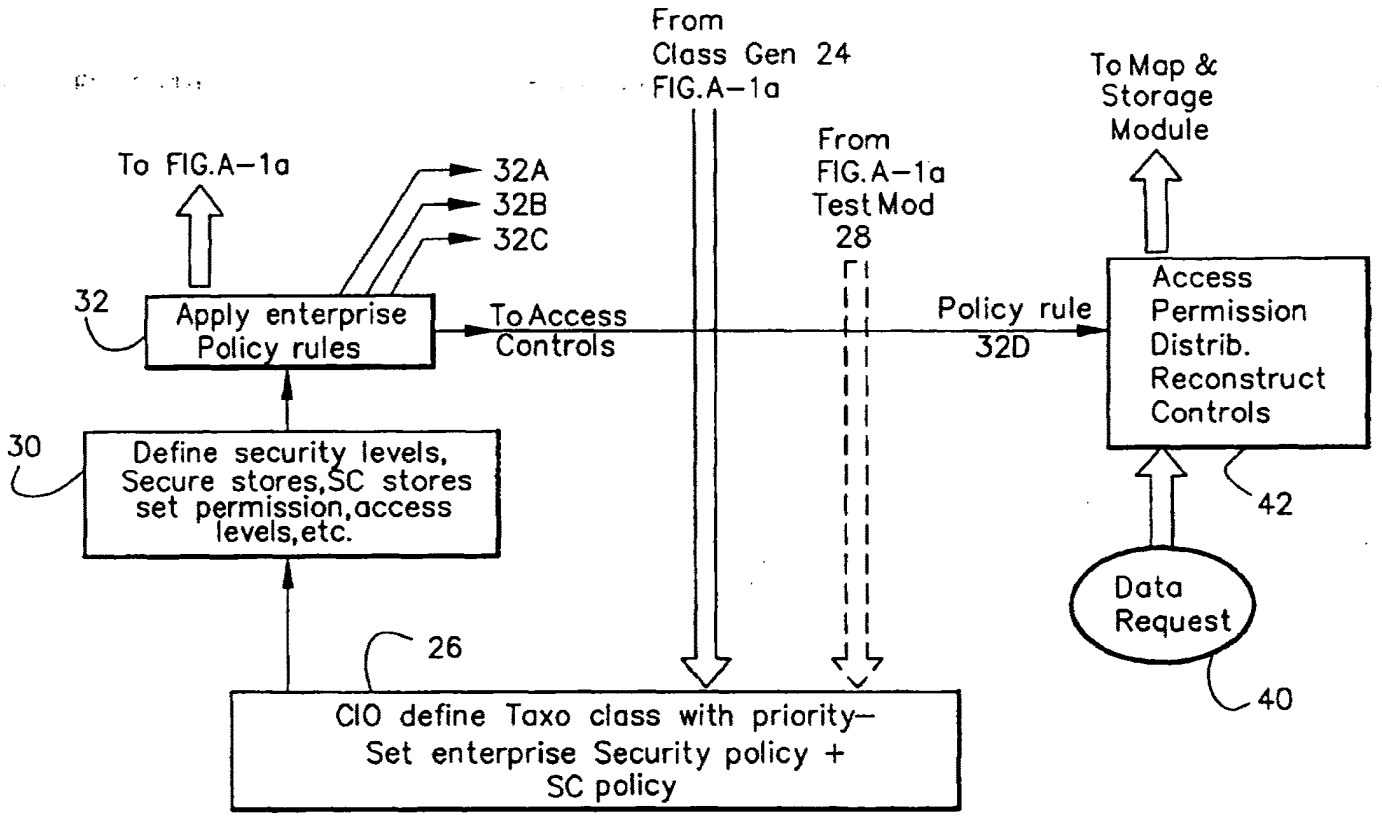

From
Class Gen 24
FIG.A–1a
To Map &
Storage
Module
From
FIG.A–1a
Test Mod
28
To FIG.A–1a
32A
32B
32C
32
Apply enterprise
Policy rules
To Access
Controls
Policy rule
32D
Access
Permission
Distrib.
Reconstruct
Controls
42
30
Define security levels,
Secure stores,SC stores
set permission,access
levels,etc.
Data
Request
40
26
CIO define Taxo class with priority–
Set enterprise Security policy +
SC policy

FIG.A–1b

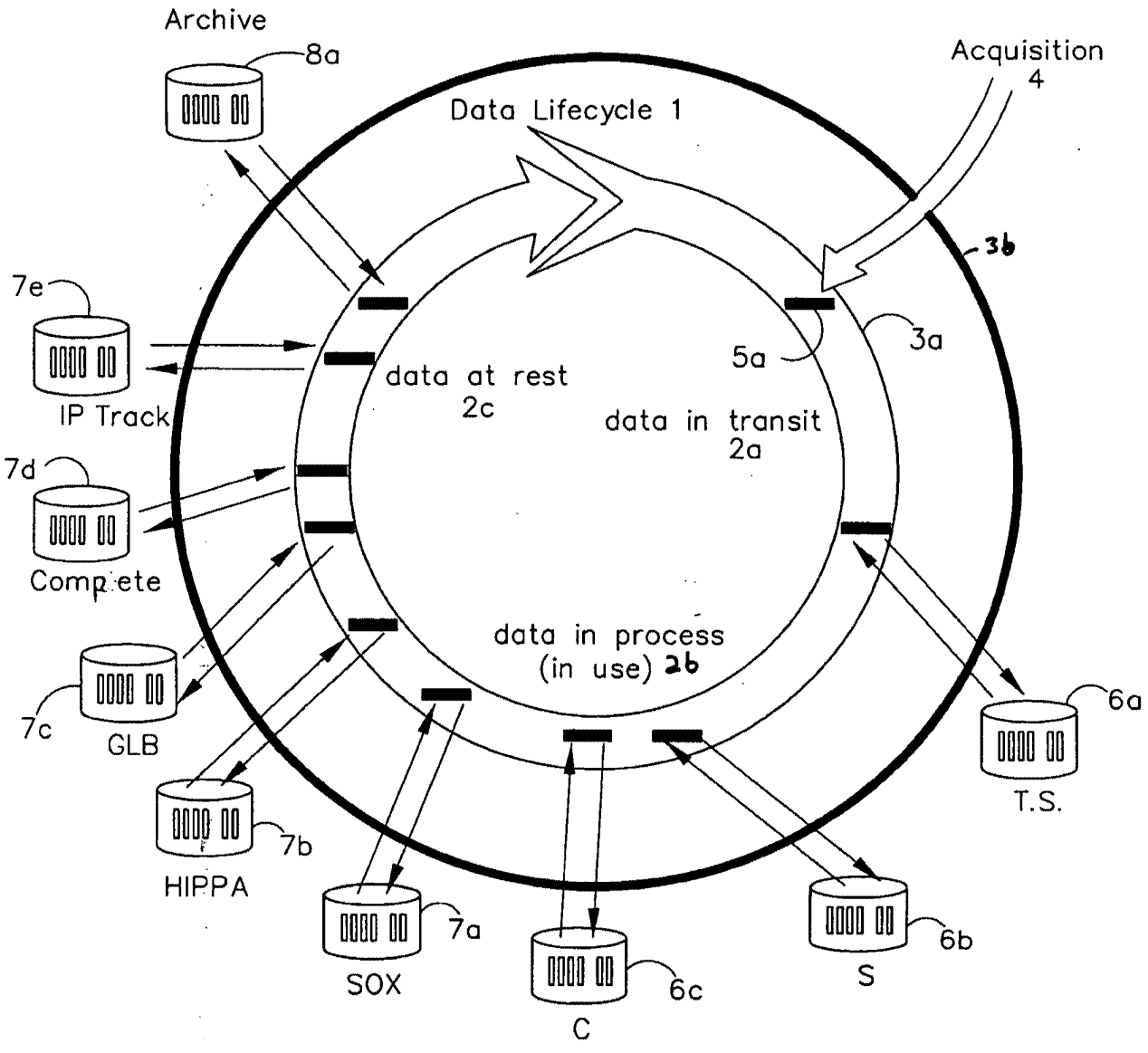

Archive
8a
Acquisition
4
Data Lifecycle 1
3b
7e
IP Track
data at rest
2c
5a
3a
data in transit
2a
7d
Comp ete
data in process
(in use) 2b
7c
GLB
6a
T.S.
7b
HIPPA
7a
SOX
6c
C
6b
S

FIG. A–2

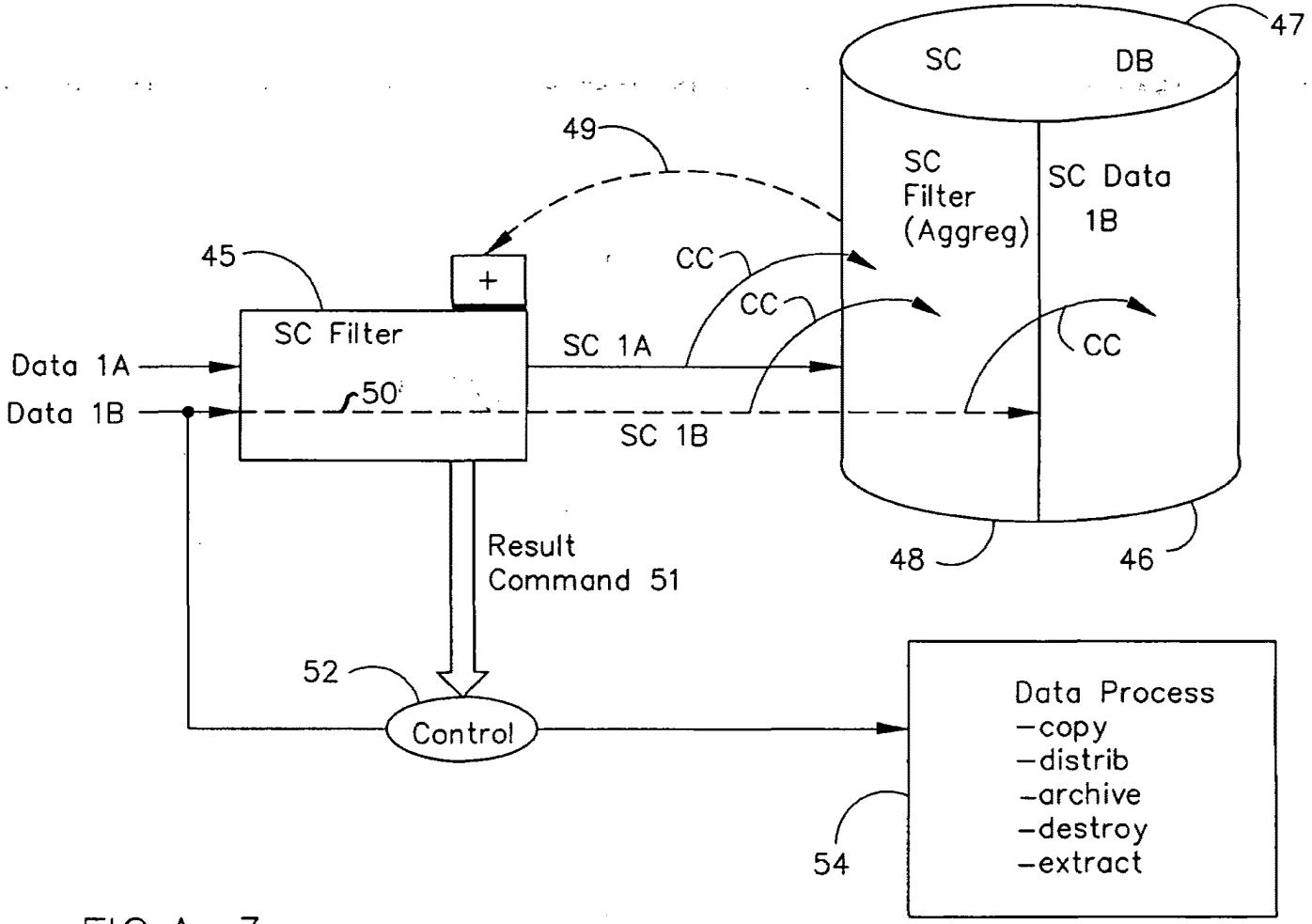

47
SC
DB
49
SC
Filter
(Aggreg)
SC Data
1B
45
+
CC
CC
CC
SC Filter
Data 1A
Data 1B
50
SC 1A
SC 1B
Result
Command 51
48
46
52
Control
Data Process
-copy
-distrib
-archive
-destroy
-extract
54

FIG. A-3

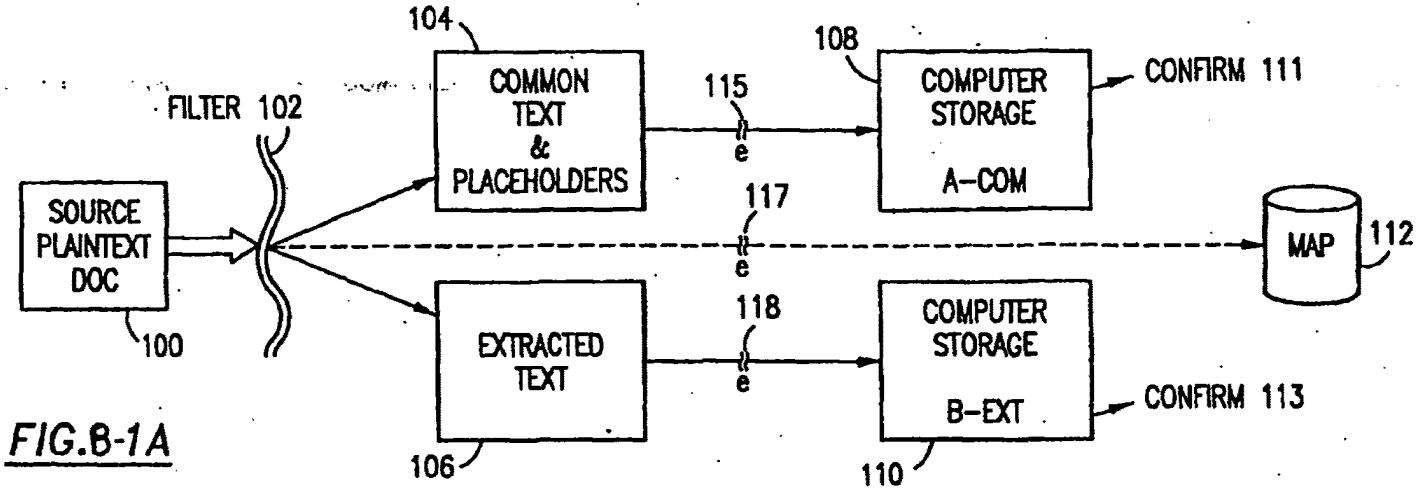


FIG.B-1A

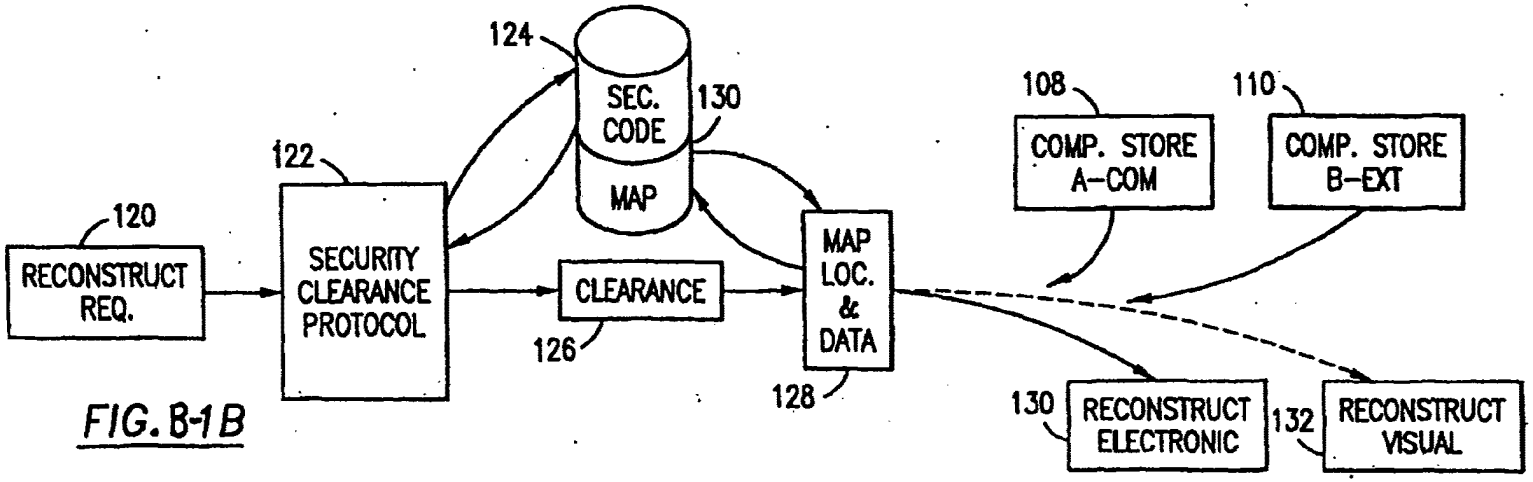


FIG. B-1B

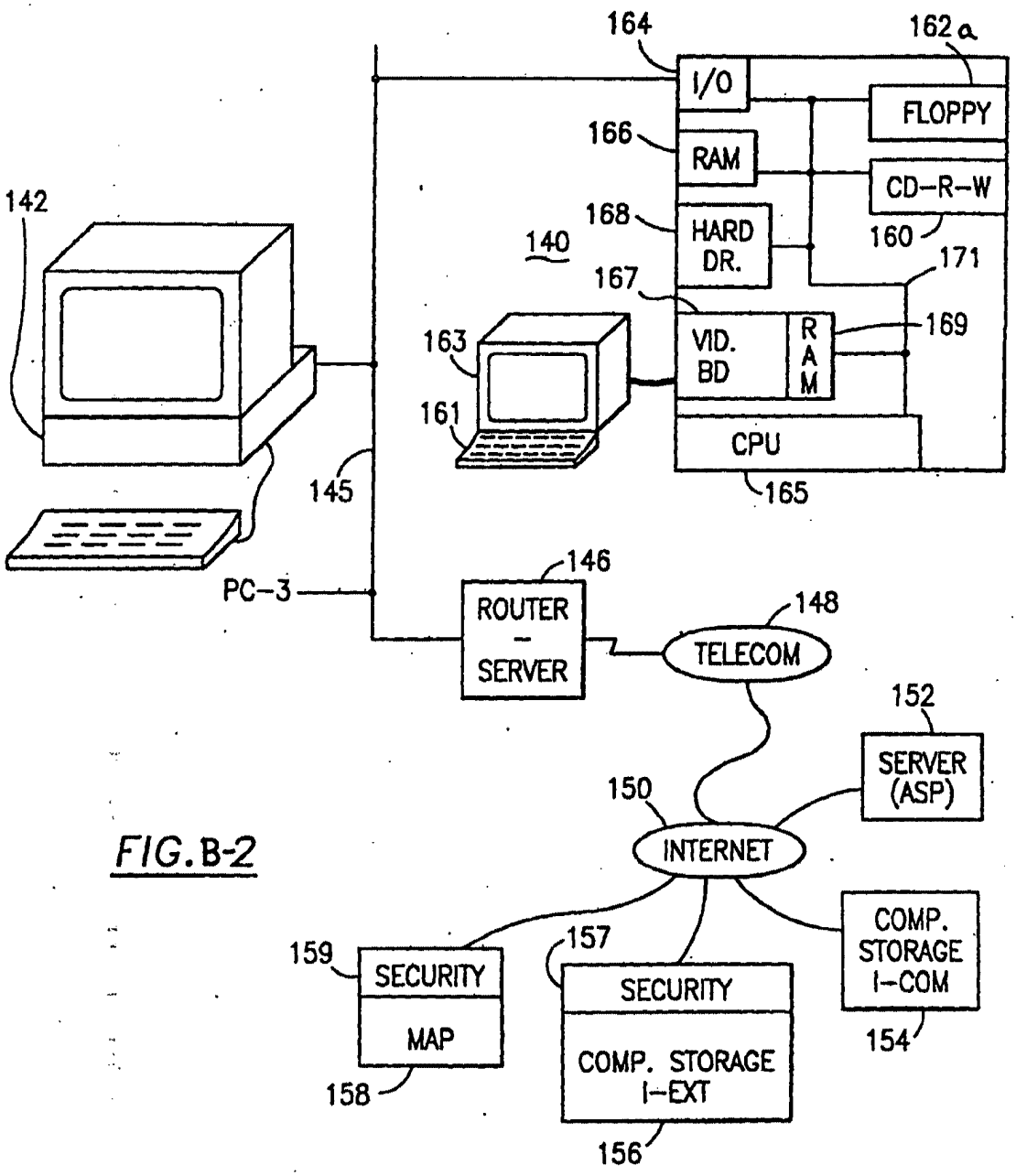

164
162a
I/O
FLOPPY
166
RAM
CD-R-W
142
168
HARD DR.
160
140
171
167
VID. BD
R A M
169
163
161
CPU
145
165
146
PC-3
ROUTER – SERVER
148
TELECOM
152
SERVER (ASP)
150
INTERNET
159
SECURITY
MAP
158
157
SECURITY
COMP. STORAGE I-EXT
156
COMP. STORAGE I-COM
154

FIG. B-2

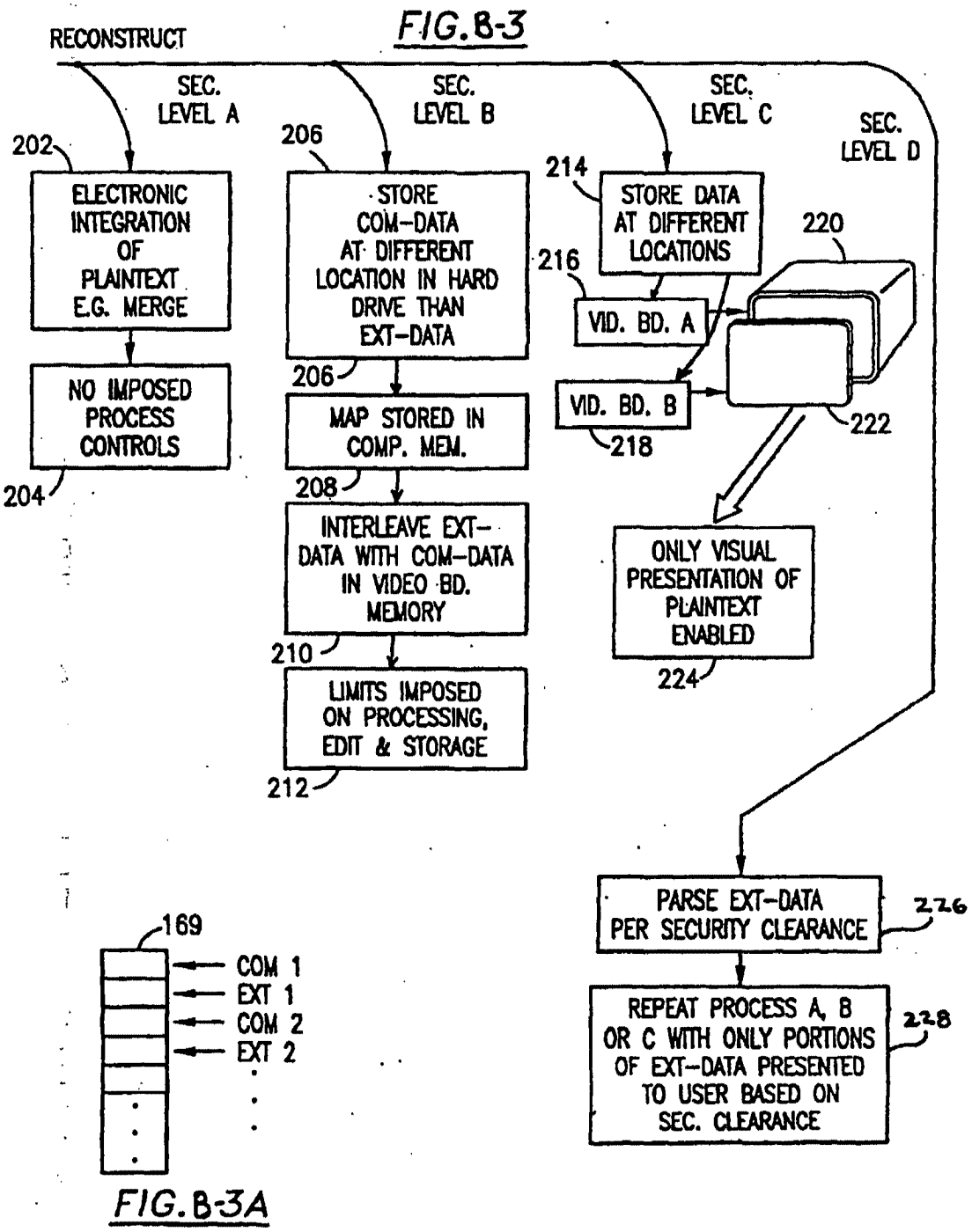
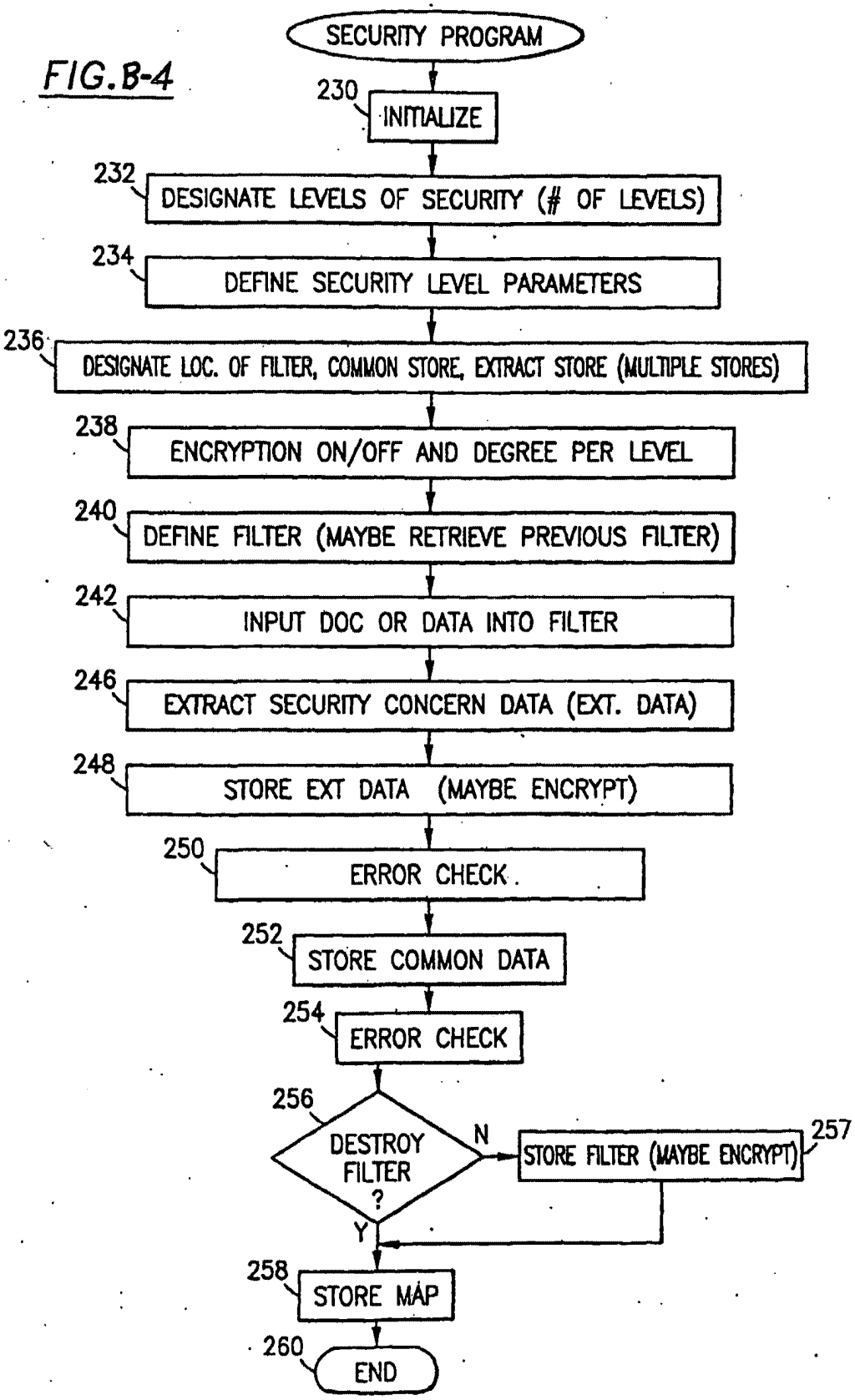


FIG. B-3

FIG. B-3A

FIG. B-4


SECURITY PROGRAM
230
INITIALIZE
232
DESIGNATE LEVELS OF SECURITY (# OF LEVELS)
234
DEFINE SECURITY LEVEL PARAMETERS
236
DESIGNATE LOC. OF FILTER, COMMON STORE, EXTRACT STORE (MULTIPLE STORES)
238
ENCRYPTION ON/OFF AND DEGREE PER LEVEL
240
DEFINE FILTER (MAYBE RETRIEVE PREVIOUS FILTER)
242
INPUT DOC OR DATA INTO FILTER
246
EXTRACT SECURITY CONCERN DATA (EXT. DATA)
248
STORE EXT DATA (MAYBE ENCRYPT)
250
ERROR CHECK.
252
STORE COMMON DATA
254
ERROR CHECK
256
DESTROY FILTER ?
N
257
STORE FILTER (MAYBE ENCRYPT)
Y
258
STORE MAP
260
END

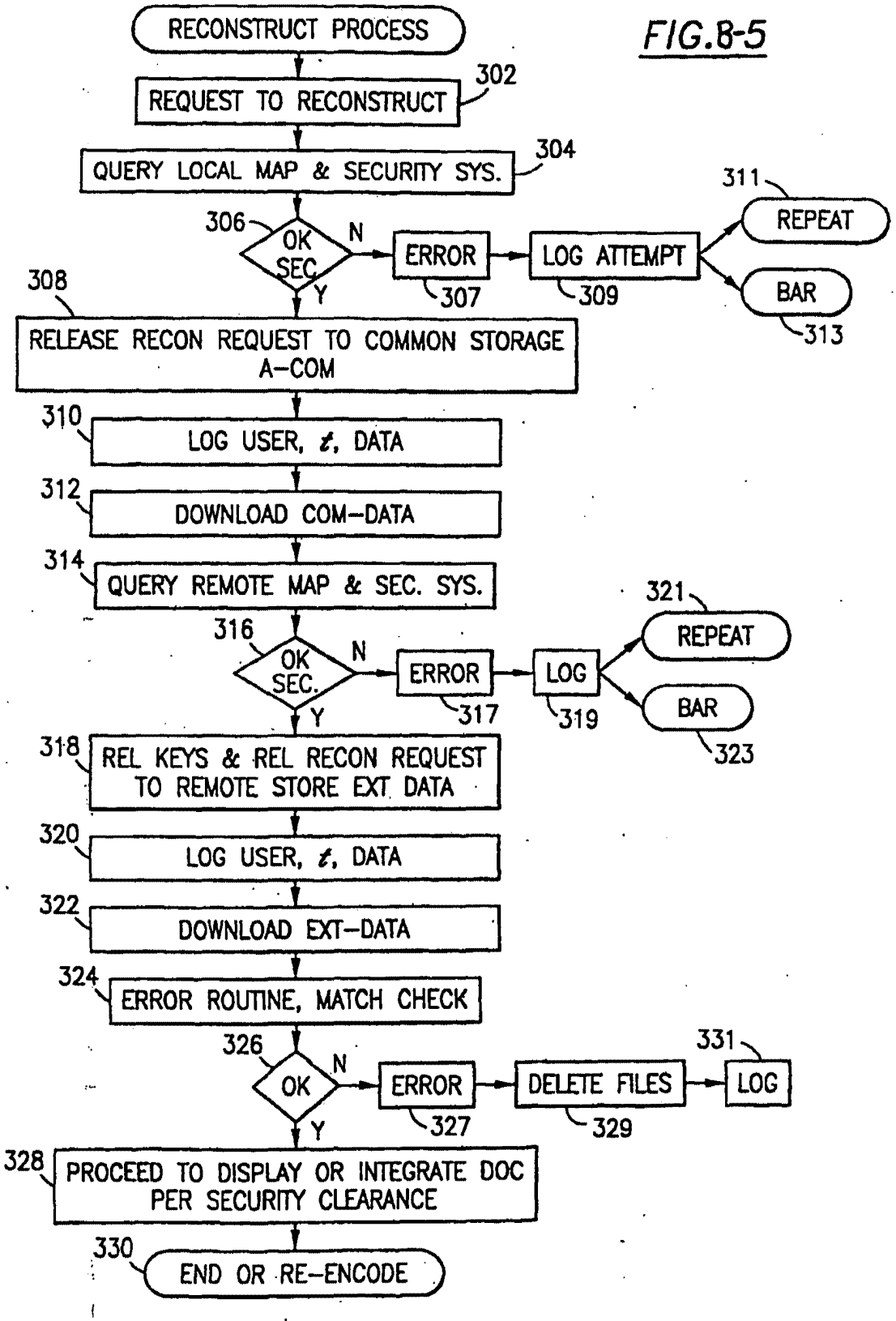

RECONSTRUCT PROCESS
REQUEST TO RECONSTRUCT
302
QUERY LOCAL MAP & SECURITY SYS.
304
306
OK SEC.
N
Y
ERROR
307
LOG ATTEMPT
309
311
REPEAT
BAR
313
308
RELEASE RECON REQUEST TO COMMON STORAGE A-COM
310
LOG USER, t, DATA
312
DOWNLOAD COM-DATA
314
QUERY REMOTE MAP & SEC. SYS.
316
OK SEC.
N
Y
ERROR
317
LOG
319
321
REPEAT
BAR
323
318
REL KEYS & REL RECON REQUEST TO REMOTE STORE EXT DATA
320
LOG USER, t, DATA
322
DOWNLOAD EXT-DATA
324
ERROR ROUTINE, MATCH CHECK
326
OK
N
Y
ERROR
327
DELETE FILES
329
331
LOG
328
PROCEED TO DISPLAY OR INTEGRATE DOC PER SECURITY CLEARANCE
330
END OR RE-ENCODE

FIG.8-5

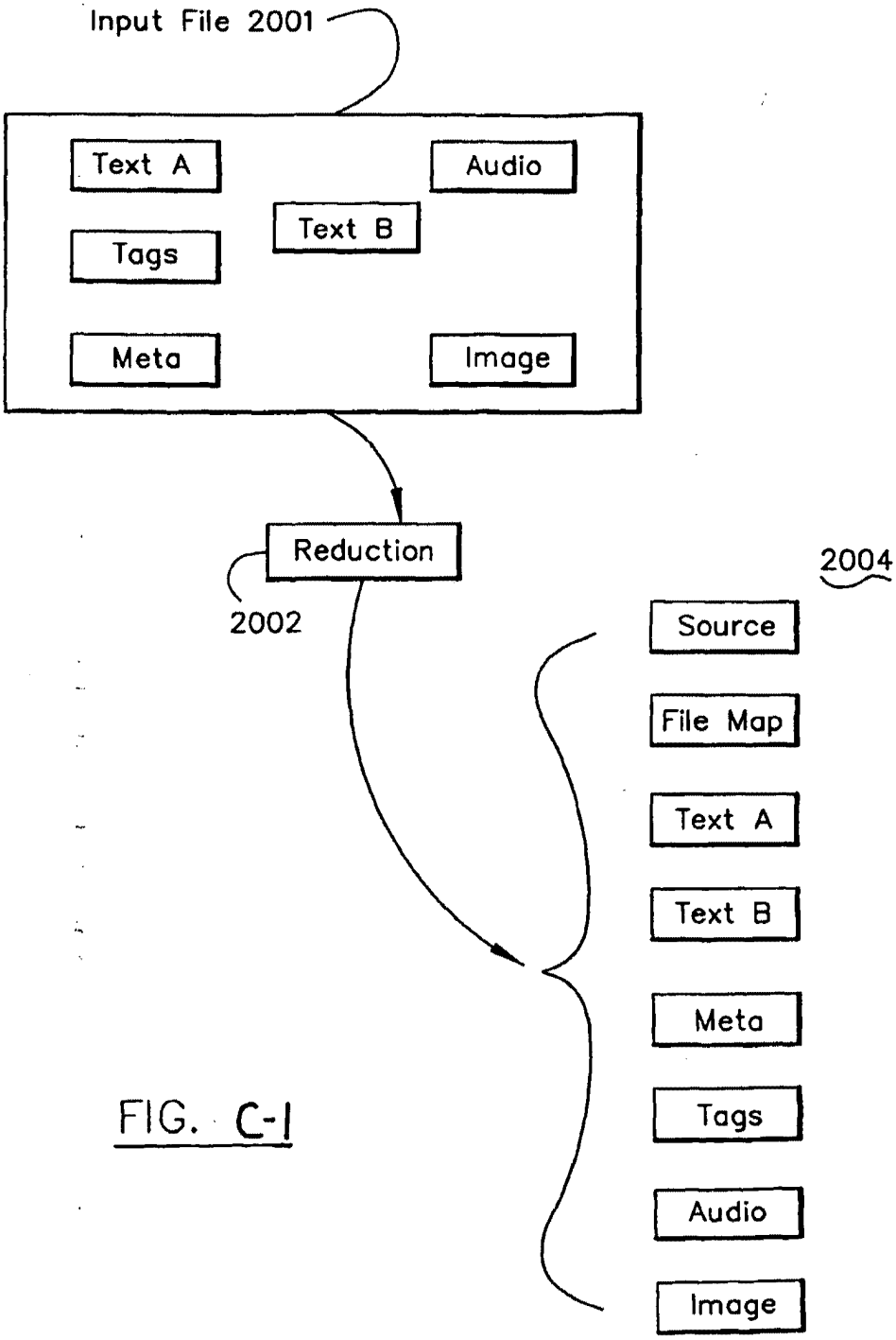

Input File 2001
Text A
Audio
Text B
Tags
Meta
Image
Reduction
2002
2004
Source
File Map
Text A
Text B
Meta
Tags
Audio
Image

FIG. C-1

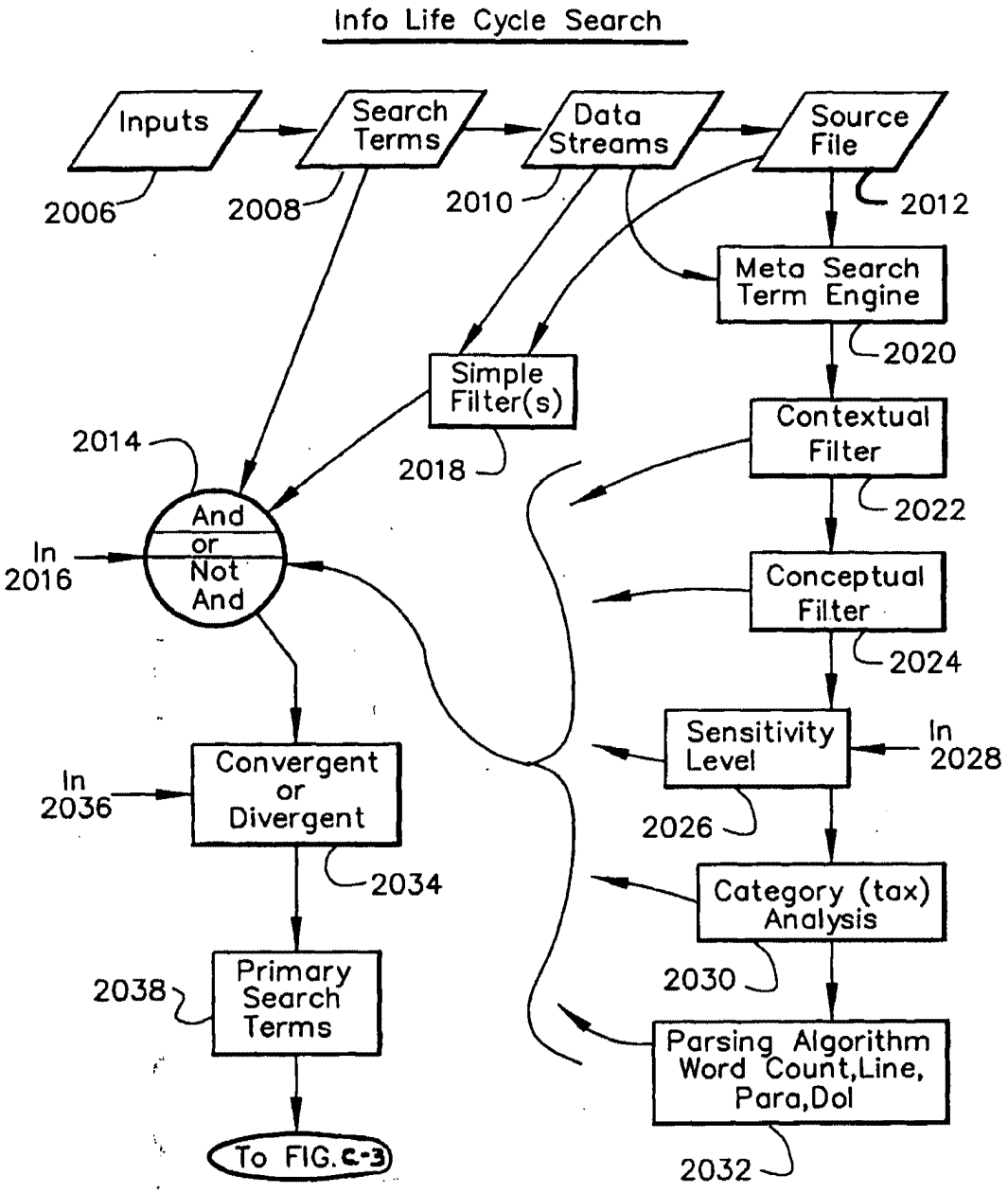

Info Life Cycle Search
Inputs
Search Terms
Data Streams
Source File
2006
2008
2010
2012
Meta Search Term Engine
2020
Simple Filter(s)
2018
Contextual Filter
2022
2014
And
or
Not
And
In
2016
Conceptual Filter
2024
Sensitivity Level
In
2028
2026
In
2036
Convergent or Divergent
2034
Category (tax) Analysis
2030
2038
Primary Search Terms
Parsing Algorithm Word Count,Line, Para,Dol
2032
To FIG. C-3

FIG. C-2

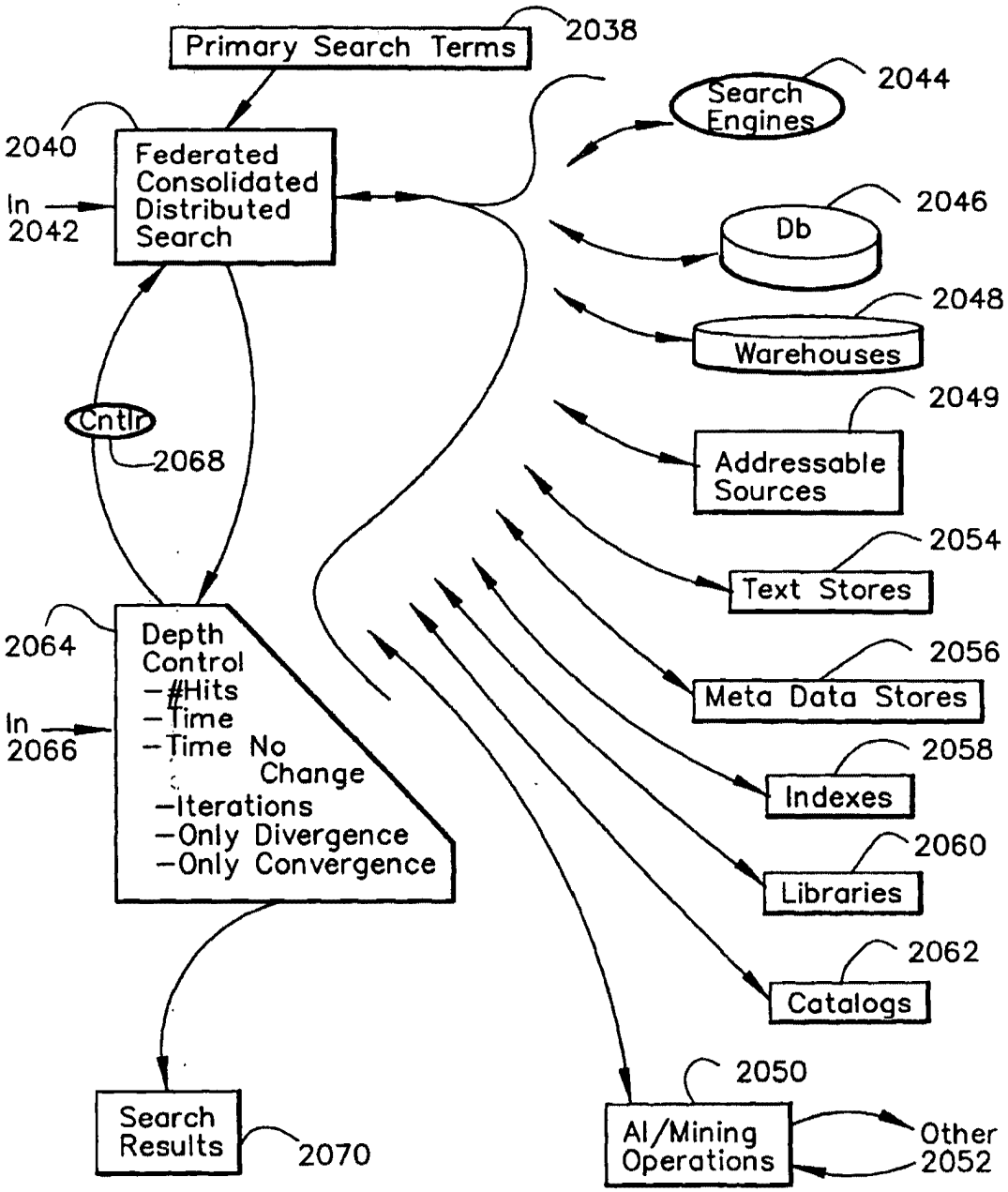

Primary Search Terms
2038
2040
Federated Consolidated Distributed Search
In 2042
Search Engines
2044
Db
2046
Warehouses
2048
Addressable Sources
2049
Text Stores
2054
Meta Data Stores
2056
Indexes
2058
Libraries
2060
Catalogs
2062
Cntlr
2068
2064
Depth Control
-#Hits
-Time
-Time No Change
-Iterations
-Only Divergence
-Only Convergence
In 2066
Search Results
2070
2050
AI/Mining Operations
Other 2052

FIG. C-3

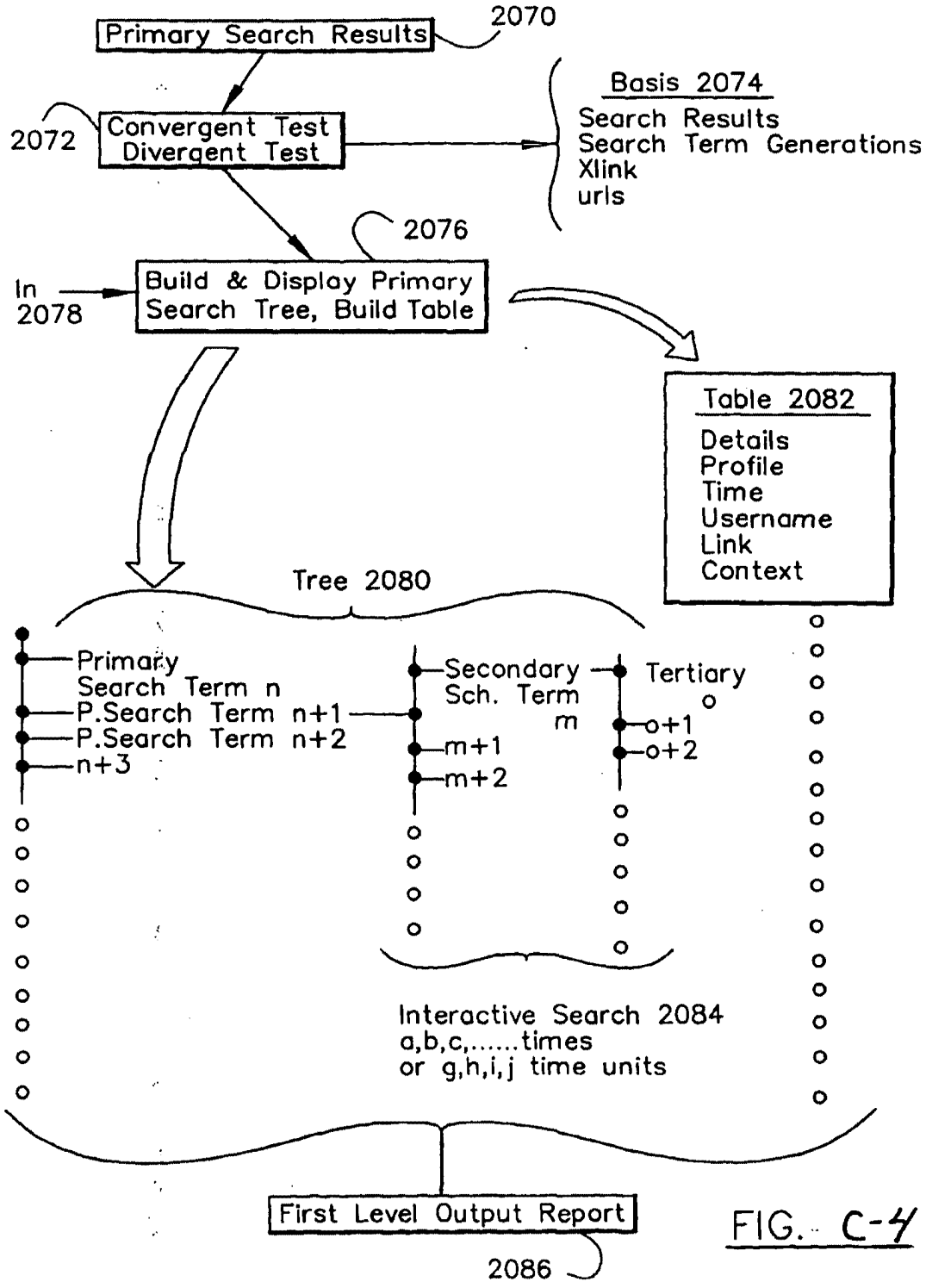

Primary Search Results
2070
Basis 2074
Search Results
Search Term Generations
Xlink
urls
2072
Convergent Test
Divergent Test
2076
In 2078
Build & Display Primary Search Tree, Build Table
Table 2082
Details
Profile
Time
Username
Link
Context
Tree 2080
Primary Search Term n
P.Search Term n+1
P.Search Term n+2
n+3
Secondary Sch. Term m
m+1
m+2
Tertiary o
o+1
o+2
Interactive Search 2084
a,b,c,.....times
or g,h,i,j time units
First Level Output Report
2086

FIG. C-4

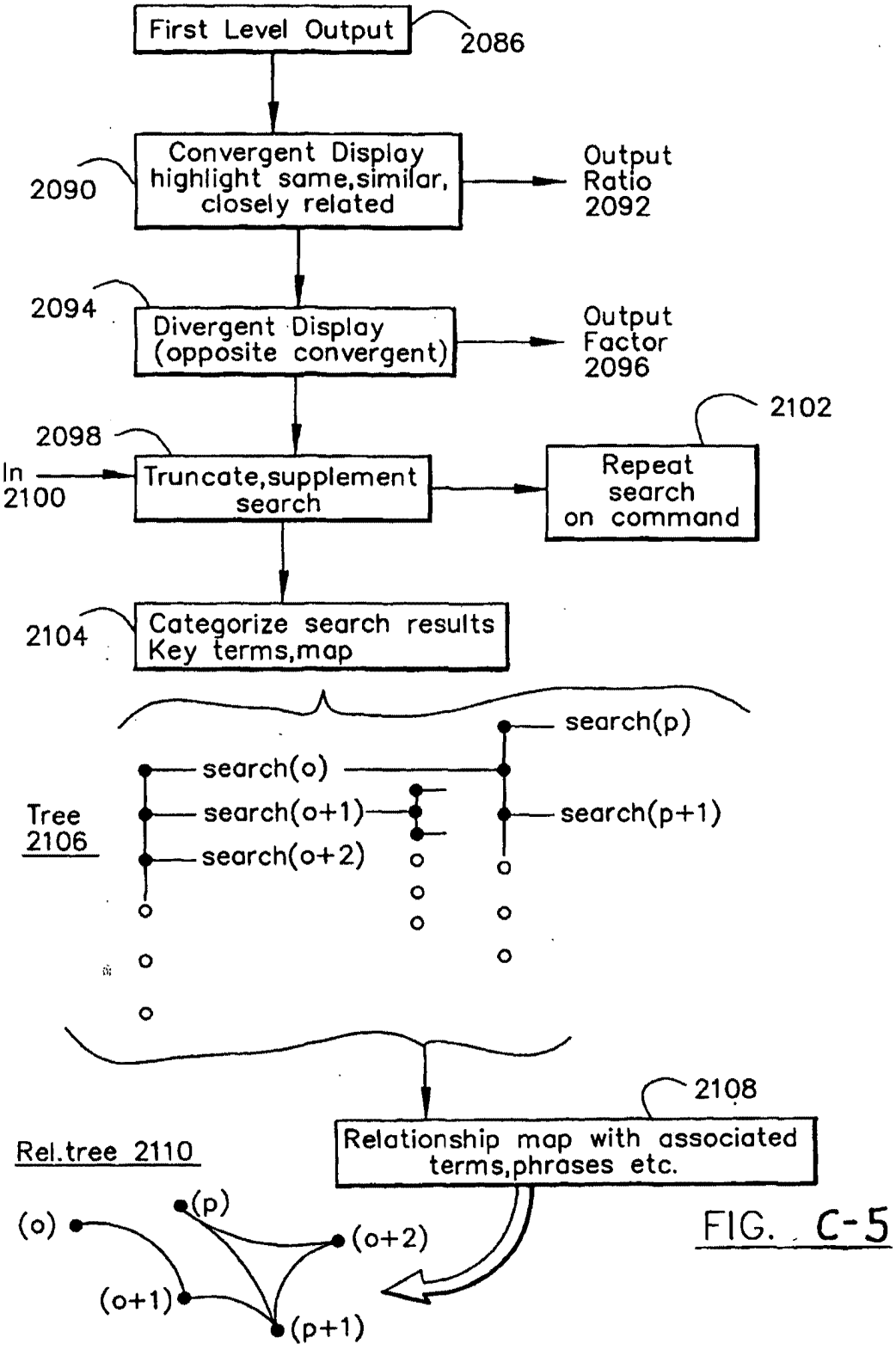

First Level Output
2086
Convergent Display highlight same,similar, closely related
2090
Output Ratio 2092
2094
Divergent Display (opposite convergent)
Output Factor 2096
2098
In 2100
Truncate,supplement search
2102
Repeat search on command
2104
Categorize search results Key terms,map
Tree 2106
search(o)
search(o+1)
search(o+2)
search(p)
search(p+1)
2108
Relationship map with associated terms,phrases etc.
Rel.tree 2110
(o)
(p)
(o+2)
(o+1)
(p+1)

FIG. C-5

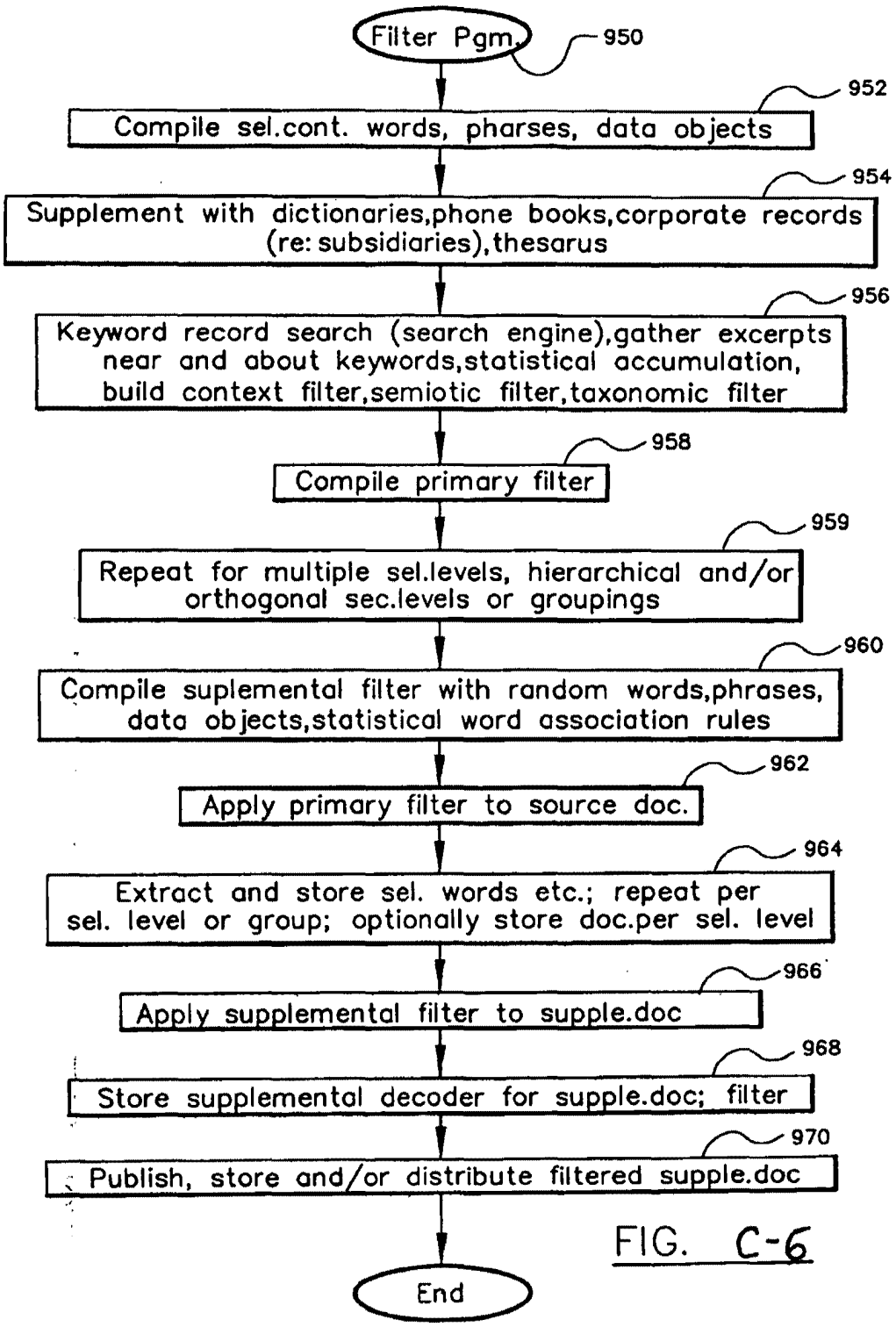

Filter Pgm.
950
952
Compile sel.cont. words, pharses, data objects
954
Supplement with dictionaries,phone books,corporate records (re: subsidiaries),thesarus
956
Keyword record search (search engine),gather excerpts near and about keywords,statistical accumulation, build context filter,semiotic filter,taxonomic filter
958
Compile primary filter
959
Repeat for multiple sel.levels, hierarchical and/or orthogonal sec.levels or groupings
960
Compile suplemental filter with random words,phrases, data objects,statistical word association rules
962
Apply primary filter to source doc.
964
Extract and store sel. words etc.; repeat per sel. level or group; optionally store doc.per sel. level
966
Apply supplemental filter to supple.doc
968
Store supplemental decoder for supple.doc; filter
970
Publish, store and/or distribute filtered supple.doc
End

FIG. C-6

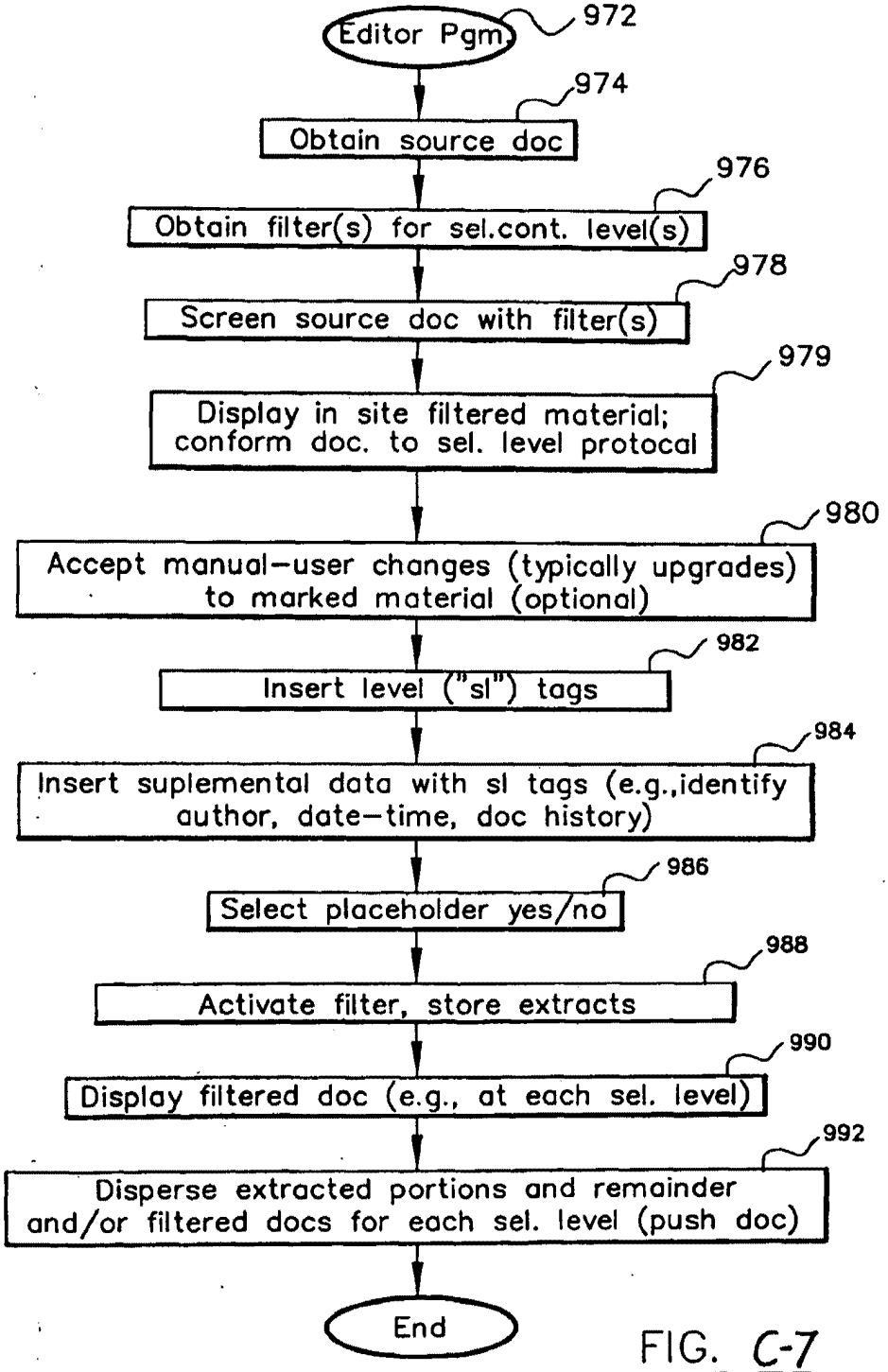

Editor Pgm
972
Obtain source doc
974
Obtain filter(s) for sel.cont. level(s)
976
Screen source doc with filter(s)
978
Display in site filtered material; conform doc. to sel. level protocal
979
Accept manual-user changes (typically upgrades) to marked material (optional)
980
Insert level ("sl") tags
982
Insert suplemental data with sl tags (e.g.,identify author, date-time, doc history)
984
Select placeholder yes/no
986
Activate filter, store extracts
988
Display filtered doc (e.g., at each sel. level)
990
Disperse extracted portions and remainder and/or filtered docs for each sel. level (push doc)
992
End

FIG. C-7

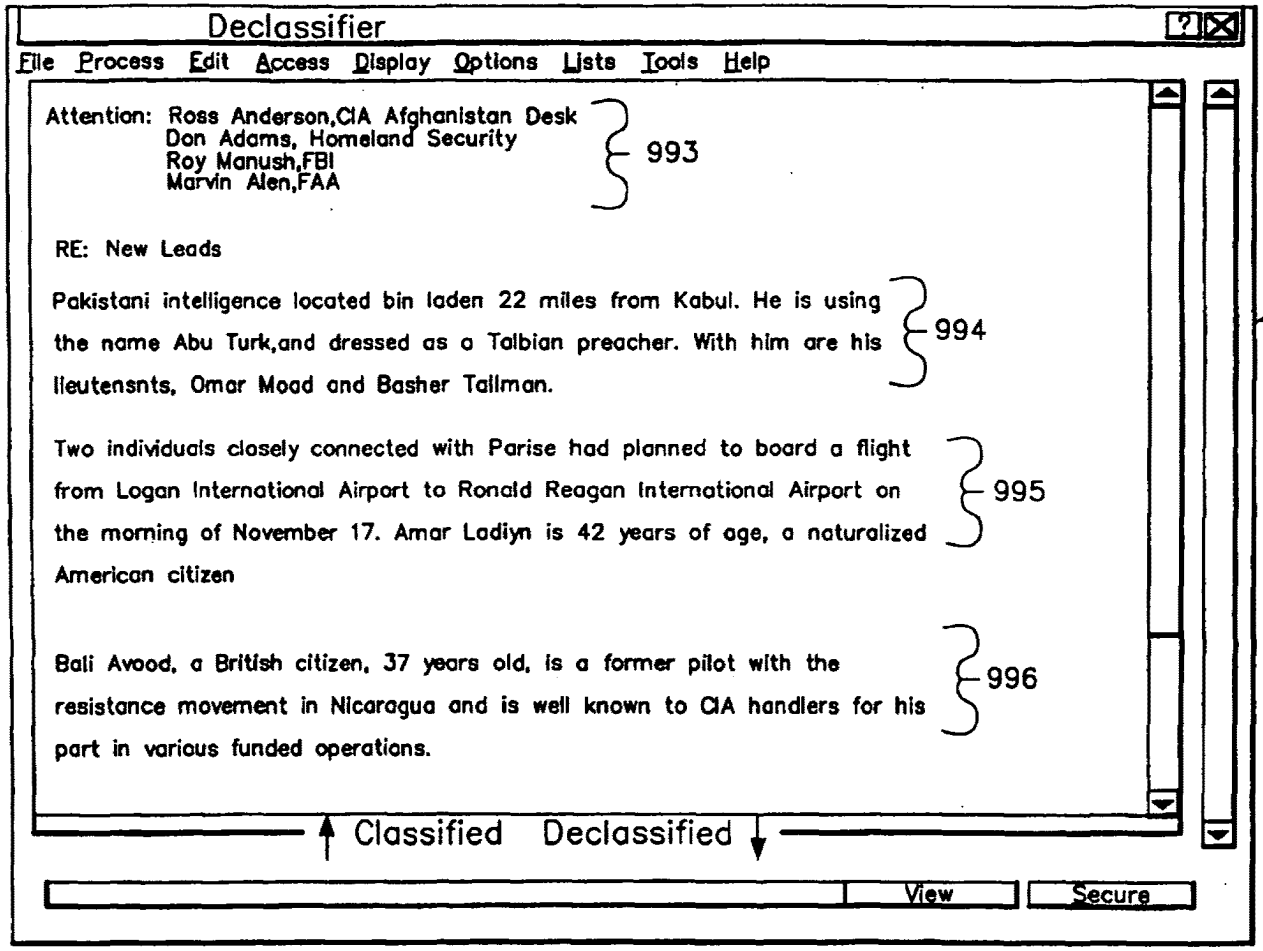

Declassifier
File Process Edit Access Display Options Lists Tools Help
Attention: Ross Anderson,CIA Afghanistan Desk
Don Adams, Homeland Security
Roy Manush,FBI
Marvin Alen,FAA
993
RE: New Leads
Pakistani intelligence located bin laden 22 miles from Kabul. He is using the name Abu Turk,and dressed as a Taibian preacher. With him are his lieutensnts, Omar Moad and Basher Tailman.
994
Two individuals closely connected with Parise had planned to board a flight from Logan International Airport to Ronald Reagan International Airport on the morning of November 17. Amar Ladiyn is 42 years of age, a naturalized American citizen
995
Bali Avood, a British citizen, 37 years old, is a former pilot with the resistance movement in Nicaragua and is well known to CIA handlers for his part in various funded operations.
996
Classified Declassified
View
Secure
991

FIG. C-8A

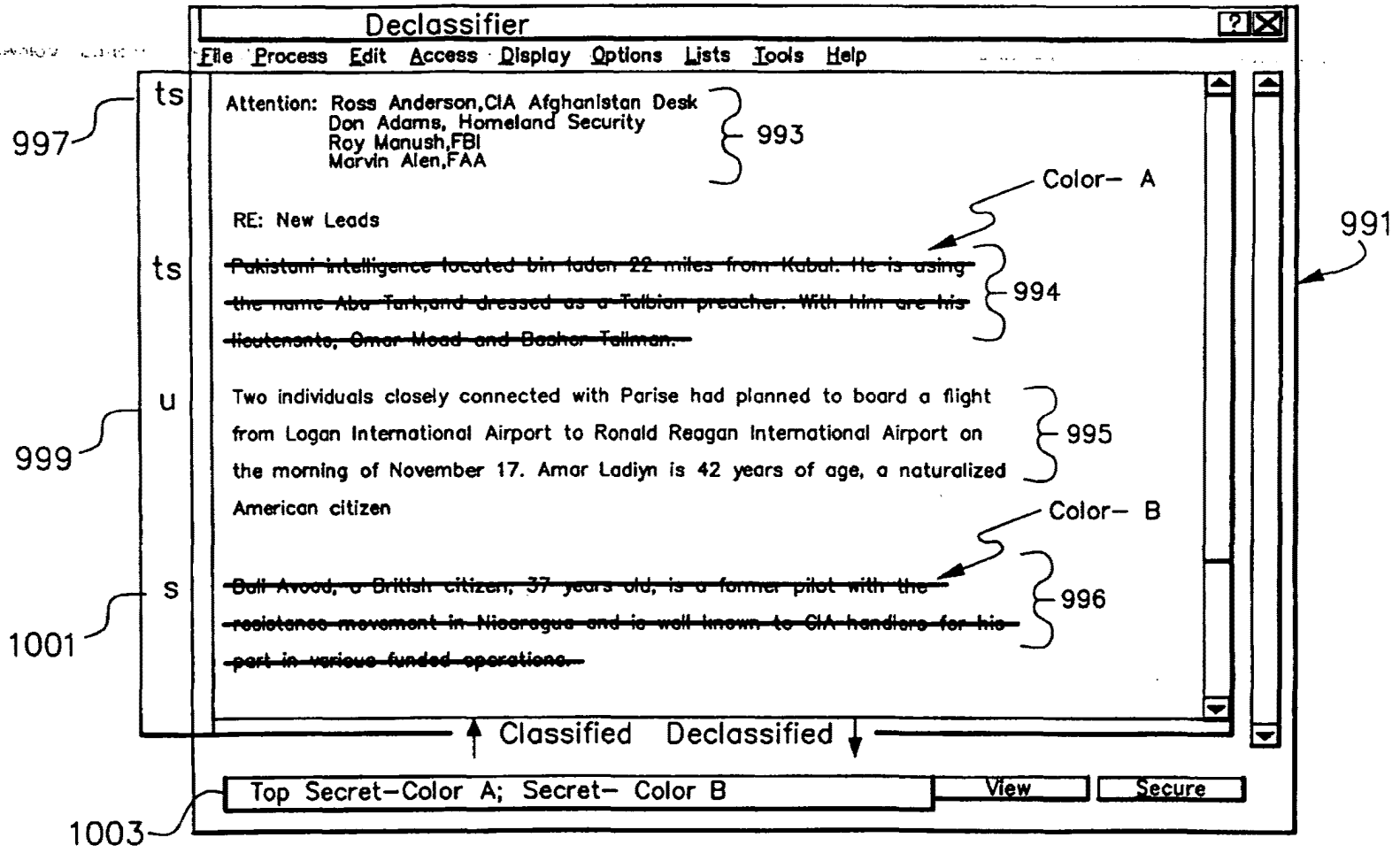

Declassifier
File Process Edit Access Display Options Lists Tools Help
ts
997
Attention: Ross Anderson,CIA Afghanistan Desk
Don Adams, Homeland Security
Roy Manush,FBI
Marvin Alen,FAA
993
Color– A
RE: New Leads
991
ts
Pakistani intelligence located bin laden 22 miles from Kabul. He is using the name Abu Tark,and dressed as a Talbian preacher. With him are his lieutenants, Omar Moad and Bashar Tallman.
994
u
999
Two individuals closely connected with Parise had planned to board a flight from Logan International Airport to Ronald Reagan International Airport on the morning of November 17. Amar Ladiyn is 42 years of age, a naturalized American citizen
995
Color– B
s
1001
Bali Avood, a British citizen, 37 years old, is a former pilot with the resistance movement in Nicaragua and is well known to CIA handlers for his part in various funded operations.
996
Classified Declassified
Top Secret–Color A; Secret– Color B
View
Secure
1003

FIG. C-8B

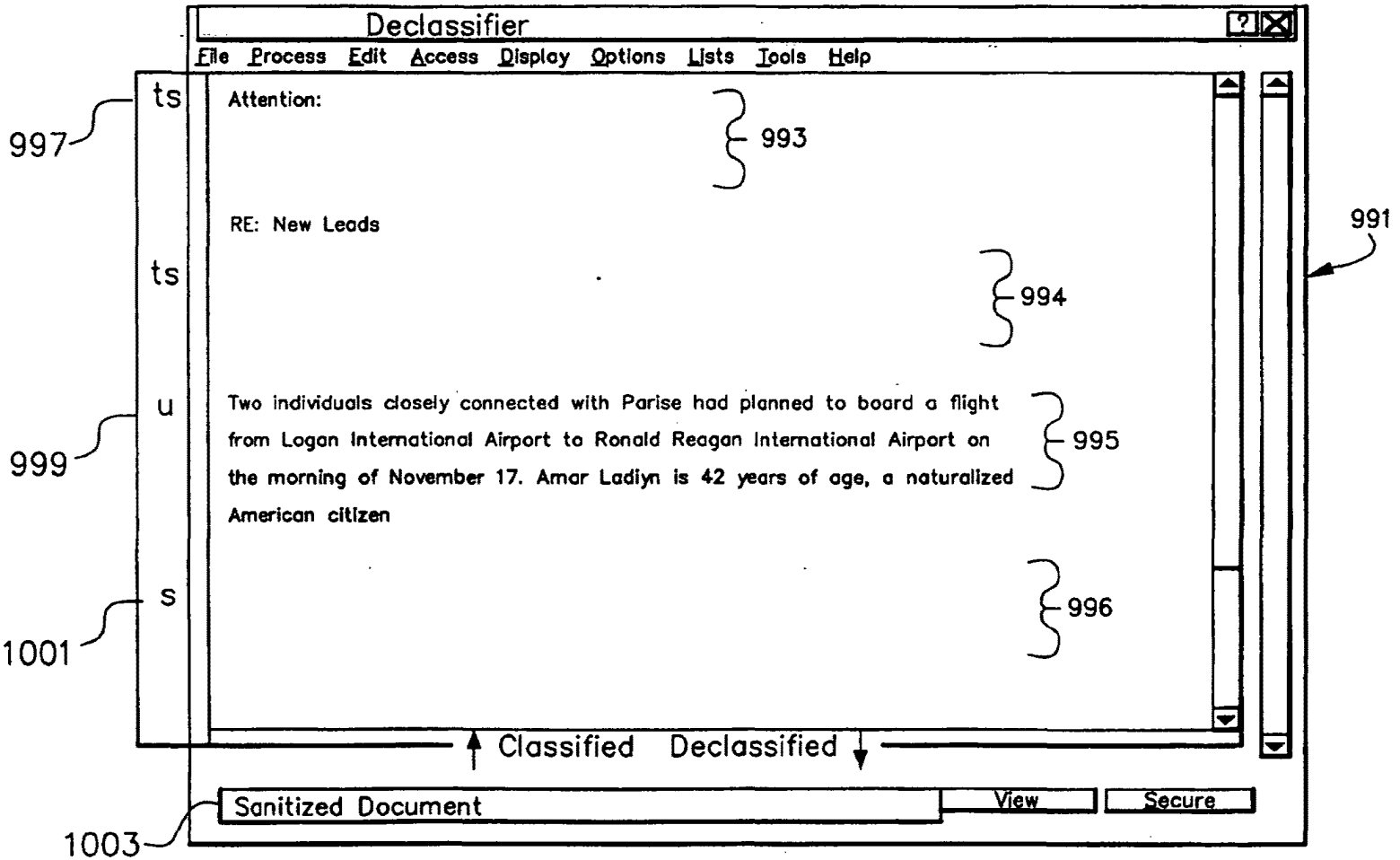

Declassifier
File Process Edit Access Display Options Lists Tools Help
ts
ts
u
s
997
999
1001
Attention:
993
RE: New Leads
994
Two individuals closely connected with Parise had planned to board a flight from Logan International Airport to Ronald Reagan International Airport on the morning of November 17. Amar Ladiyn is 42 years of age, a naturalized American citizen
995
996
991
Classified Declassified
Sanitized Document
View
Secure
1003

FIG. C-8C

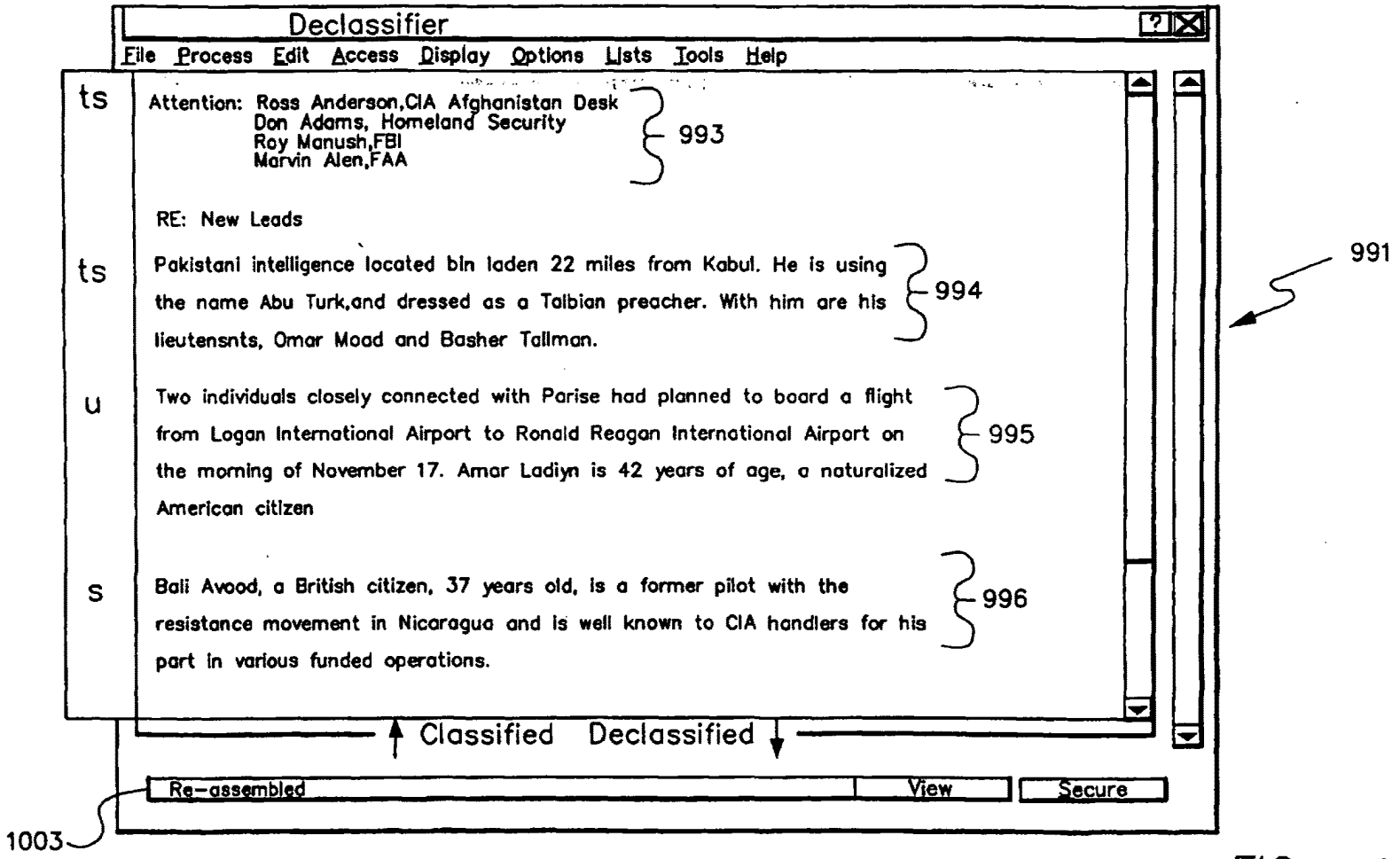

Declassifier
File Process Edit Access Display Options Lists Tools Help
ts
Attention: Ross Anderson,CIA Afghanistan Desk
Don Adams, Homeland Security
Roy Manush,FBI
Marvin Alen,FAA
993
RE: New Leads
ts
Pakistani intelligence located bin laden 22 miles from Kabul. He is using the name Abu Turk,and dressed as a Talbian preacher. With him are his lieutensnts, Omar Moad and Basher Tallman.
994
u
Two individuals closely connected with Parise had planned to board a flight from Logan International Airport to Ronald Reagan International Airport on the morning of November 17. Amar Ladiyn is 42 years of age, a naturalized American citizen
995
s
Bali Avood, a British citizen, 37 years old, is a former pilot with the resistance movement in Nicaragua and is well known to CIA handlers for his part in various funded operations.
996
991
Classified Declassified
Re-assembled
View
Secure
1003

FIG. C-8D

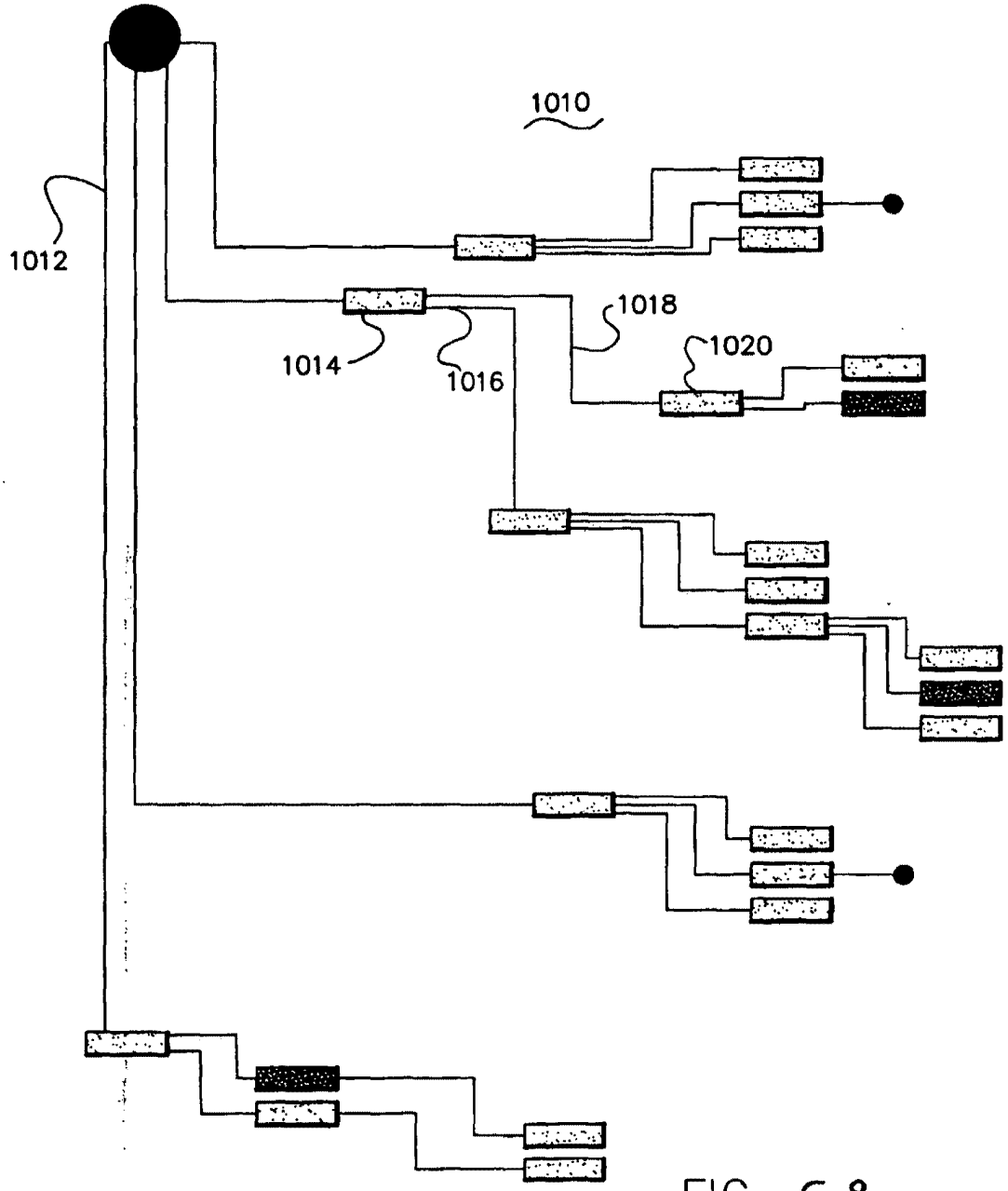


FIG. C-9

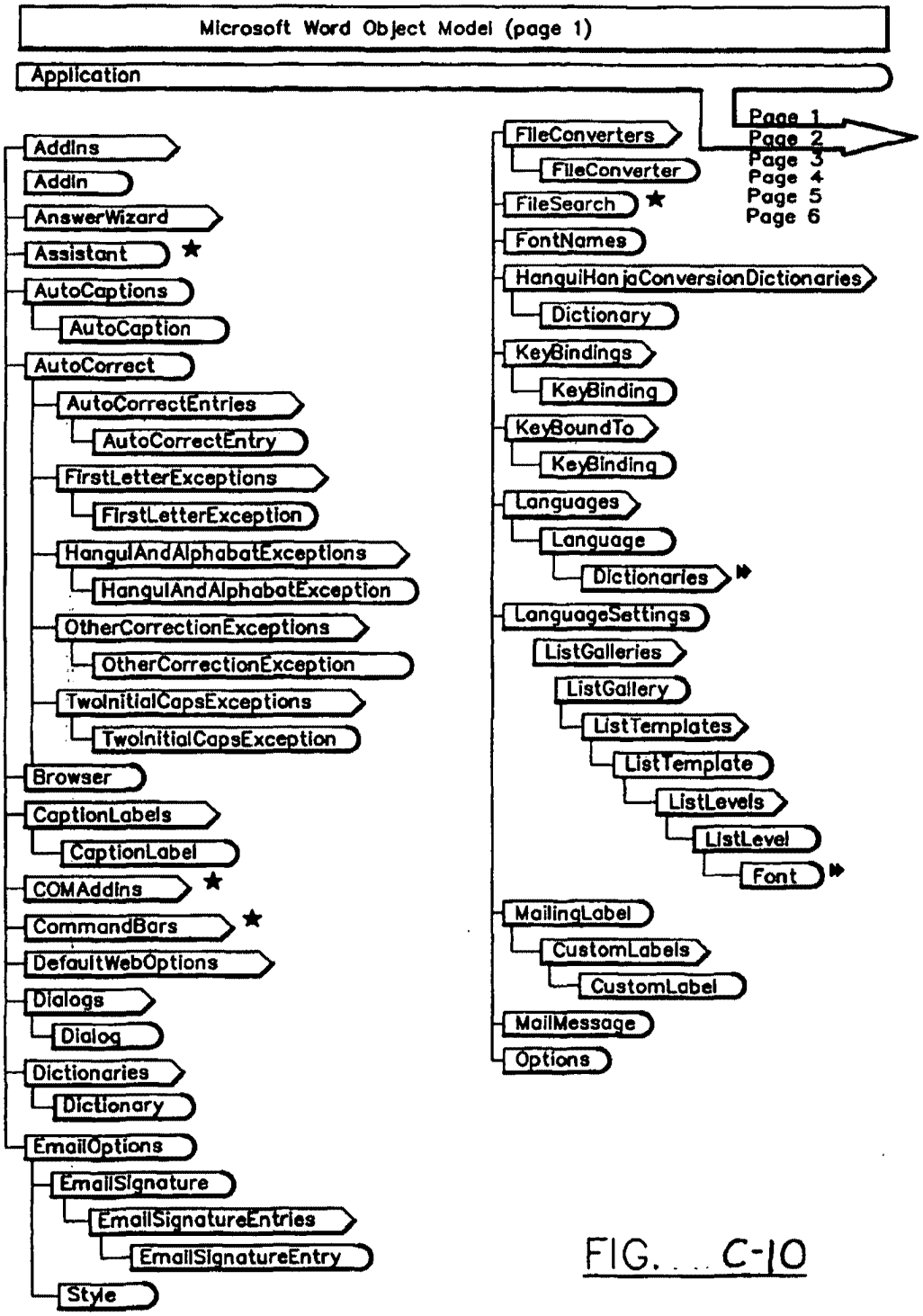

Microsoft Word Object Model (page 1)
Application
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
AddIns
AddIn
AnswerWizard
Assistant
AutoCaptions
AutoCaption
AutoCorrect
AutoCorrectEntries
AutoCorrectEntry
FirstLetterExceptions
FirstLetterException
HangulAndAlphabatExceptions
HangulAndAlphabatException
OtherCorrectionExceptions
OtherCorrectionException
TwoInitialCapsExceptions
TwoInitialCapsException
Browser
CaptionLabels
CaptionLabel
COMAddIns
CommandBars
DefaultWebOptions
Dialogs
Dialog
Dictionaries
Dictionary
EmailOptions
EmailSignature
EmailSignatureEntries
EmailSignatureEntry
Style
FileConverters
FileConverter
FileSearch
FontNames
HanqulHanjaConversionDictionaries
Dictionary
KeyBindings
KeyBinding
KeyBoundTo
KeyBinding
Languages
Language
Dictionaries
LanguageSettings
ListGalleries
ListGallery
ListTemplates
ListTemplate
ListLevels
ListLevel
Font
MailingLabel
CustomLabels
CustomLabel
MailMessage
Options

FIG. C-10

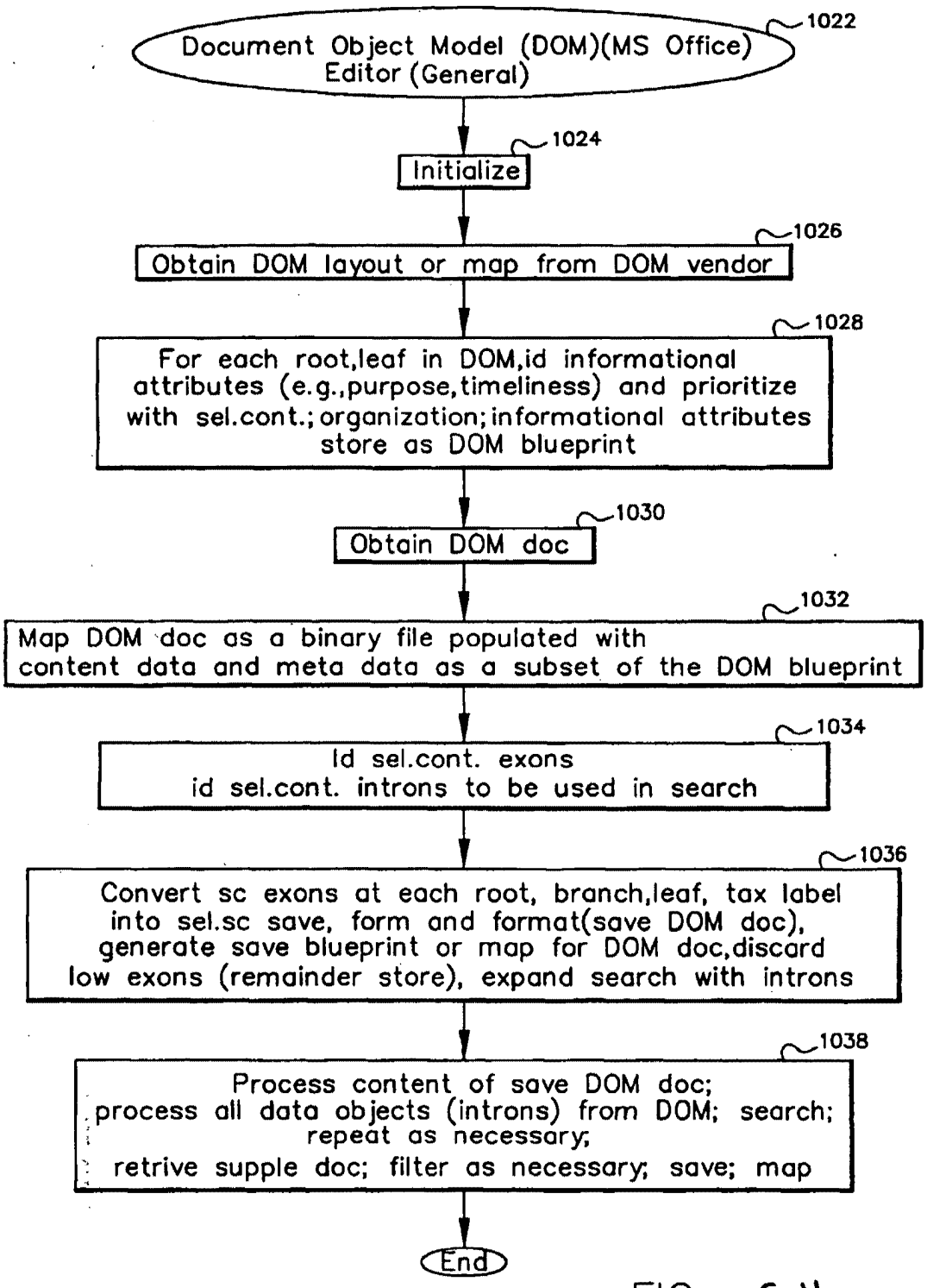

Document Object Model (DOM)(MS Office) Editor (General)
1022
Initialize
1024
Obtain DOM layout or map from DOM vendor
1026
For each root,leaf in DOM,id informational attributes (e.g.,purpose,timeliness) and prioritize with sel.cont.; organization; informational attributes store as DOM blueprint
1028
Obtain DOM doc
1030
Map DOM doc as a binary file populated with content data and meta data as a subset of the DOM blueprint
1032
Id sel.cont. exons
id sel.cont. introns to be used in search
1034
Convert sc exons at each root, branch,leaf, tax label into sel.sc save, form and format(save DOM doc), generate save blueprint or map for DOM doc,discard low exons (remainder store), expand search with introns
1036
Process content of save DOM doc;
process all data objects (introns) from DOM; search; repeat as necessary;
retrive supple doc; filter as necessary; save; map
1038
End

FIG. C-11

FIG. C-12A

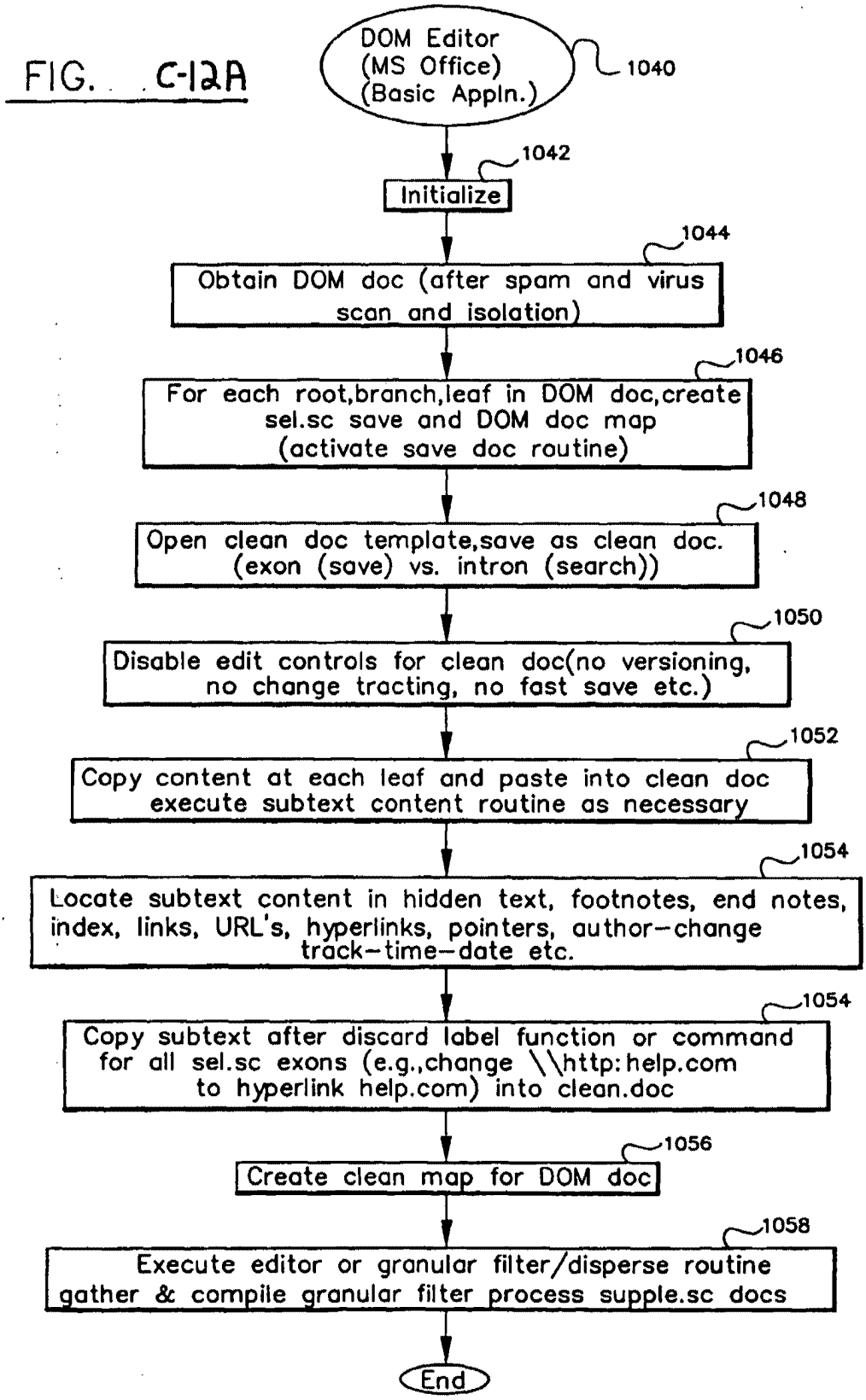

DOM Editor
(MS Office)
(Basic Appln.)
1040
1042
Initialize
1044
Obtain DOM doc (after spam and virus scan and isolation)
1046
For each root,branch,leaf in DOM doc,create sel.sc save and DOM doc map (activate save doc routine)
1048
Open clean doc template,save as clean doc. (exon (save) vs. intron (search))
1050
Disable edit controls for clean doc(no versioning, no change tracting, no fast save etc.)
1052
Copy content at each leaf and paste into clean doc execute subtext content routine as necessary
1054
Locate subtext content in hidden text, footnotes, end notes, index, links, URL's, hyperlinks, pointers, author-change track-time-date etc.
1054
Copy subtext after discard label function or command for all sel.sc exons (e.g.,change \\http:help.com to hyperlink help.com) into clean.doc
1056
Create clean map for DOM doc
1058
Execute editor or granular filter/disperse routine gather & compile granular filter process supple.sc docs
End

FIG. C-12B -- Table 1: Categorical Identifiers For Personal Identity

name
address(es) and variants
telephone number(s) and variants
username
biometrics
gene typing
photograph
date of birth
age
marital status
gender
sexual orientation
sexual proclivities
disabilities
tattoos
scars
visible or functional injuries
age/age range
hair color
eye color
race profile
educational history
employment history
home price
ethnicity
personal interests
personal descriptive information (e.g., SWHM 38, Professional)
physical stigmata
skill set
credit history
credit reports (formal NCR, etc.)
social security number
patient ID or other location- or process-specific
user assignment
insurance number
credit card numbers
birthplace
heritage
health history
political party
political beliefs
association(s)
frequent flyer/buyer club info
remittance advice
investigation evidence
court evidence
EDI/EDIFACT records
applications
personal web sites
Chain of trust forms
Chain of custody forms
skill set
religion
personal interests
travel log
number of siblings
business partners
business name
profession
account numbers (banking, services, suppliers)
service providers (physicians, insurers, hospitals, clinics, etc.)
X-rays
surveillance
dental charts
medical records
account balances
account transfer or transaction amounts
income range
neighborhood/city/region/country
license (driver, occupational, professional)
vehicle registration (license, tag, plate, etc.)
vehicle identification
vehicle make, type, model, color, year
date of life events
incident reports (legal, criminal, health services, news)
accident reports (auto, OSHA, EPA, EEOC, etc.)
criminal convictions
court records
abuse records
divorce proceedings
bankruptcy records
organization registrations
Corporation officers and registrations
tax records (chattel, land, local, state, Federal, and special use taxes)
property ownership
permit applications
donor lists
news reports
family history
family relationships
family health history
legal documents
consent forms
newsgroup postings

FIG. C-12C -- Table 2: Ranked Identity Category

name	disabilities
address	tattoos
telephone	scars
username	injuries
biometrics	age range
gene typing	hair color
photograph	eye color
date of birth	race profile
age	education
marital status	employment
sex	personal interests
sexual orientation	
sexual proclivities	

FIG. C-12D -- Table 3: Categorical Identifiers for Manufacturing Business

- Manufacturing
  - product brand names
  - product generic name
  - drawings
  - tools (brand names and generic names)
    - hand tools
    - software
    - machines
  - software, computer programs, etc.
  - Research and Development
    - competitors products, competitor names, patent numbers, patent titles, project names, project personnel
- Sales
  - personnel
  - competitors
  - sales data
    - quantity
    - geographic distribution
  - customers
    - names, addresses, contacts
  - sales projections
- Financial
  - chart of accounts
  - ledgers
  - financial statements
  - tax returns
- Human Resources
  - see categorical identifiers for personal identity

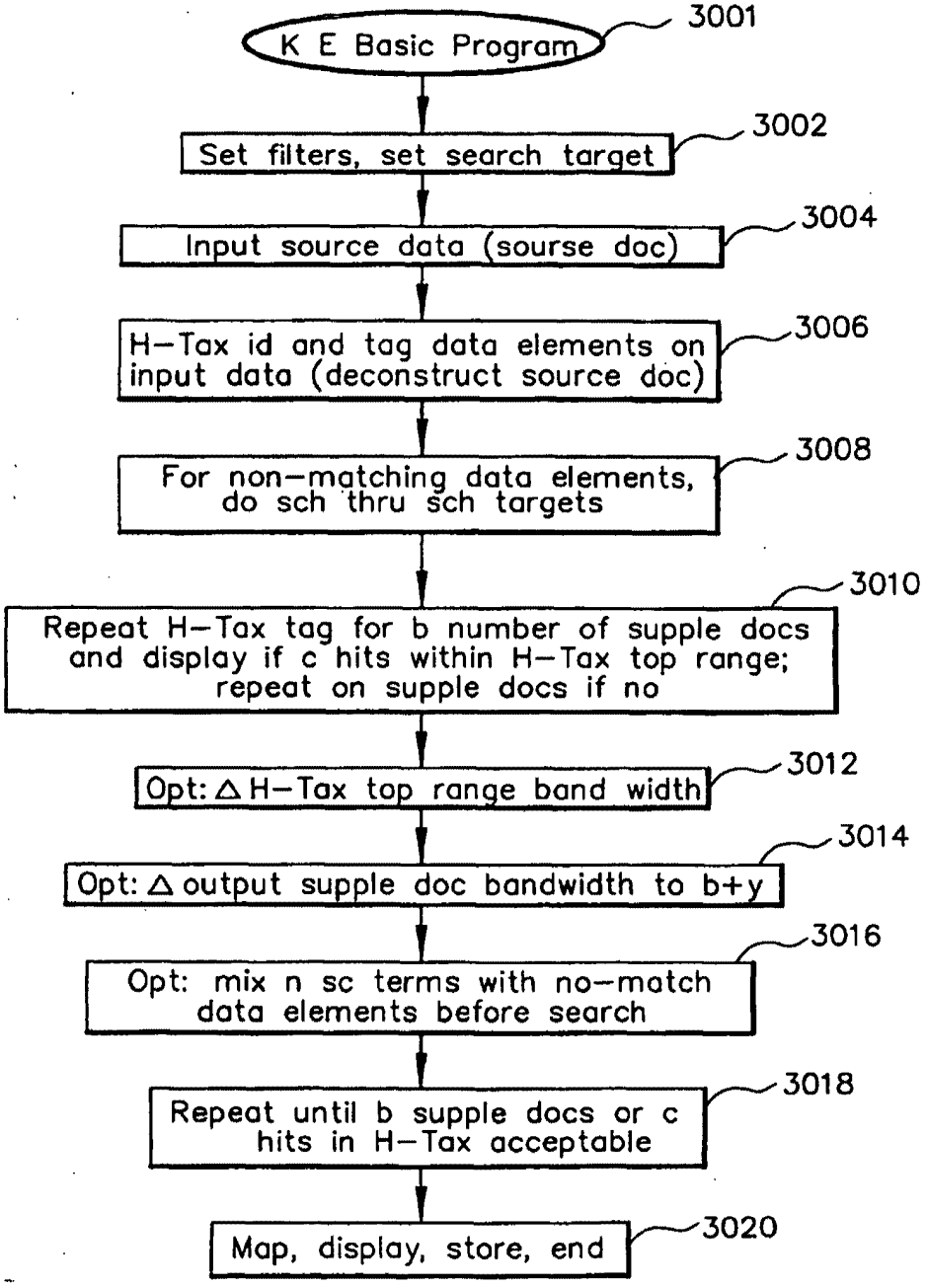

K E Basic Program
3001
Set filters, set search target
3002
Input source data (sourse doc)
3004
H–Tax id and tag data elements on input data (deconstruct source doc)
3006
For non–matching data elements, do sch thru sch targets
3008
Repeat H–Tax tag for b number of supple docs and display if c hits within H–Tax top range; repeat on supple docs if no
3010
Opt: Δ H–Tax top range band width
3012
Opt: Δ output supple doc bandwidth to b+y
3014
Opt: mix n sc terms with no–match data elements before search
3016
Repeat until b supple docs or c hits in H–Tax acceptable
3018
Map, display, store, end
3020

FIG. C-13

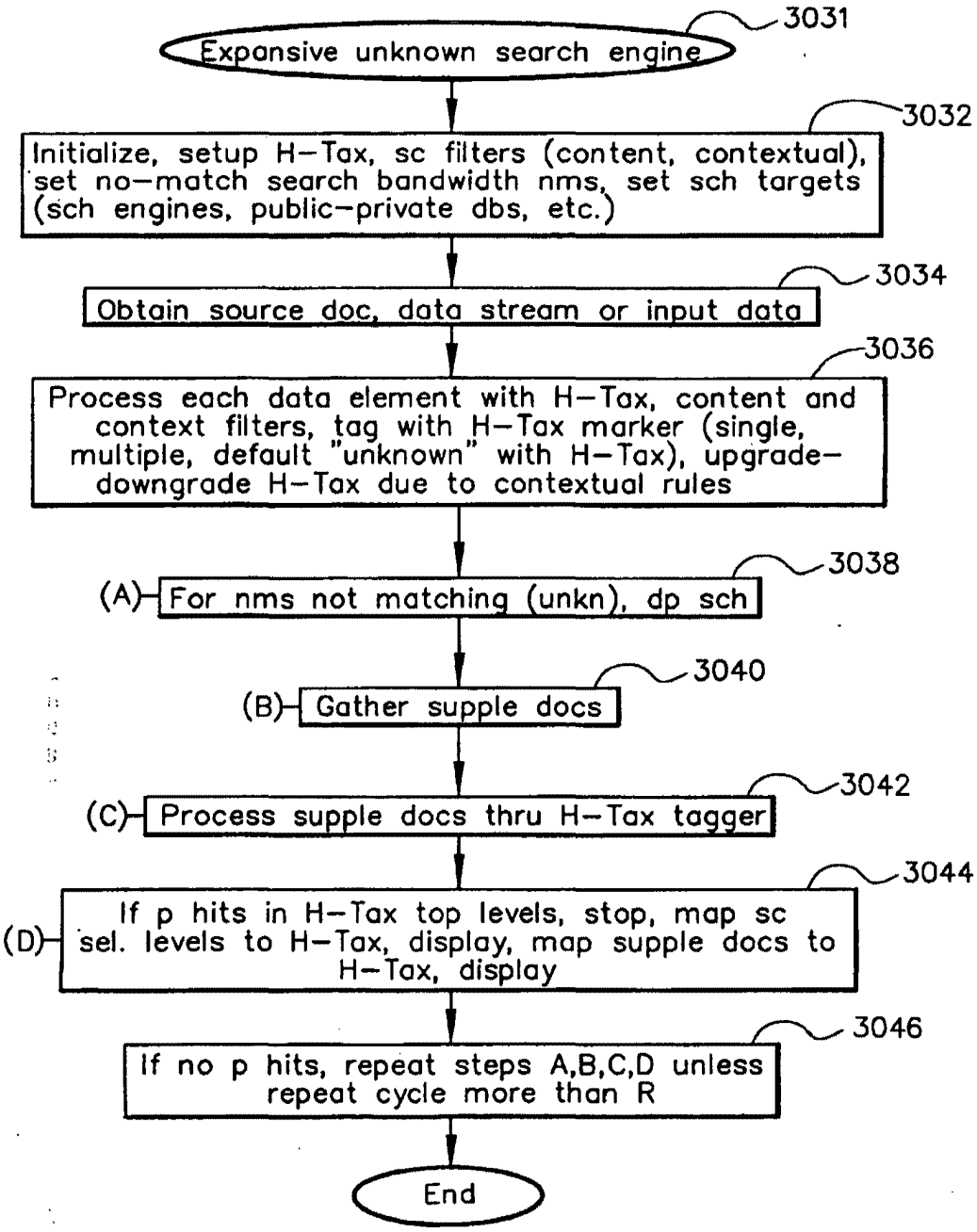

3031
Expansive unknown search engine
3032
Initialize, setup H-Tax, sc filters (content, contextual), set no-match search bandwidth nms, set sch targets (sch engines, public-private dbs, etc.)
3034
Obtain source doc, data stream or input data
3036
Process each data element with H-Tax, content and context filters, tag with H-Tax marker (single, multiple, default "unknown" with H-Tax), upgrade-downgrade H-Tax due to contextual rules
3038
(A) For nms not matching (unkn), dp sch
3040
(B) Gather supple docs
3042
(C) Process supple docs thru H-Tax tagger
3044
(D) If p hits in H-Tax top levels, stop, map sc sel. levels to H-Tax, display, map supple docs to H-Tax, display
3046
If no p hits, repeat steps A,B,C,D unless repeat cycle more than R
End

FIG. C-14

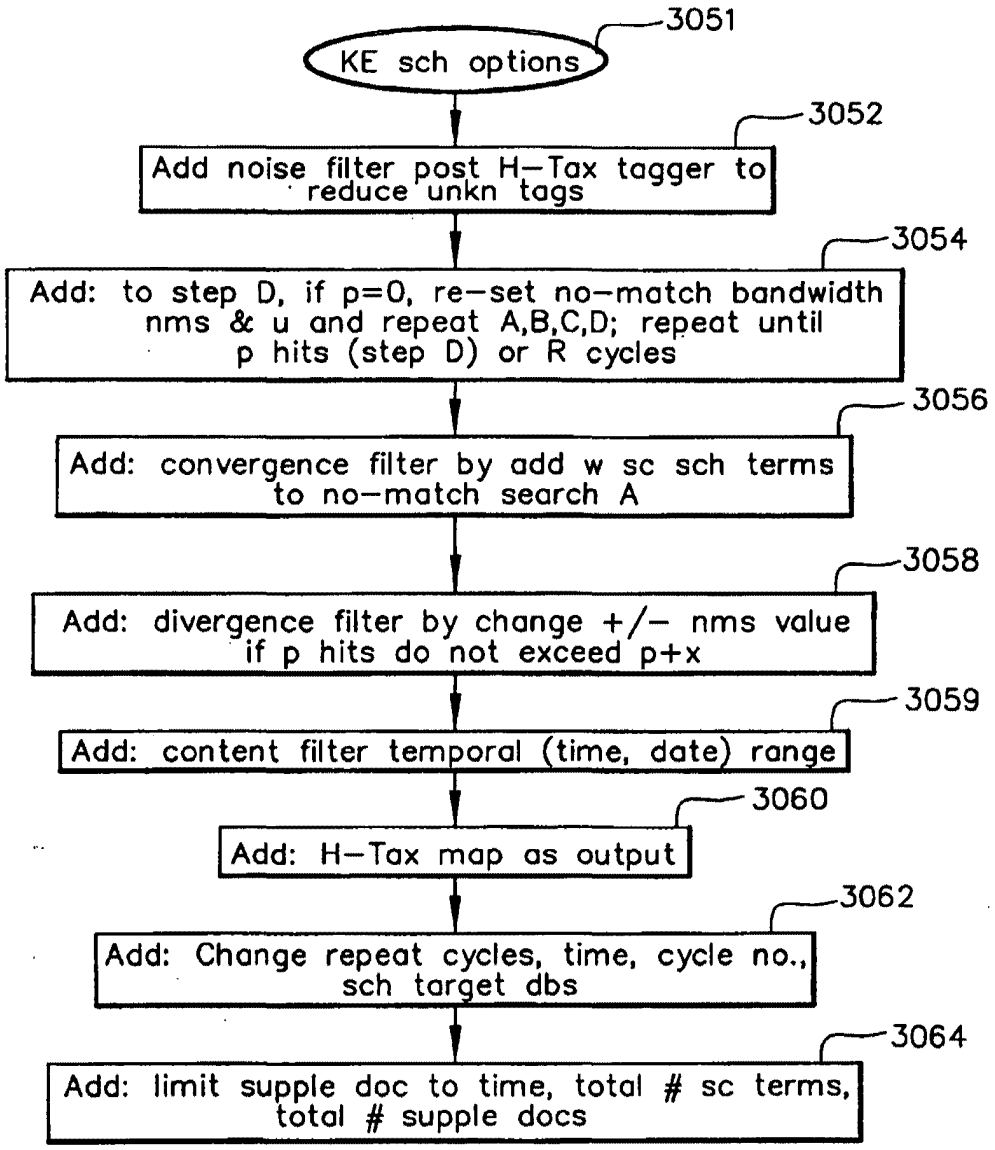

3051
KE sch options
3052
Add noise filter post H–Tax tagger to reduce unkn tags
3054
Add: to step D, if p=0, re–set no–match bandwidth nms & u and repeat A,B,C,D; repeat until p hits (step D) or R cycles
3056
Add: convergence filter by add w sc sch terms to no–match search A
3058
Add: divergence filter by change +/– nms value if p hits do not exceed p+x
3059
Add: content filter temporal (time, date) range
3060
Add: H–Tax map as output
3062
Add: Change repeat cycles, time, cycle no., sch target dbs
3064
Add: limit supple doc to time, total # sc terms, total # supple docs

FIG. C-15

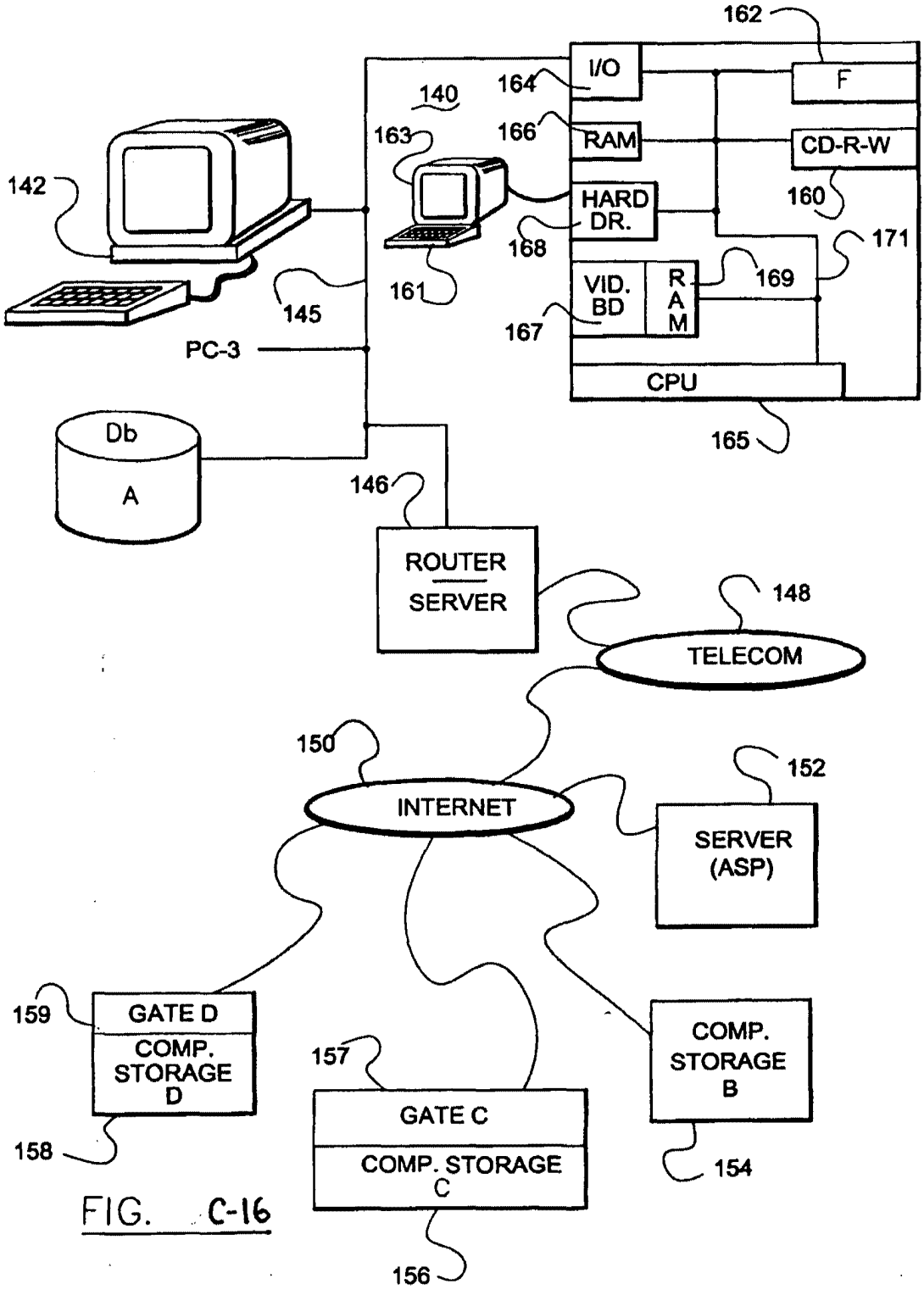

162
I/O
F
164
140
163
166
RAM
CD-R-W
142
HARD DR.
160
168
171
VID. BD
R A M
169
167
145
161
PC-3
CPU
165
Db
A
146
ROUTER SERVER
148
TELECOM
150
152
INTERNET
SERVER (ASP)
159
GATE D
COMP. STORAGE D
157
COMP. STORAGE B
GATE C
COMP. STORAGE C
158
154
156

FIG. C-16

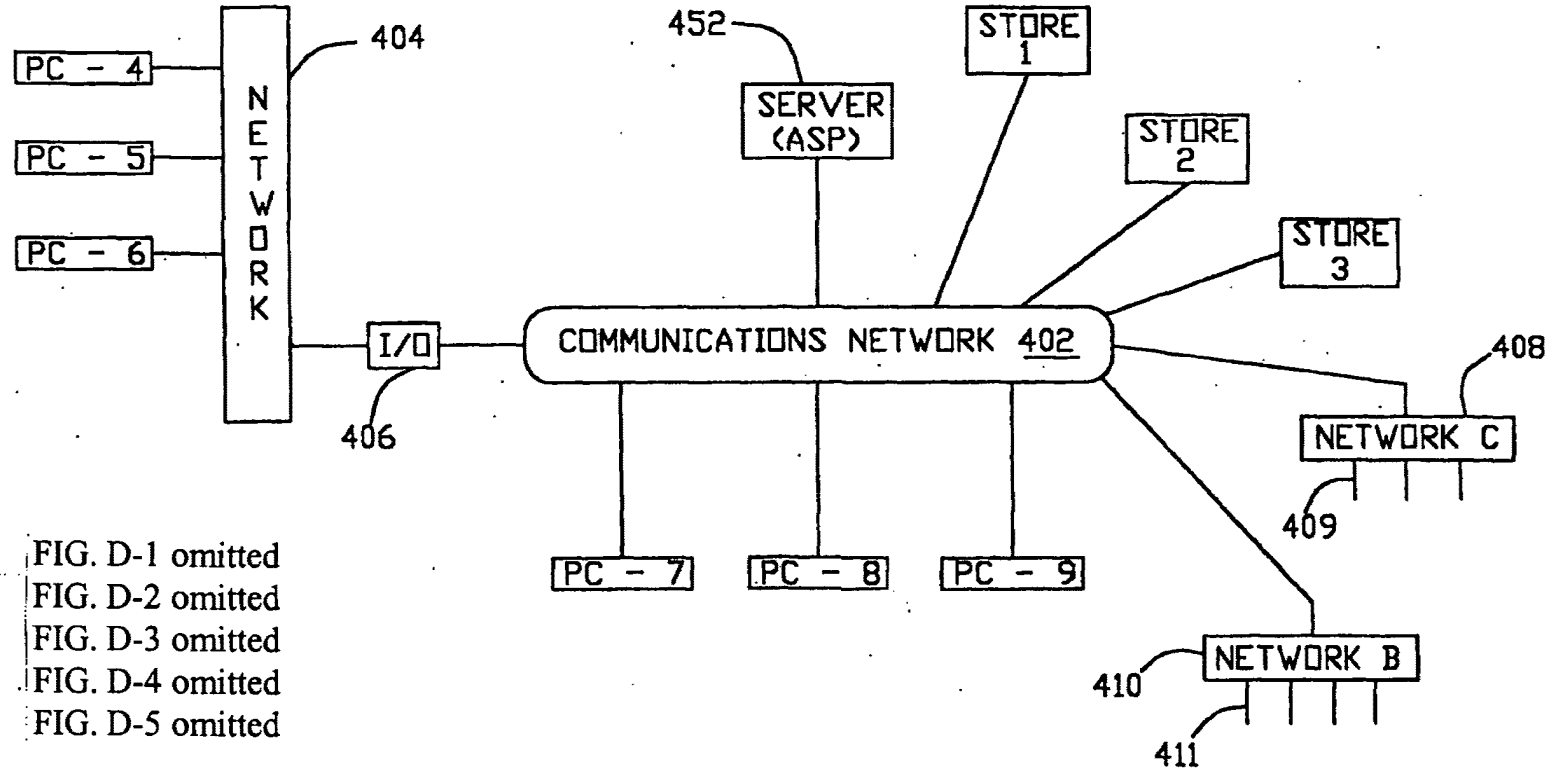


FIG. D-1 omitted
FIG. D-2 omitted
FIG. D-3 omitted
FIG. D-4 omitted
FIG. D-5 omitted

FIG. D-6

FIG. D-7a

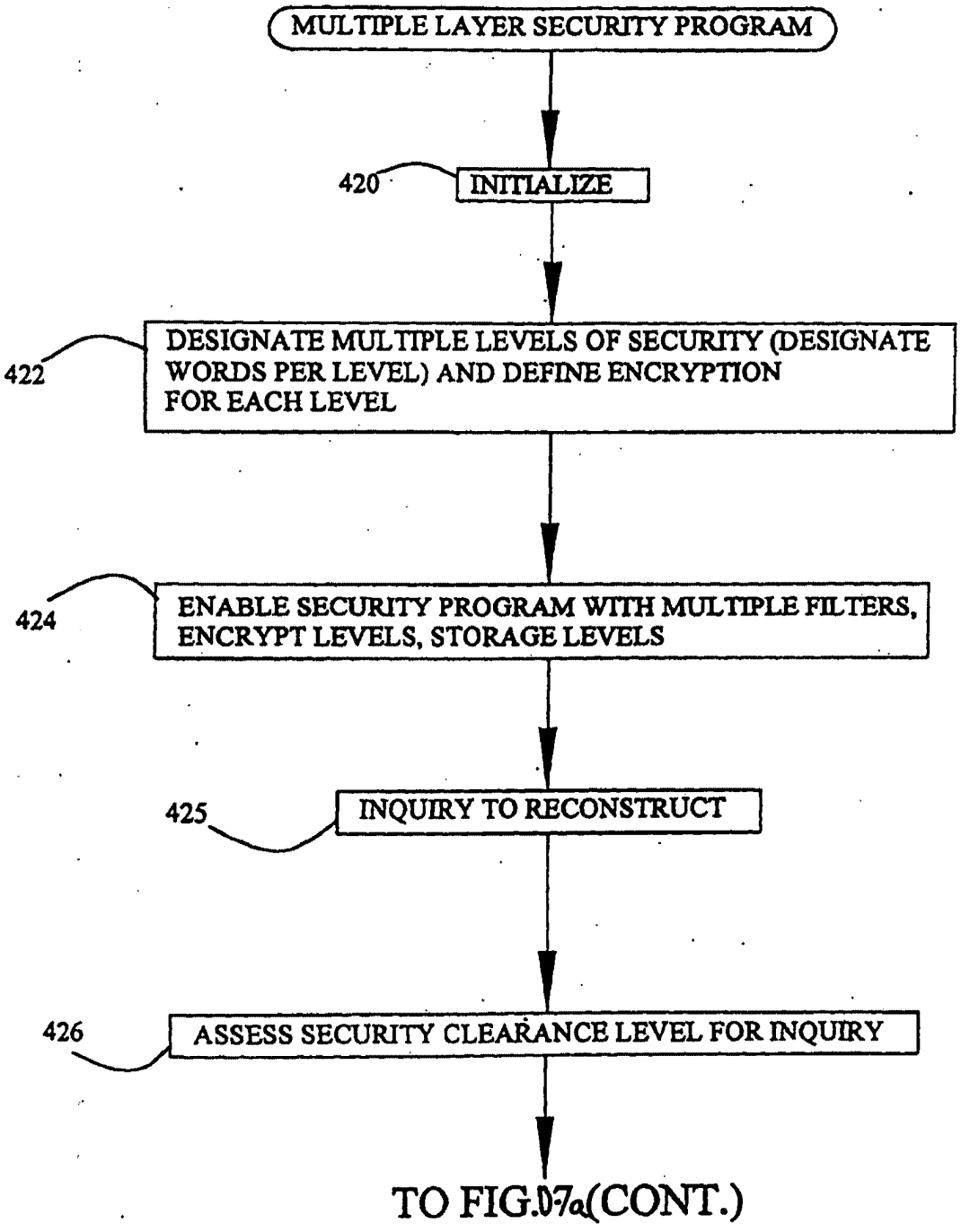
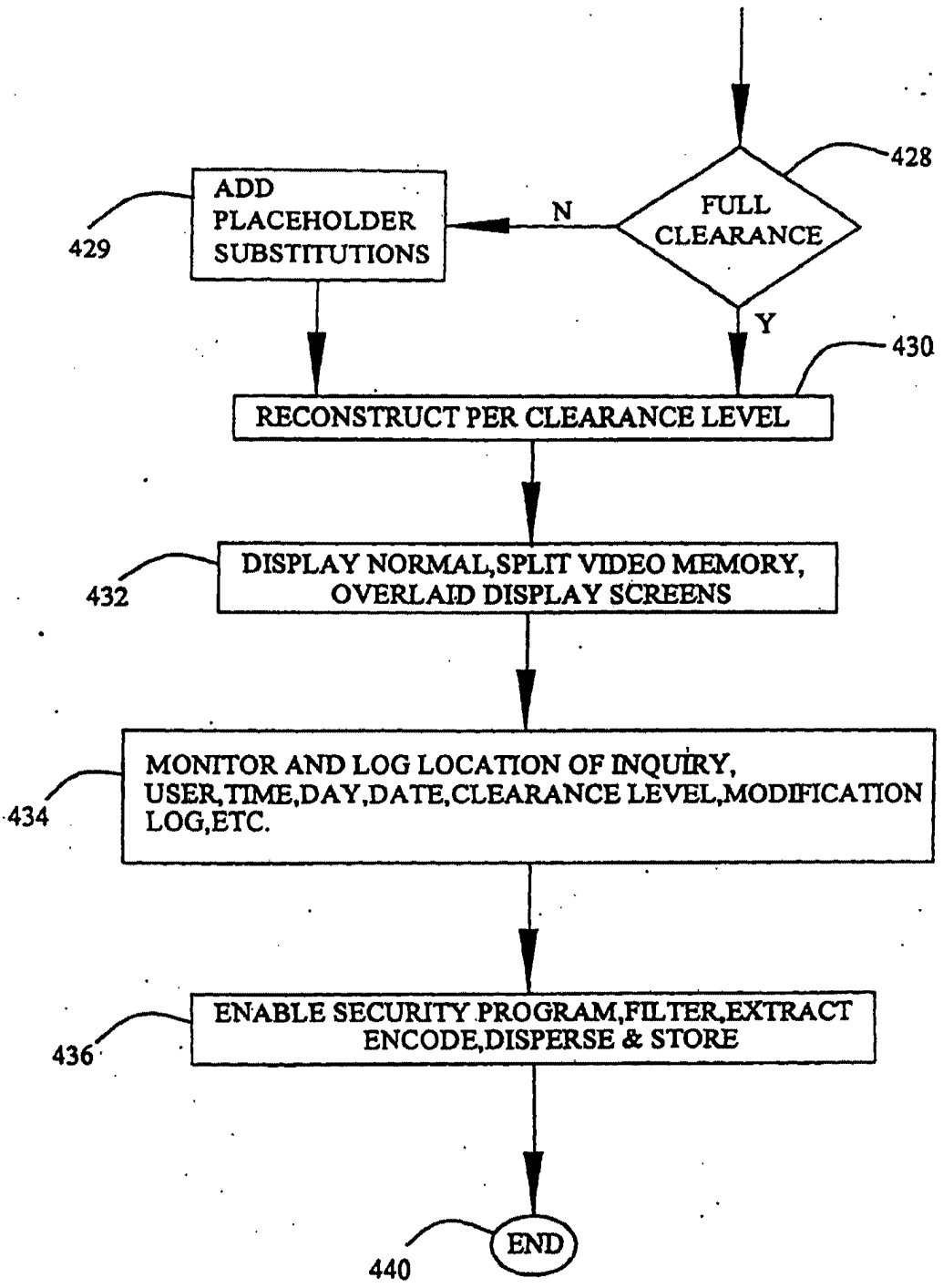

MULTIPLE LAYER SECURITY PROGRAM
420
INITIALIZE
422
DESIGNATE MULTIPLE LEVELS OF SECURITY (DESIGNATE WORDS PER LEVEL) AND DEFINE ENCRYPTION FOR EACH LEVEL
424
ENABLE SECURITY PROGRAM WITH MULTIPLE FILTERS, ENCRYPT LEVELS, STORAGE LEVELS
425
INQUIRY TO RECONSTRUCT
426
ASSESS SECURITY CLEARANCE LEVEL FOR INQUIRY
TO FIG.07a(CONT.)

FIG.7a(CONT.)


428
FULL CLEARANCE
N
ADD PLACEHOLDER SUBSTITUTIONS
429
Y
430
RECONSTRUCT PER CLEARANCE LEVEL
DISPLAY NORMAL,SPLIT VIDEO MEMORY, OVERLAID DISPLAY SCREENS
432
MONITOR AND LOG LOCATION OF INQUIRY, USER,TIME,DAY,DATE,CLEARANCE LEVEL,MODIFICATION LOG,ETC.
434
ENABLE SECURITY PROGRAM,FILTER,EXTRACT ENCODE,DISPERSE & STORE
436
END
440

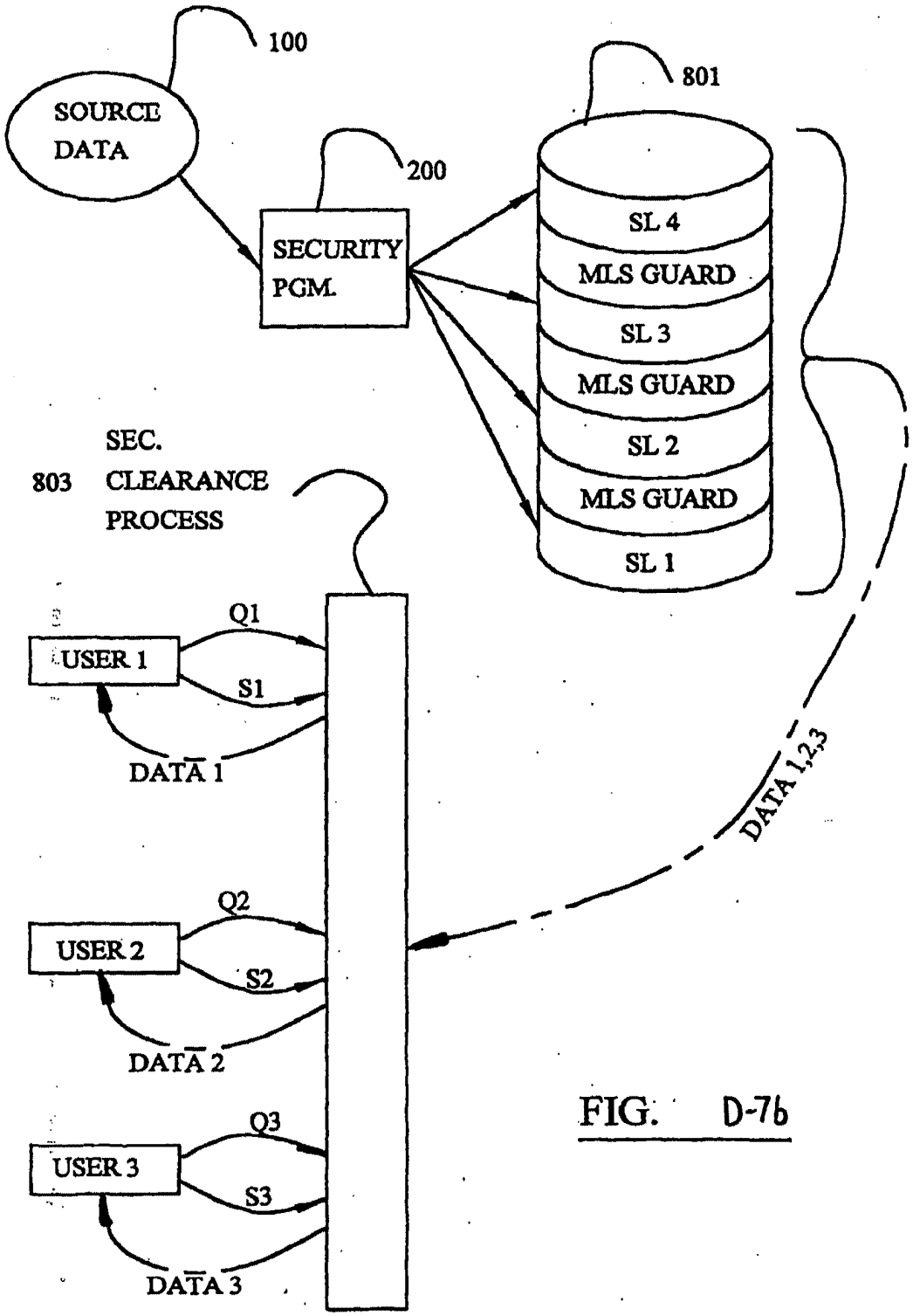

100
SOURCE DATA
200
SECURITY PGM.
801
SL 4
MLS GUARD
SL 3
MLS GUARD
SL 2
MLS GUARD
SL 1
SEC. CLEARANCE PROCESS
803
Q1
USER 1
S1
DATA 1
Q2
USER 2
S2
DATA 2
Q3
USER 3
S3
DATA 3
DATA 1,2,3

FIG. D-7b

FIG. D-8

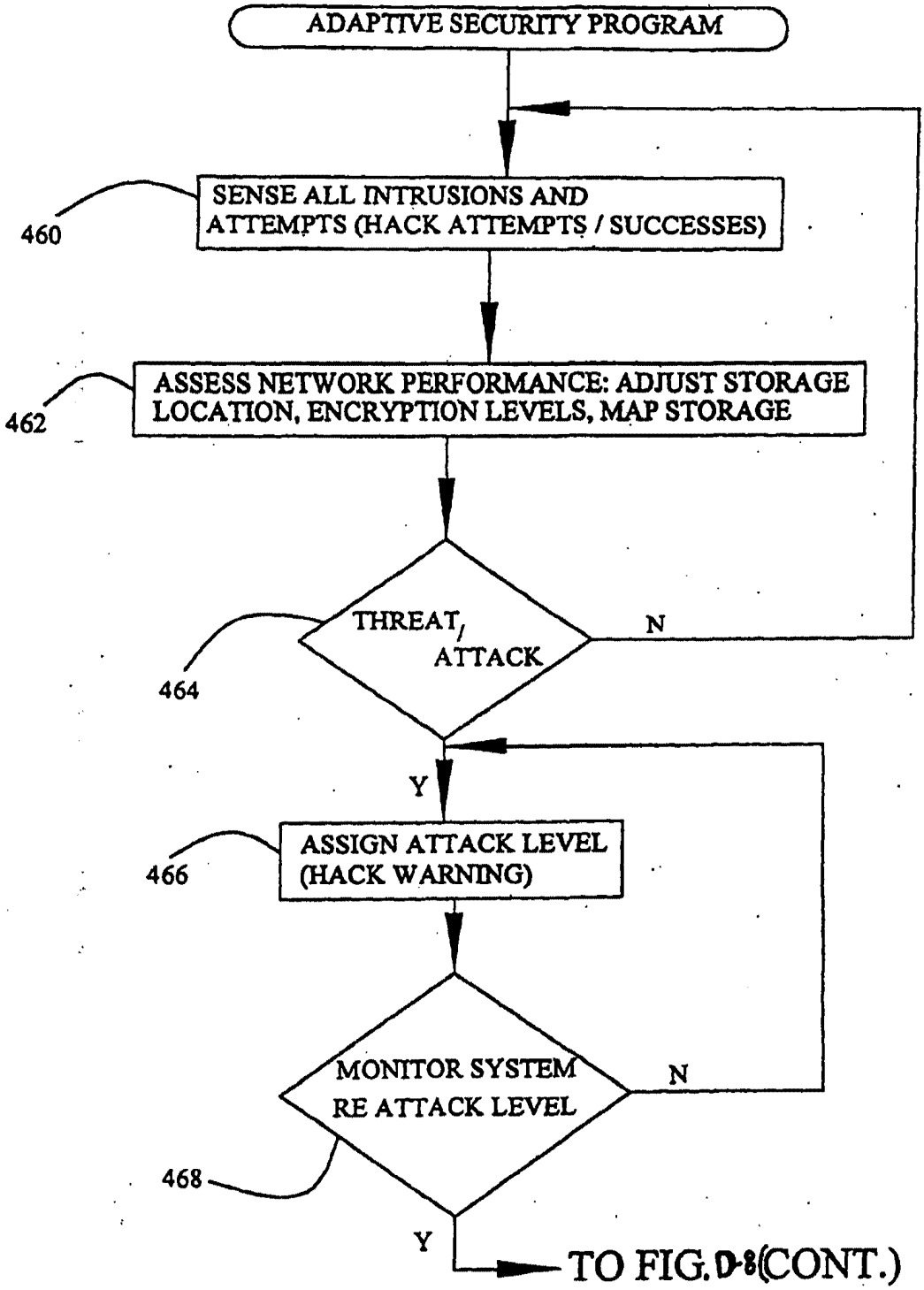

ADAPTIVE SECURITY PROGRAM
460
SENSE ALL INTRUSIONS AND ATTEMPTS (HACK ATTEMPTS / SUCCESSES)
462
ASSESS NETWORK PERFORMANCE: ADJUST STORAGE LOCATION, ENCRYPTION LEVELS, MAP STORAGE
464
THREAT/ ATTACK
N
Y
466
ASSIGN ATTACK LEVEL (HACK WARNING)
MONITOR SYSTEM RE ATTACK LEVEL
N
468
Y
TO FIG. D-8 (CONT.)

FIG. D-8 (CONT.)

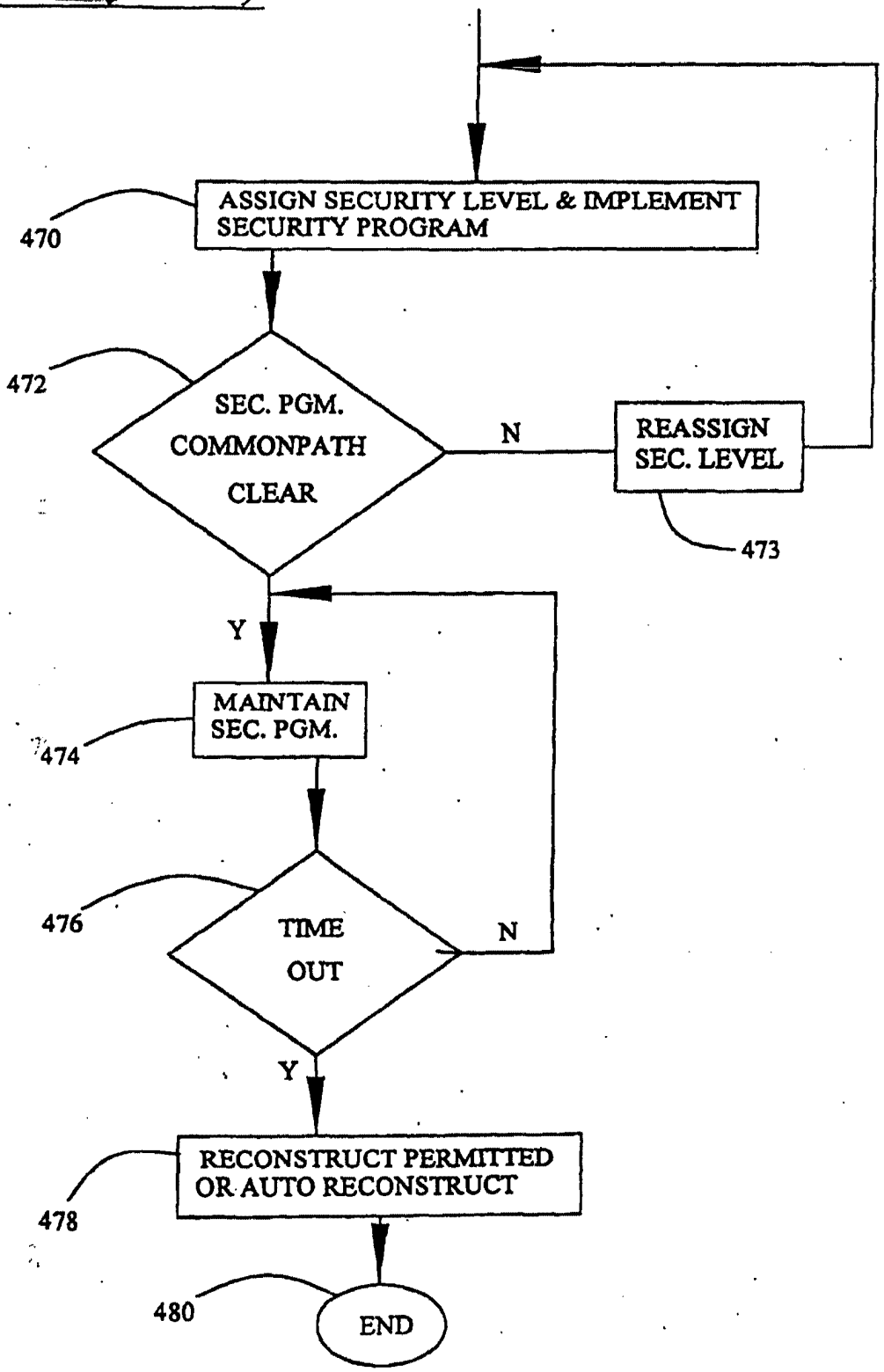

ASSIGN SECURITY LEVEL & IMPLEMENT SECURITY PROGRAM
470
SEC. PGM. COMMONPATH CLEAR
472
N
REASSIGN SEC. LEVEL
473
Y
MAINTAIN SEC. PGM.
474
TIME OUT
476
N
Y
RECONSTRUCT PERMITTED OR AUTO RECONSTRUCT
478
END
480

FIG. D-8A

Standard Automatic Defenses Matrix Table

Mode	Normal	Threat	Attack
Encryption	Targeted Encryption	Full Encryption	Multi Type Encryption
Extraction	Plain-text Extraction	Extraction of Encrypted Data	Extraction of Multi Type Encryption
Distributed Dispersion	Single Storage Location	Several Storage Locations	Many Storage Locations
Display	Single display	Color/Dither Protection	Multiple Displays

FIG. D-8B Optional Automatic Defenses Matrix Table

Mode	Normal	Threat	Attack
Substitution of Code Words	None	Partial	Many
Substitution of Misinformation	None	Partial	Many
Controlled Release-Storage	Full	Partial	Conditional
Storage Locations	2	4	10 or more
Time for release	Anytime	Working Hours	Conditional
Authorized Users	Many	Partial	Conditional
What to Release	All	Partial	Conditional
Secret Sharing	None	Two Users	As Configured

FIG. D-8C

Security Meter Module Table

	Normal Mode	Threat Mode	Attack Mode
ENCRYPTION	Targeted encryption	Full encryption	Multi layer encryption
	(Secret sharing)	(Secret sharing)	(Secret sharing)
EXTRACTION	Plain-text extraction	Extraction of encrypted Data	Extraction of multi encryption
Distributed Storage	1 critical storage	few critical storage	many critical storage
Controlled Release-Storage			
	Storage # ID		
	Time for release		
	Authorized Users		
	What to release		
	Special conditions	2 users online	3 or more users
Display	single display	single display	multiple displays
Substitution of code words	No	No	No

FIG. D-8D

Normal Work Mode Table

	Extraction Level 1	Level 2	Level 3	Level 4	Storage Web	Offline	Remote	Removable	Local
social security	X						X		
credit card	X						X		
last name	X						X		
number	X						X		
telephone			X					X	
name			X					X	
URL			X					X	
e-mail			X					X	
uppercase			X					X	
initial cap			X					X	
currency				X					X
postal code				X					X
address				X					X
location				X					X
date				X					X

FIG. D-8E

Threat Mode Table

	Extraction				Storage				
	Level 1	Level 2	Level 3	Level 4	Web	Offline	Remote	Removable	Local
social security	X							X	
credit card	X							X	
last name	X							X	
number	X							X	
telephone		X						X	
name		X						X	
URL		X						X	
e-mail		X						X	
uppercase		X						X	
initial capital		X						X	
currency			X						X
postal code			X						X
address			X						X
location			X						X
date			X						X

FIG. D-8F

Attack Mode Table

	Extraction Level 1	Level 2	Level 3	Level 4	Storage Web	Offline	Remote	Removable
social security	X					X		
credit card	X					X		
last name	X					X		
number	X					X		
telephone	X						X	
name	X						X	
URL	X						X	
e-mail	X						X	
uppercase	X						X	
initial capital	X						X	
currency		X					X	
postal code		X					X	
address		X					X	
location		X					X	
date		X					X	

FIG. D-9

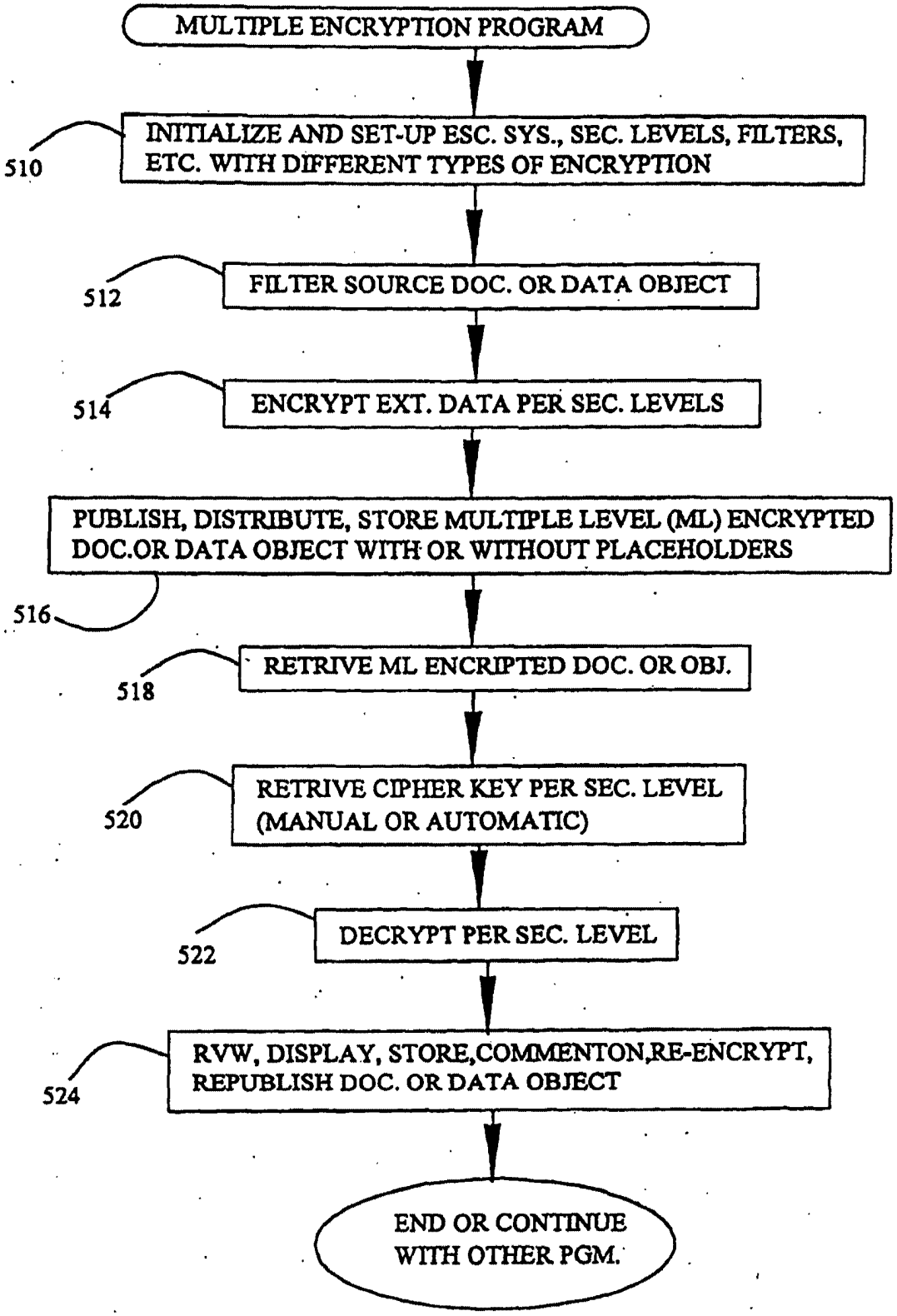

MULTIPLE ENCRYPTION PROGRAM
510
INITIALIZE AND SET-UP ESC. SYS., SEC. LEVELS, FILTERS, ETC. WITH DIFFERENT TYPES OF ENCRYPTION
512
FILTER SOURCE DOC. OR DATA OBJECT
514
ENCRYPT EXT. DATA PER SEC. LEVELS
516
PUBLISH, DISTRIBUTE, STORE MULTIPLE LEVEL (ML) ENCRYPTED DOC.OR DATA OBJECT WITH OR WITHOUT PLACEHOLDERS
518
RETRIVE ML ENCRIPTED DOC. OR OBJ.
520
RETRIVE CIPHER KEY PER SEC. LEVEL (MANUAL OR AUTOMATIC)
522
DECRYPT PER SEC. LEVEL
524
RVW, DISPLAY, STORE,COMMENTON,RE-ENCRYPT, REPUBLISH DOC. OR DATA OBJECT
END OR CONTINUE WITH OTHER PGM.

FIG. D-10

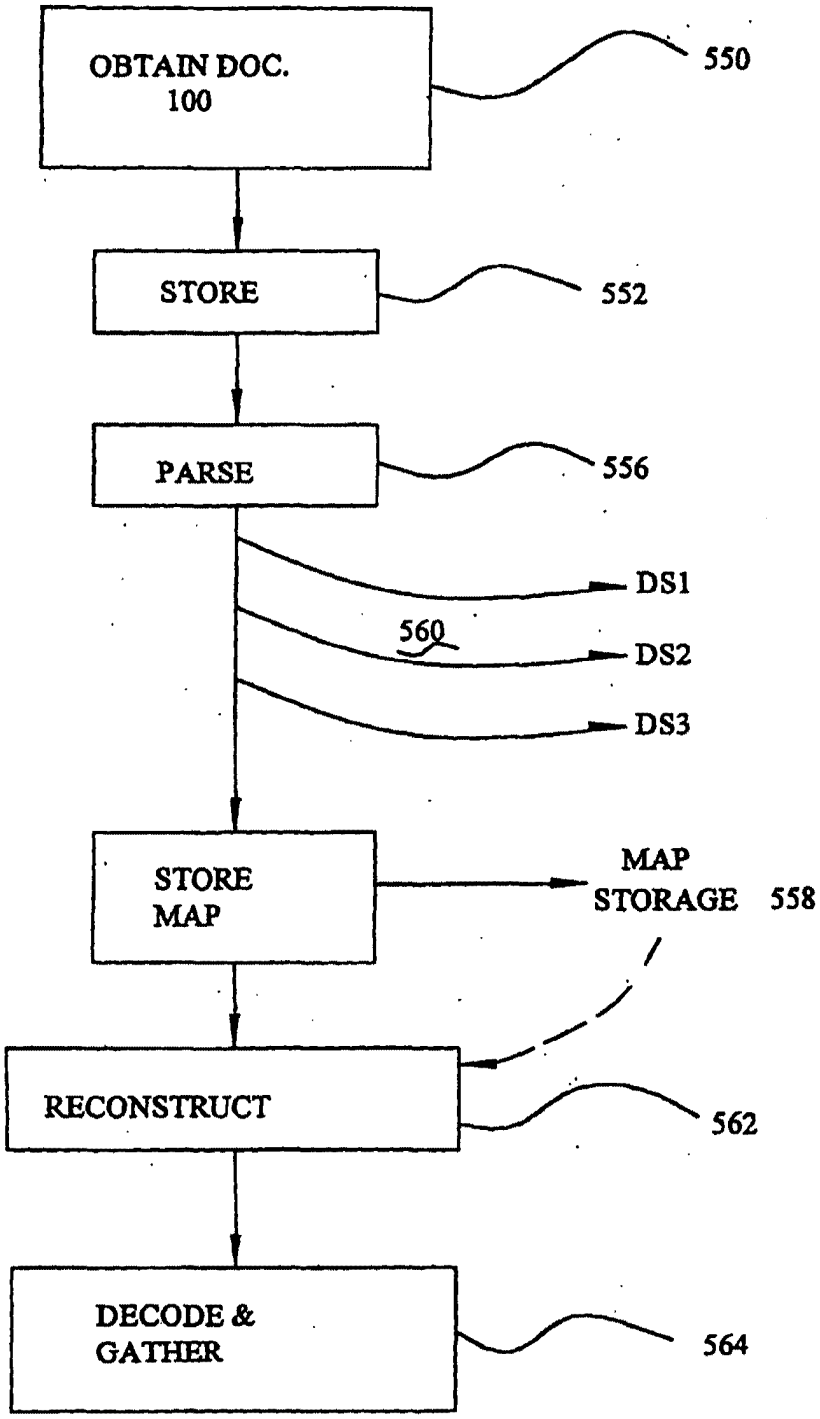

OBTAIN DOC. 100
550
STORE
552
PARSE
556
DS1
560
DS2
DS3
STORE MAP
MAP STORAGE 558
RECONSTRUCT
562
DECODE & GATHER
564

FIG. 11A

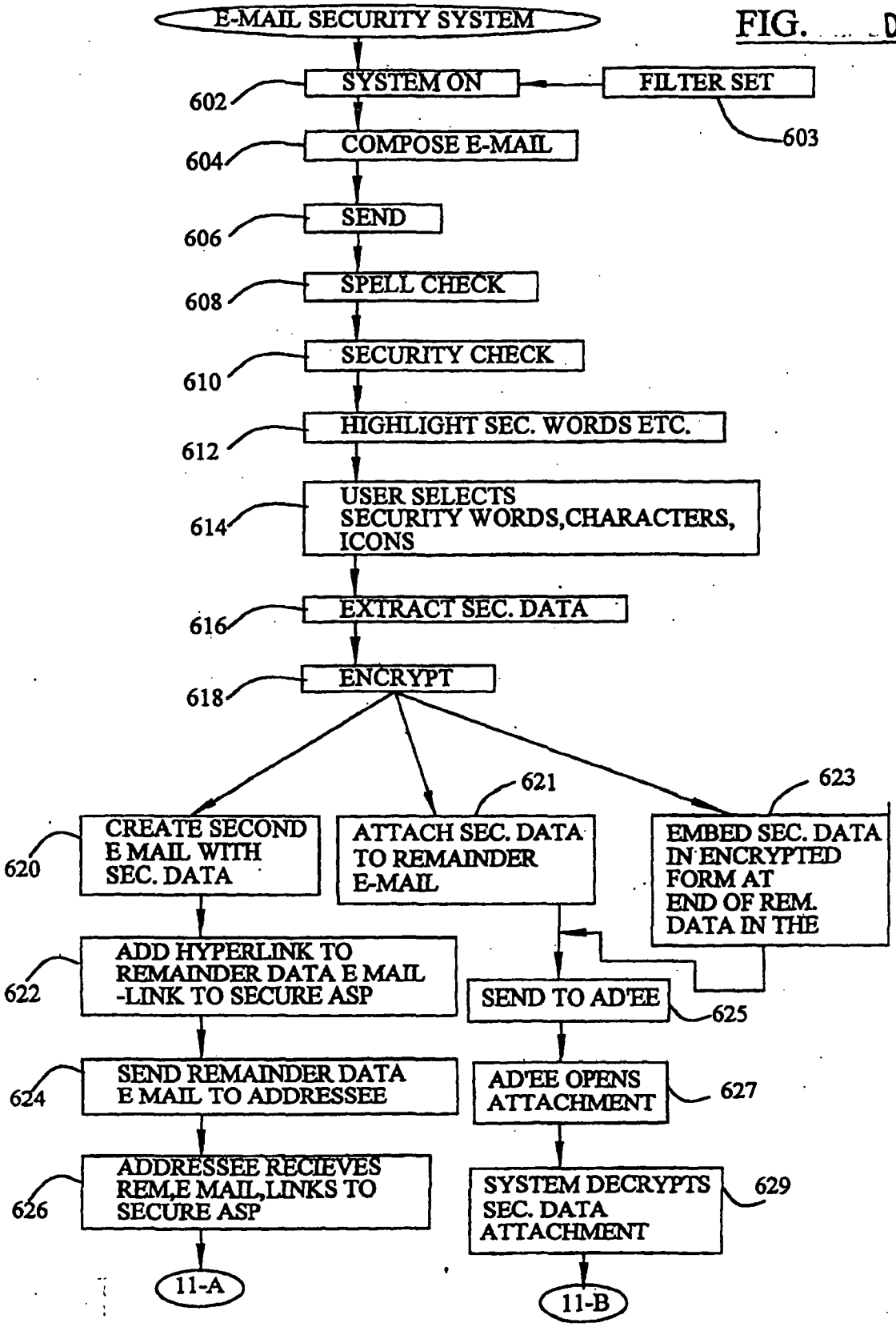

E-MAIL SECURITY SYSTEM
SYSTEM ON
FILTER SET
602
603
COMPOSE E-MAIL
604
SEND
606
SPELL CHECK
608
SECURITY CHECK
610
HIGHLIGHT SEC. WORDS ETC.
612
USER SELECTS SECURITY WORDS,CHARACTERS, ICONS
614
EXTRACT SEC. DATA
616
ENCRYPT
618
621
623
CREATE SECOND E MAIL WITH SEC. DATA
620
ATTACH SEC. DATA TO REMAINDER E-MAIL
EMBED SEC. DATA IN ENCRYPTED FORM AT END OF REM. DATA IN THE
ADD HYPERLINK TO REMAINDER DATA E MAIL -LINK TO SECURE ASP
622
SEND TO AD'EE
625
SEND REMAINDER DATA E MAIL TO ADDRESSEE
624
AD'EE OPENS ATTACHMENT
627
ADDRESSEE RECIEVES REM,E MAIL,LINKS TO SECURE ASP
626
SYSTEM DECRYPTS SEC. DATA ATTACHMENT
629
11-A
11-B

FIG. D-11B

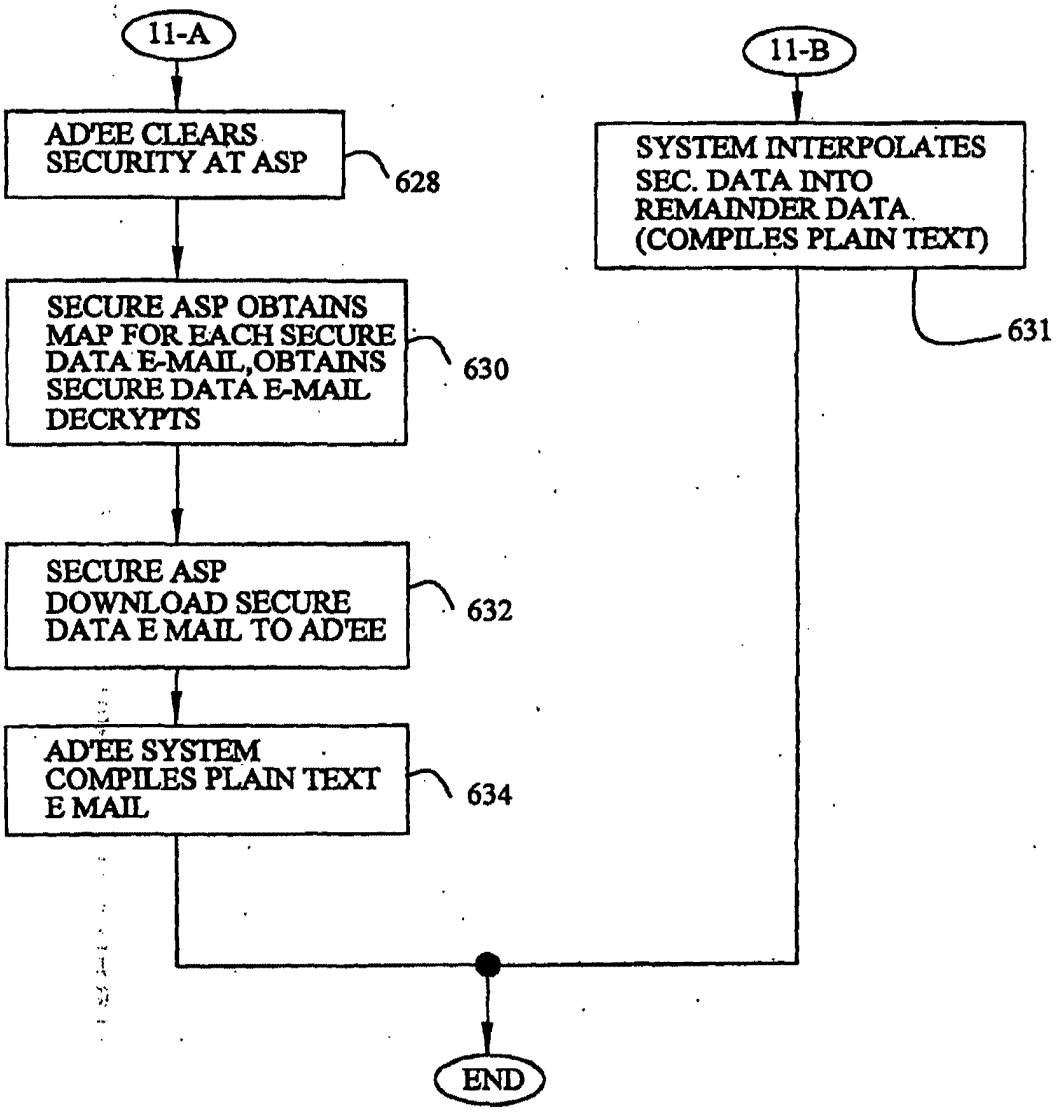

11-A
AD'EE CLEARS SECURITY AT ASP
628
SECURE ASP OBTAINS MAP FOR EACH SECURE DATA E-MAIL, OBTAINS SECURE DATA E-MAIL DECRYPTS
630
SECURE ASP DOWNLOAD SECURE DATA E MAIL TO AD'EE
632
AD'EE SYSTEM COMPILES PLAIN TEXT E MAIL
634
11-B
SYSTEM INTERPOLATES SEC. DATA INTO REMAINDER DATA. (COMPILES PLAIN TEXT)
631
END

FIG. D-12A

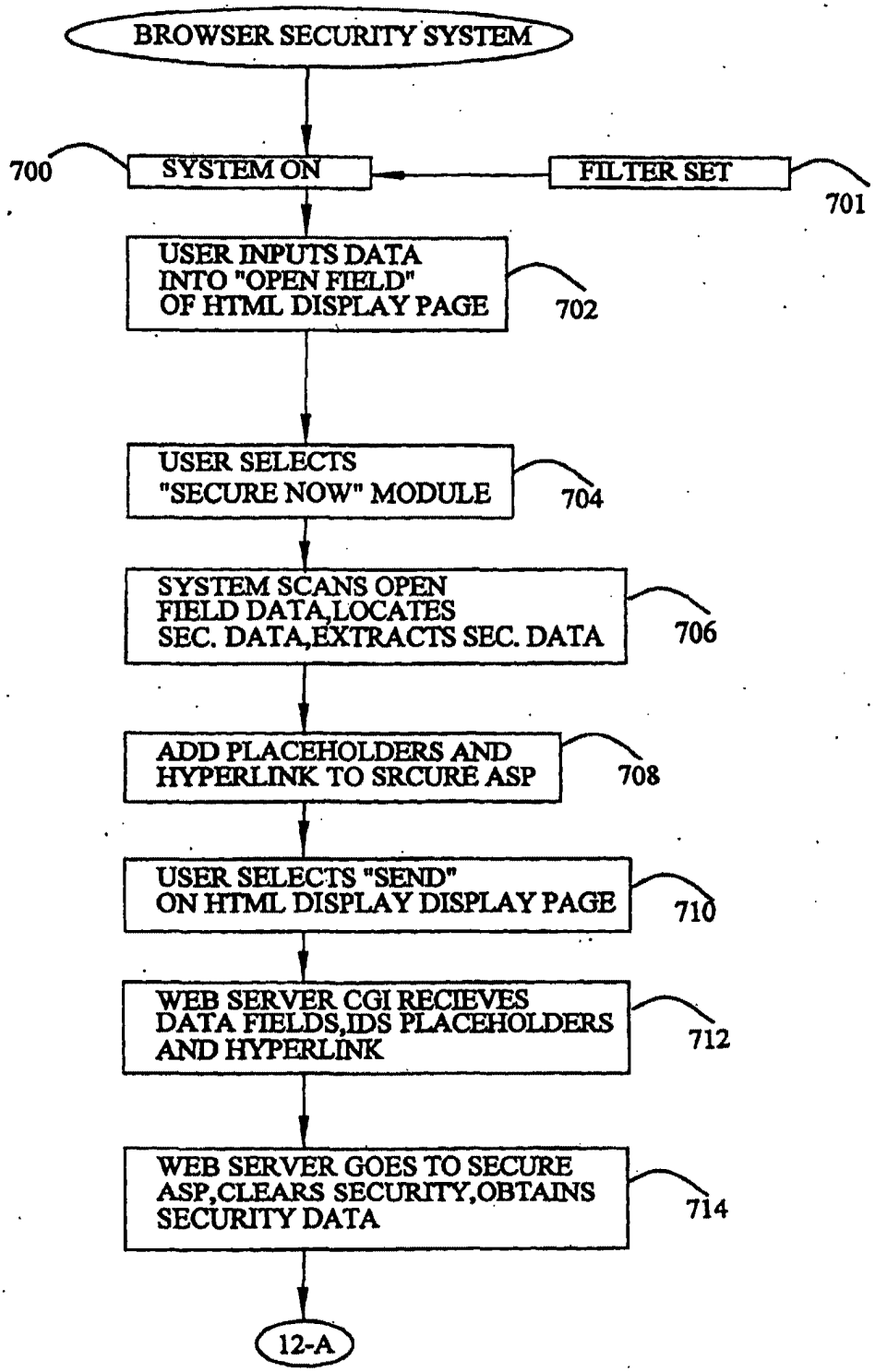

BROWSER SECURITY SYSTEM
700
SYSTEM ON
FILTER SET
701
USER INPUTS DATA INTO "OPEN FIELD" OF HTML DISPLAY PAGE
702
USER SELECTS "SECURE NOW" MODULE
704
SYSTEM SCANS OPEN FIELD DATA,LOCATES SEC. DATA,EXTRACTS SEC. DATA
706
ADD PLACEHOLDERS AND HYPERLINK TO SRCURE ASP
708
USER SELECTS "SEND" ON HTML DISPLAY DISPLAY PAGE
710
WEB SERVER CGI RECIEVES DATA FIELDS,IDS PLACEHOLDERS AND HYPERLINK
712
WEB SERVER GOES TO SECURE ASP,CLEARS SECURITY,OBTAINS SECURITY DATA
714
12-A

FIG. D-12B

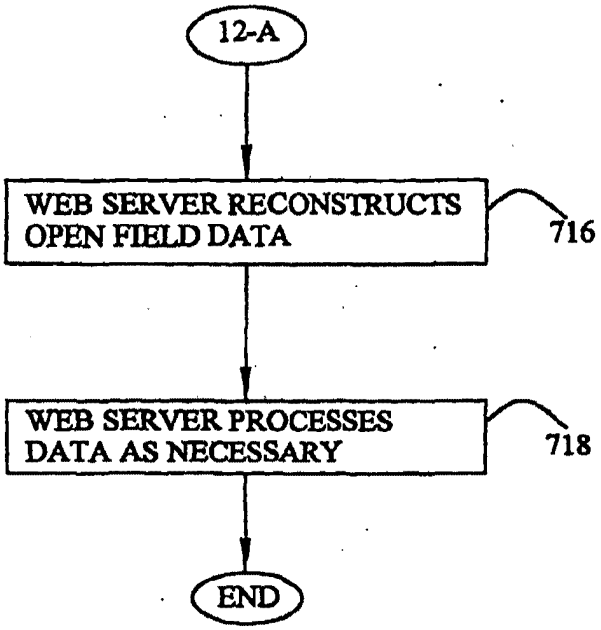
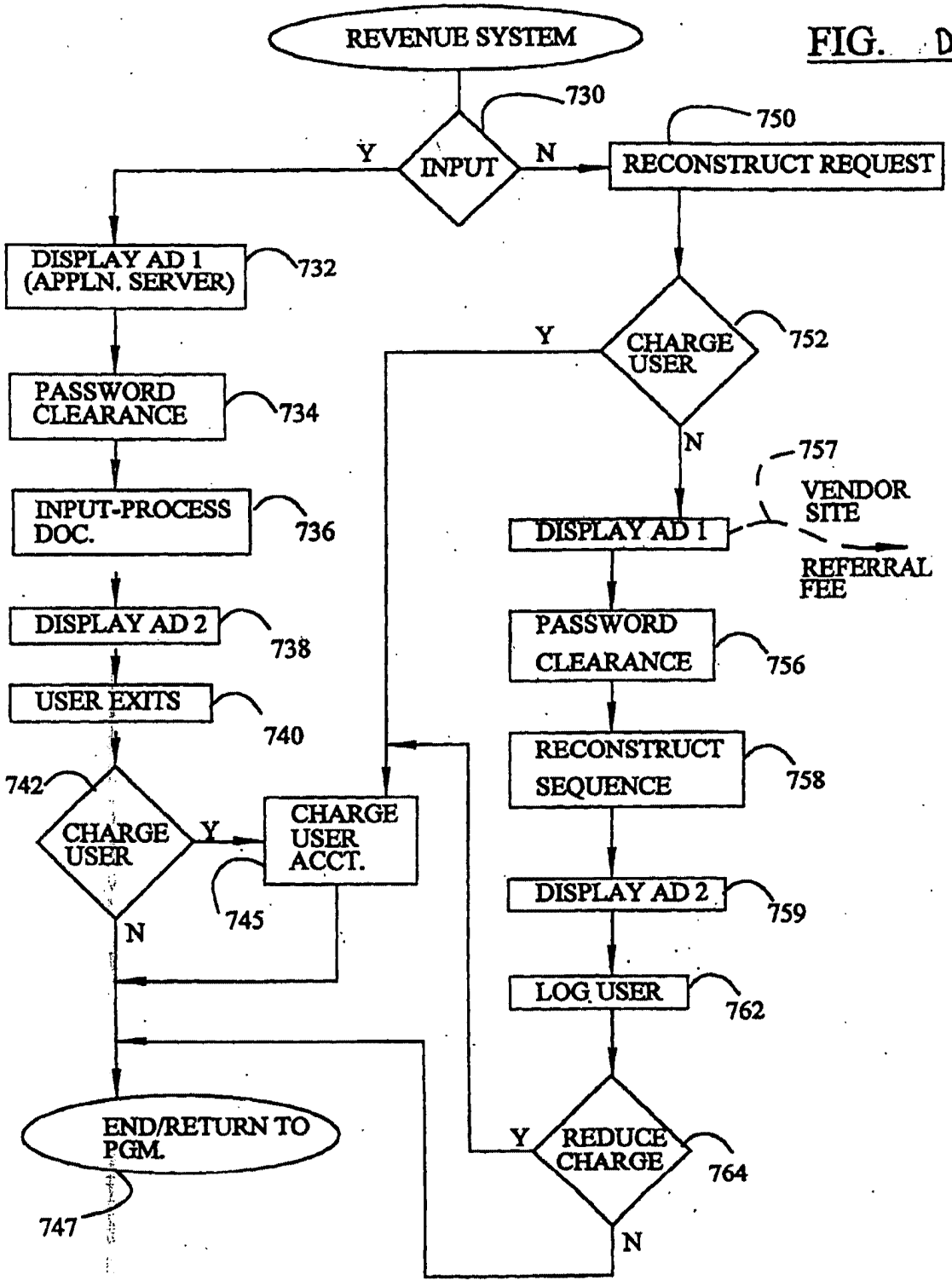

12-A
WEB SERVER RECONSTRUCTS OPEN FIELD DATA
716
WEB SERVER PROCESSES DATA AS NECESSARY
718
END

FIG. D-13


REVENUE SYSTEM
730
INPUT
Y
N
750
RECONSTRUCT REQUEST
DISPLAY AD 1 (APPLN. SERVER)
732
PASSWORD CLEARANCE
734
INPUT-PROCESS DOC.
736
DISPLAY AD 2
738
USER EXITS
740
742
CHARGE USER
Y
CHARGE USER ACCT.
745
N
747
END/RETURN TO PGM.
Y
CHARGE USER
752
N
757
VENDOR SITE
DISPLAY AD 1
REFERRAL FEE
PASSWORD CLEARANCE
756
RECONSTRUCT SEQUENCE
758
DISPLAY AD 2
759
LOG USER
762
Y
REDUCE CHARGE
764
N

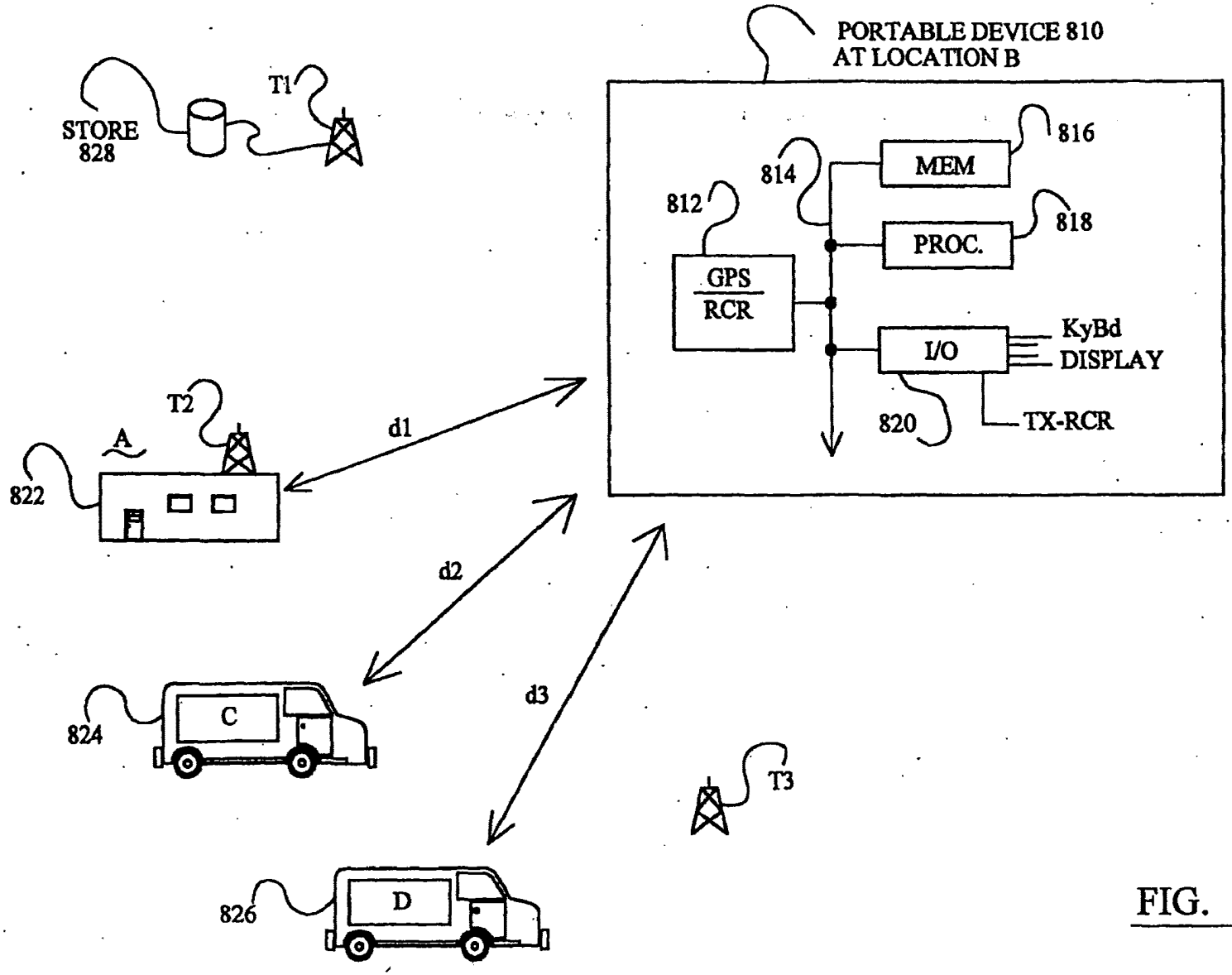

PORTABLE DEVICE 810
AT LOCATION B
STORE
828
T1
816
MEM
814
812
818
PROC.
GPS
RCR
KyBd
DISPLAY
I/O
820
TX-RCR
T2
A
d1
822
d2
d3
C
824
T3
D
826

FIG. D-14

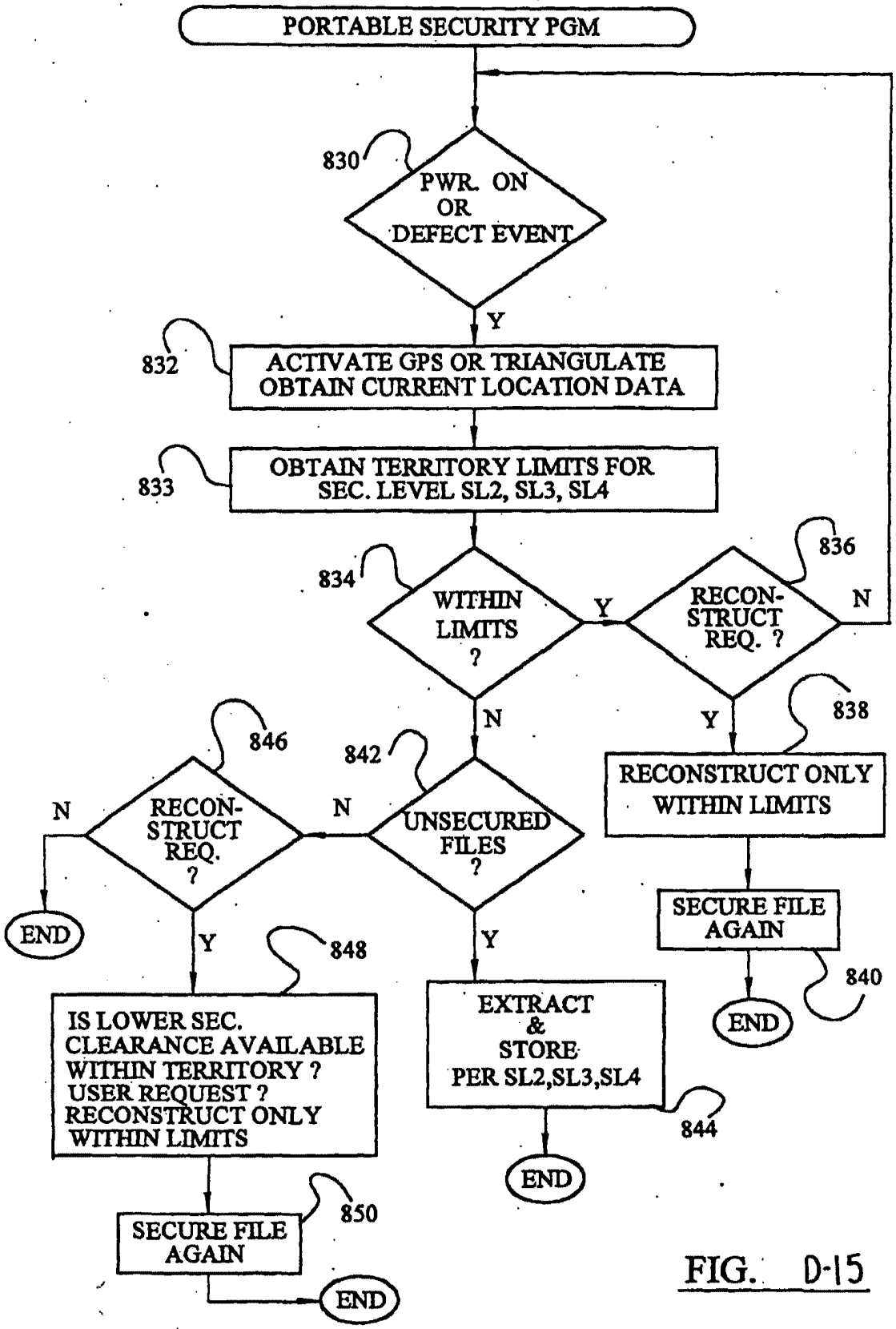

PORTABLE SECURITY PGM
830
PWR. ON OR DEFECT EVENT
Y
832
ACTIVATE GPS OR TRIANGULATE OBTAIN CURRENT LOCATION DATA
833
OBTAIN TERRITORY LIMITS FOR SEC. LEVEL SL2, SL3, SL4
834
WITHIN LIMITS ?
Y
836
RECON-STRUCT REQ. ?
N
N
Y
838
RECONSTRUCT ONLY WITHIN LIMITS
SECURE FILE AGAIN
840
END
842
UNSECURED FILES ?
N
Y
846
RECON-STRUCT REQ. ?
N
END
Y
848
IS LOWER SEC. CLEARANCE AVAILABLE WITHIN TERRITORY ? USER REQUEST ? RECONSTRUCT ONLY WITHIN LIMITS
EXTRACT & STORE PER SL2,SL3,SL4
844
END
850
SECURE FILE AGAIN
END

FIG. D-15

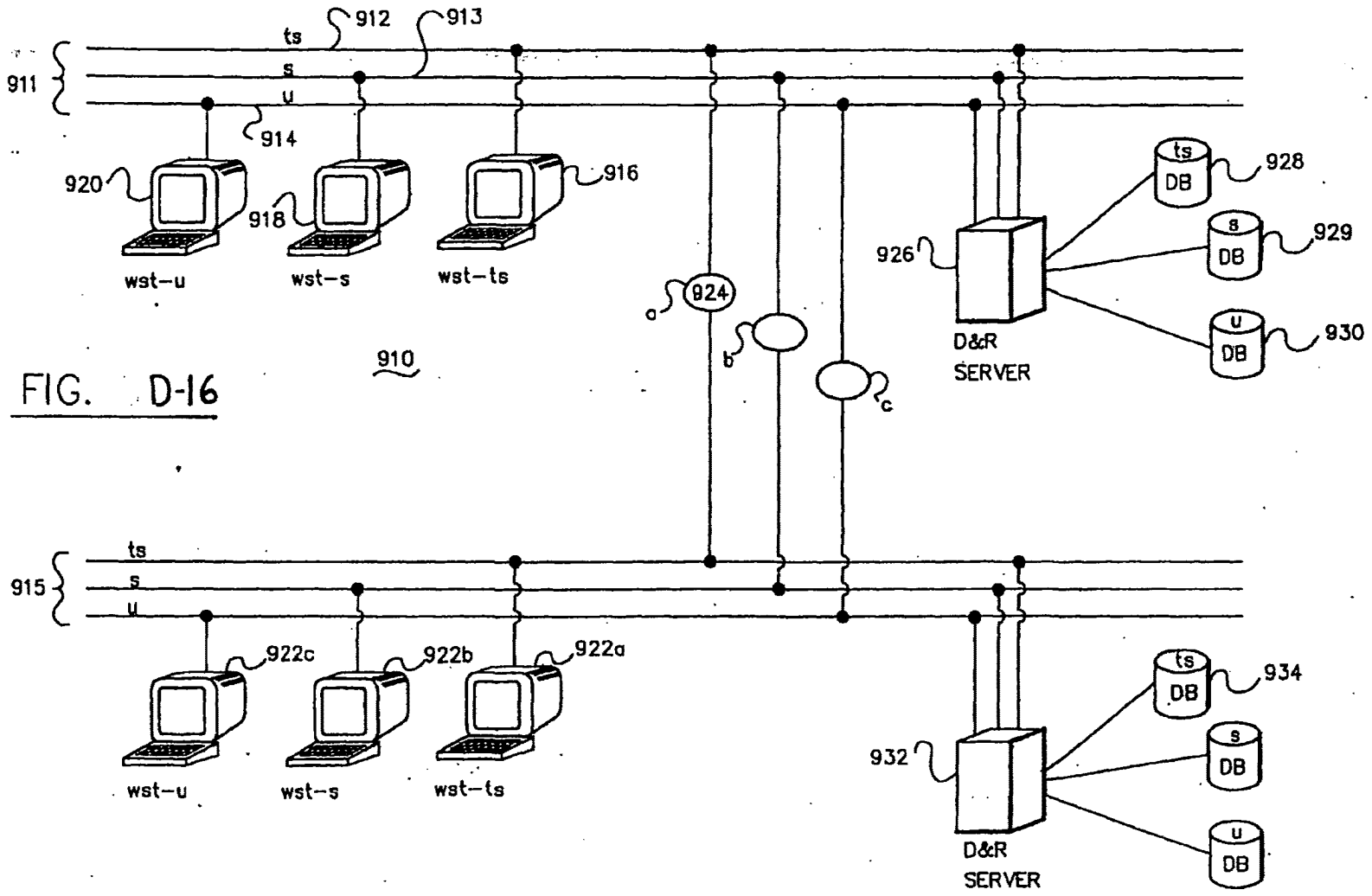

911
ts
s
u
912
913
914
920
918
916
wst-u
wst-s
wst-ts
924
a
b
c
910
926
D&R
SERVER
ts
DB
928
s
DB
929
u
DB
930
915
ts
s
u
922c
922b
922a
wst-u
wst-s
wst-ts
932
D&R
SERVER
ts
DB
934
s
DB
u
DB

FIG. D-16

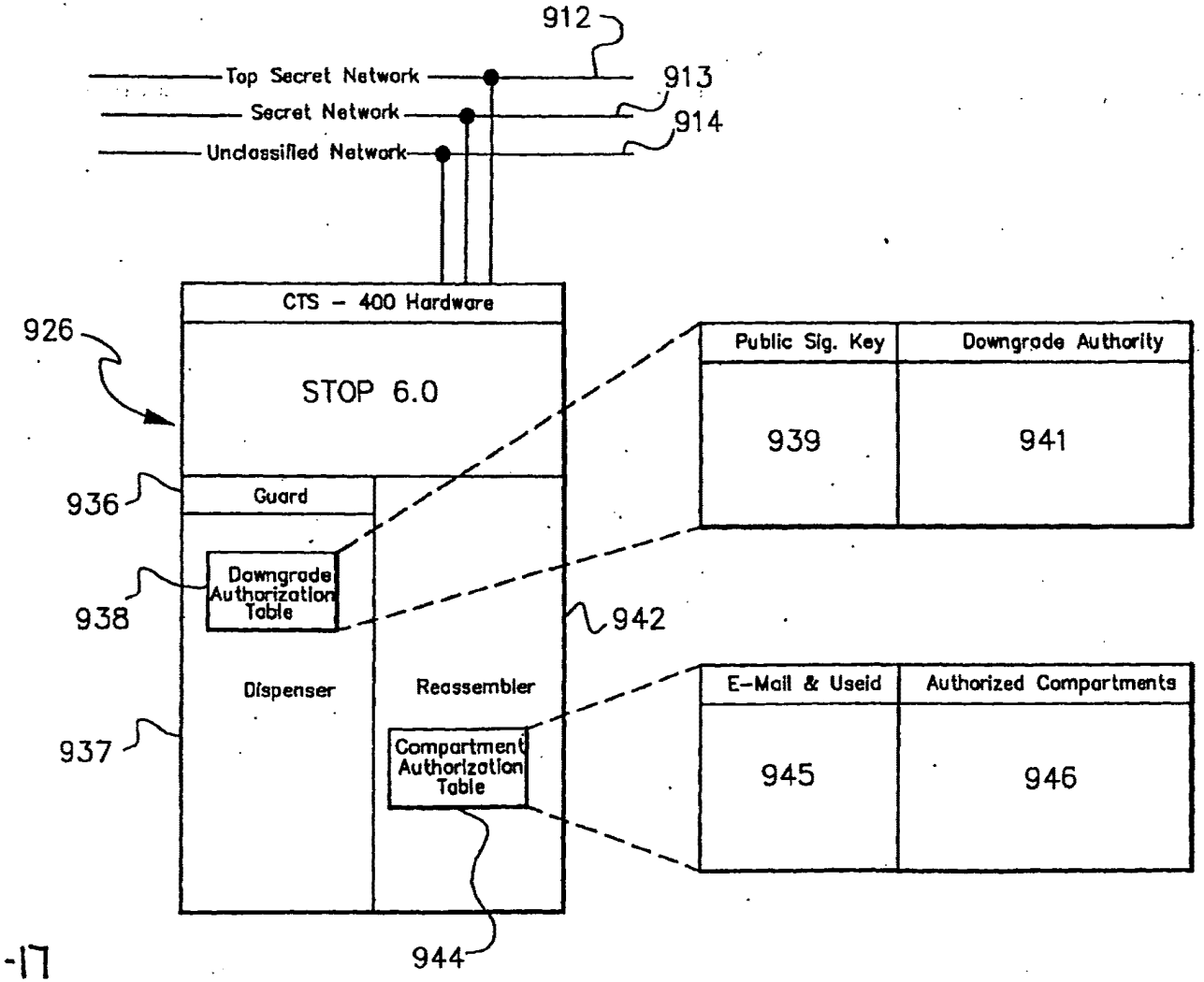

912
Top Secret Network
913
Secret Network
914
Unclassified Network
CTS – 400 Hardware
926
STOP 6.0
936
Guard
938
Downgrade Authorization Table
942
Dispenser
Reassembler
937
Compartment Authorization Table
944
Public Sig. Key
Downgrade Authority
939
941
E-Mail & Useid
Authorized Compartments
945
946

FIG. D-17

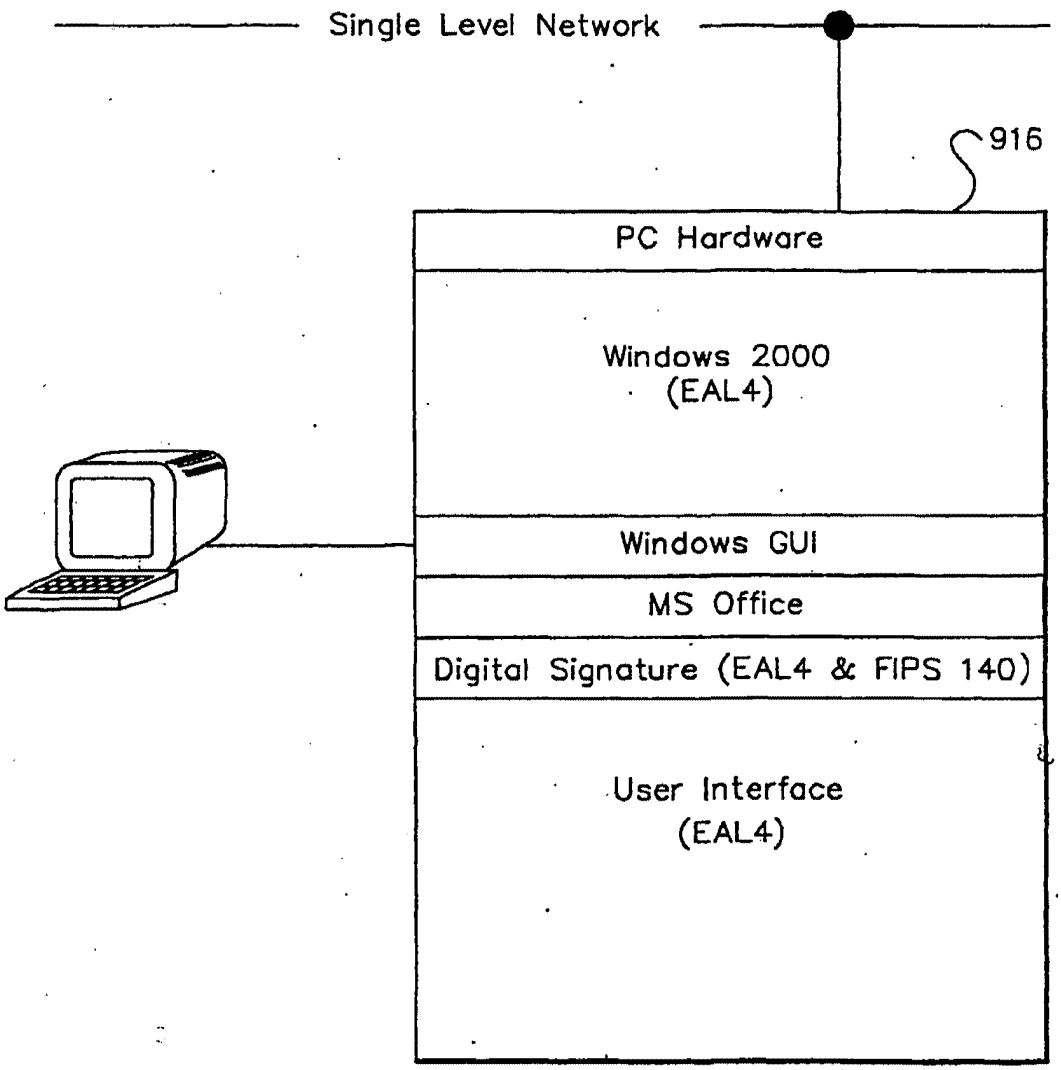

Single Level Network
916
PC Hardware
Windows 2000
(EAL4)
Windows GUI
MS Office
Digital Signature (EAL4 & FIPS 140)
User Interface
(EAL4)

FIG. D-18

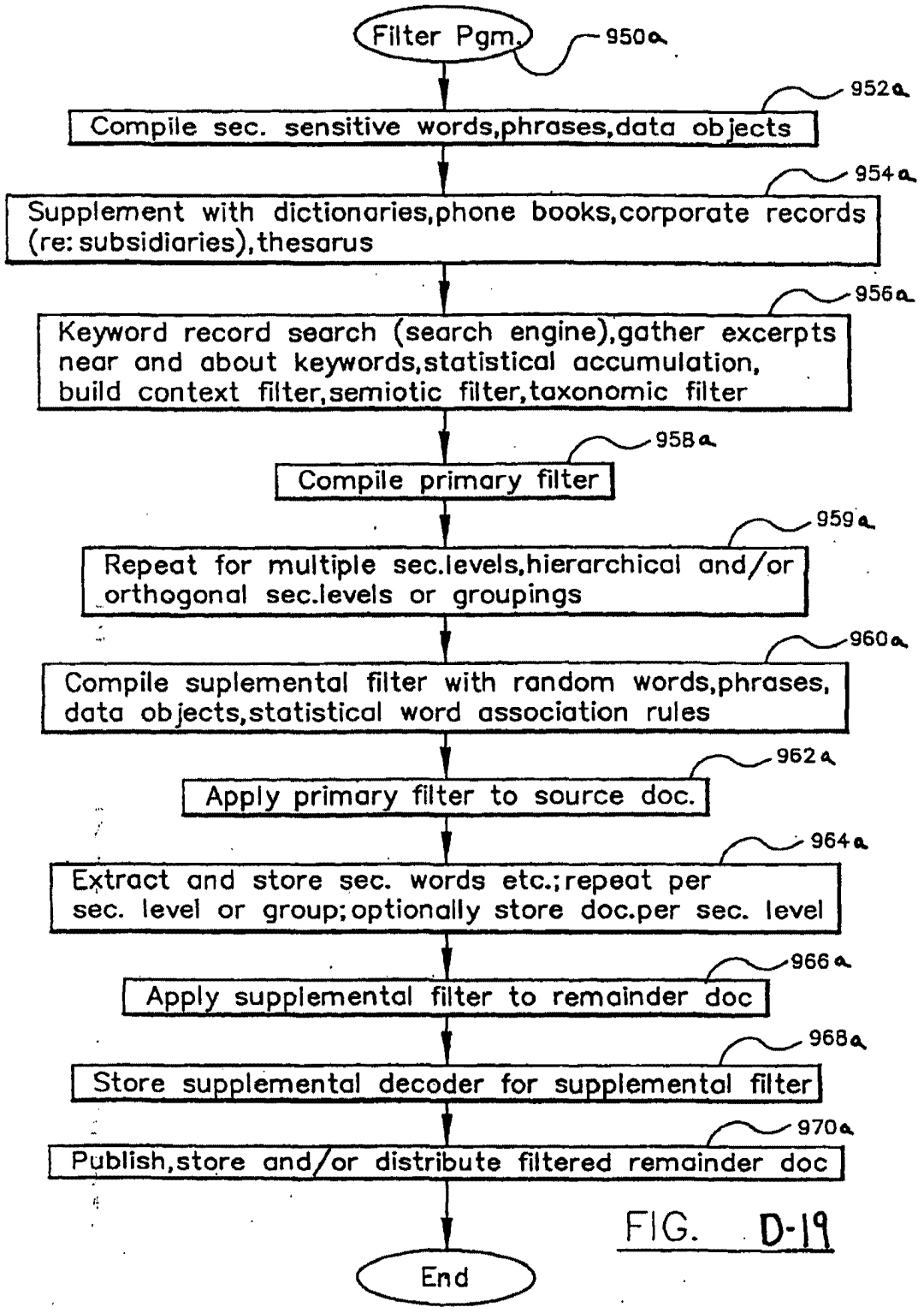

Filter Pgm.
950a
952a
Compile sec. sensitive words,phrases,data objects
954a
Supplement with dictionaries,phone books,corporate records (re: subsidiaries),thesarus
956a
Keyword record search (search engine),gather excerpts near and about keywords,statistical accumulation, build context filter,semiotic filter,taxonomic filter
958a
Compile primary filter
959a
Repeat for multiple sec.levels,hierarchical and/or orthogonal sec.levels or groupings
960a
Compile suplemental filter with random words,phrases, data objects,statistical word association rules
962a
Apply primary filter to source doc.
964a
Extract and store sec. words etc.;repeat per sec. level or group;optionally store doc.per sec. level
966a
Apply supplemental filter to remainder doc
968a
Store supplemental decoder for supplemental filter
970a
Publish,store and/or distribute filtered remainder doc
End

FIG. D-19

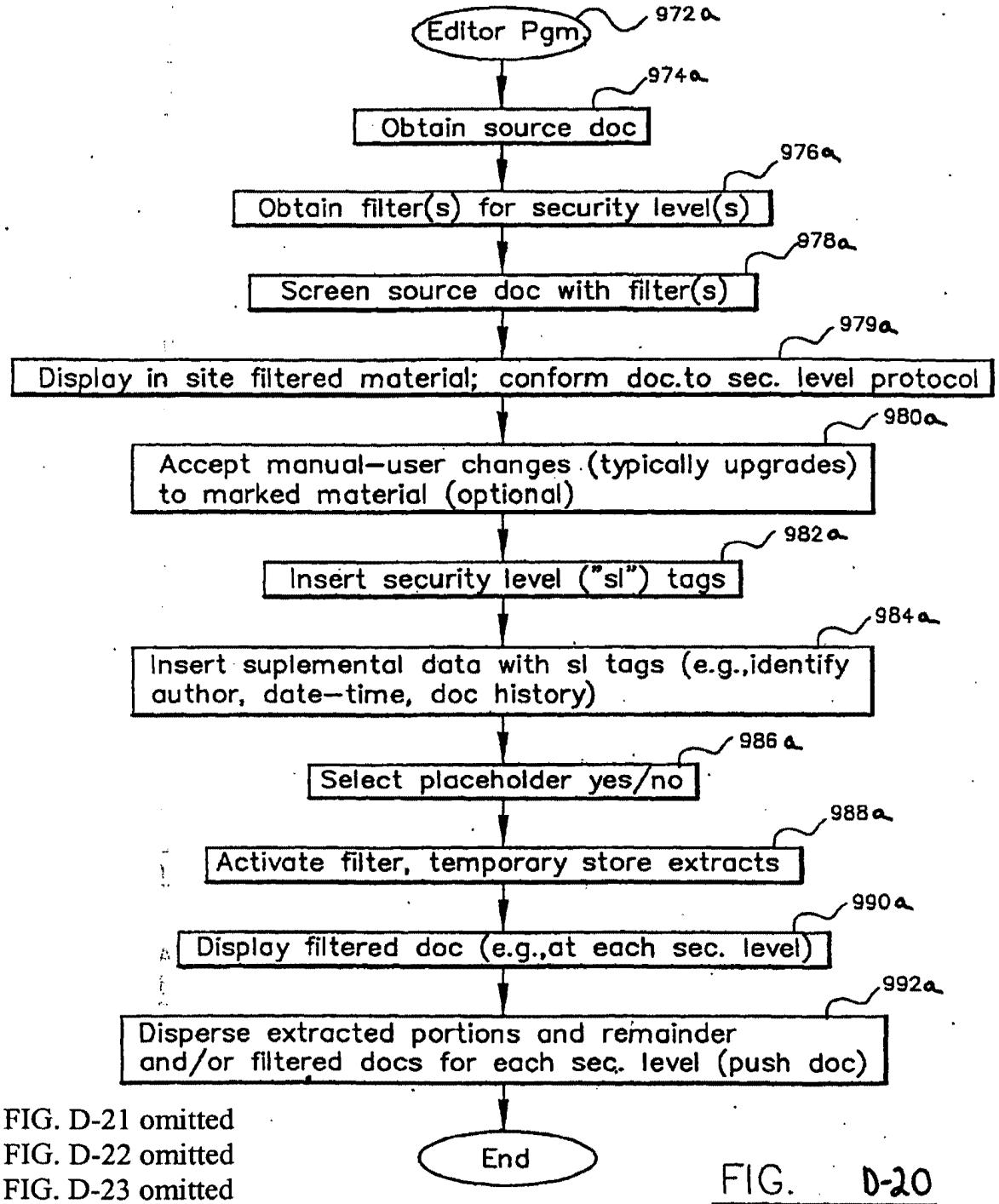


FIG. D-20

FIG. D-21 omitted
FIG. D-22 omitted
FIG. D-23 omitted

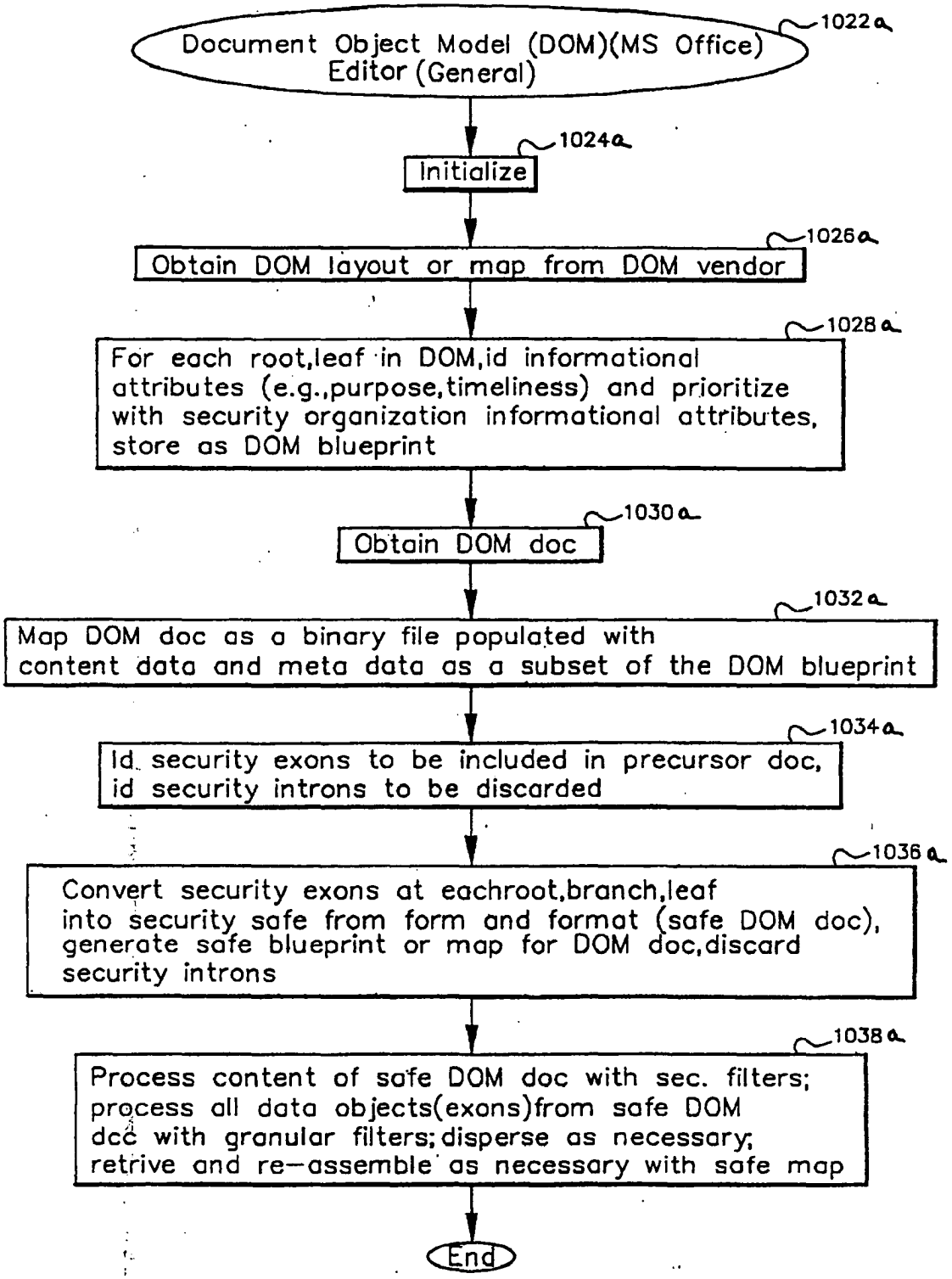
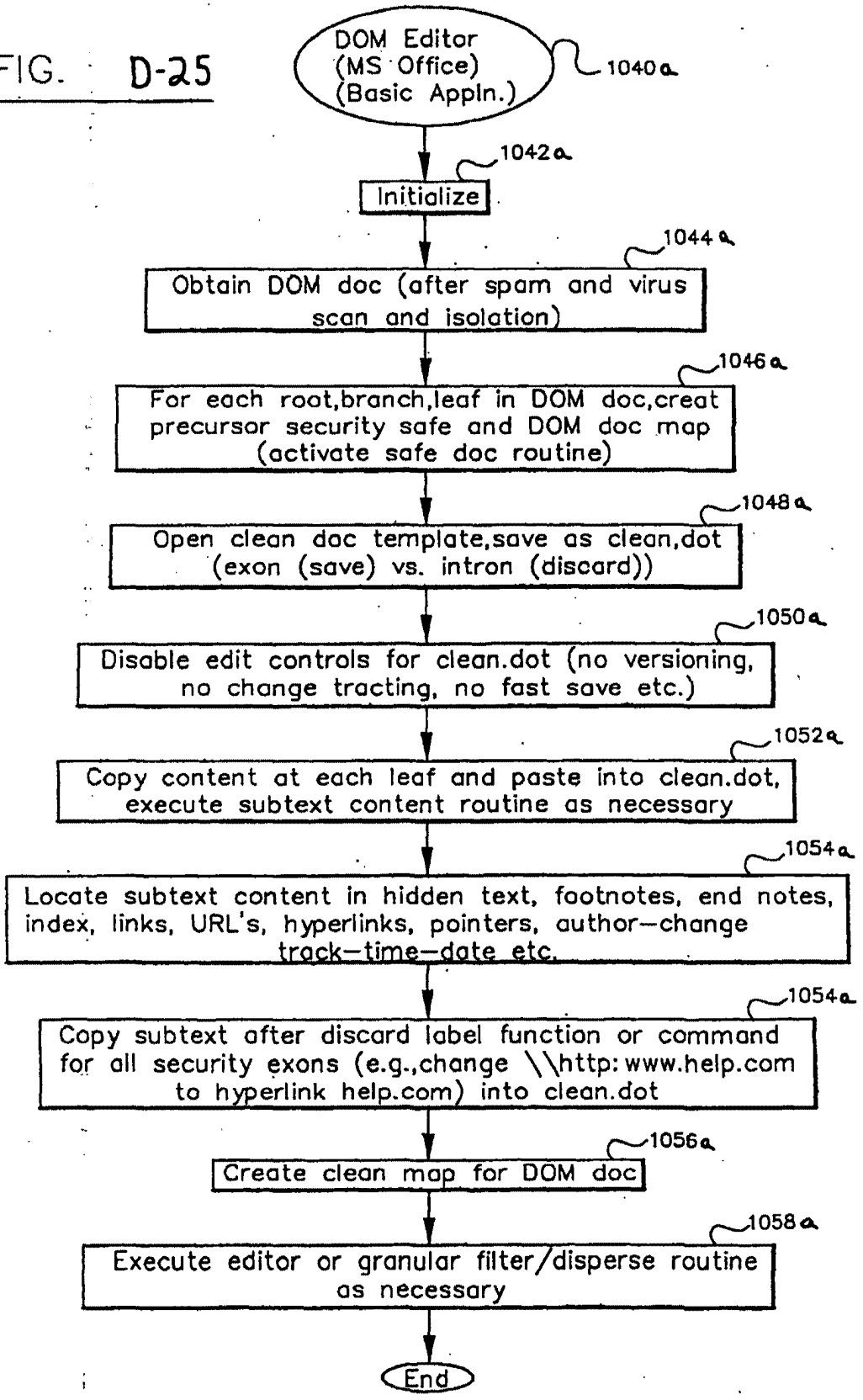

Document Object Model (DOM)(MS Office) Editor (General)
1022a
Initialize
1024a
Obtain DOM layout or map from DOM vendor
1026a
For each root,leaf in DOM,id informational attributes (e.g.,purpose,timeliness) and prioritize with security organization informational attributes, store as DOM blueprint
1028a
Obtain DOM doc
1030a
Map DOM doc as a binary file populated with content data and meta data as a subset of the DOM blueprint
1032a
Id security exons to be included in precursor doc, id security introns to be discarded
1034a
Convert security exons at eachroot,branch,leaf into security safe from form and format (safe DOM doc), generate safe blueprint or map for DOM doc,discard security introns
1036a
Process content of safe DOM doc with sec. filters; process all data objects(exons)from safe DOM doc with granular filters; disperse as necessary; retrive and re-assemble as necessary with safe map
1038a
End

FIG. D-24

FIG. D-25


DOM Editor (MS Office) (Basic Appln.)
1040a
1042a
Initialize
1044a
Obtain DOM doc (after spam and virus scan and isolation)
1046a
For each root,branch,leaf in DOM doc,creat precursor security safe and DOM doc map (activate safe doc routine)
1048a
Open clean doc template,save as clean,dot (exon (save) vs. intron (discard))
1050a
Disable edit controls for clean.dot (no versioning, no change tracting, no fast save etc.)
1052a
Copy content at each leaf and paste into clean.dot, execute subtext content routine as necessary
1054a
Locate subtext content in hidden text, footnotes, end notes, index, links, URL's, hyperlinks, pointers, author—change track—time—date etc.
1054a
Copy subtext after discard label function or command for all security exons (e.g.,change \\http: www.help.com to hyperlink help.com) into clean.dot
1056a
Create clean map for DOM doc
1058a
Execute editor or granular filter/disperse routine as necessary
End

Best Available Copy

Digital Information Infrastructure and Method

[0001] This is a continuation-in-part patent application based upon and claiming the benefit of Serial No. 11/746,440 filed May 9, 2007 which was regular patent application based upon provisional patent application serial no. 60/883,522 filed January 5, 2007, the contents of both applications are incorporated herein by reference thereto.

A.0 Introduction

[0002] The present invention relates to information management data processing modules and computer software system and method enabling protection of confidential information, identification of such information, identification of select content, classification for the same, and security system for the same. Businesses and enterprises use information in many different types of forms and format. Structured information is typically data found in database or spreadsheet formats or stored in self-defining formats wherein the information is managed by a software application. A further example of self-defining format is XML documents in format schematics. Some information is "semi-structured" wherein the information has certain defined fields such as email and the structured data fields from, to, cc, bcc and subject, but the body of the email message is unstructured. Unstructured information includes documents generated by end users with the assistance of applications program such as Microsoft Office, word processing programs, image processing programs (Adobe), files downloaded from the Internet, audio and video files etc. Enterprises focus their attention on mission critical applications and data relative thereto which is often located in structured databases or other structured data formats. However, a large majority of information is contained in unstructured or semi-structured formats. There are significant risks involved in handling and locating confidential information in these information files as well as identifying and handling intellectual property data in those files.

Best Available Copy

[0003] The content of these information files are important from a legal standpoint, a data security standpoint, and a government regulation or compliance standpoint. Although the content of the information files may be searched and indexed, the resulting information must then be classified. The enterprise must establish policies which dictate how the information files are handled in light of legal, security and compliance risk factors. Further, enterprise policies regarding information or document retention and immutability (the degree and extent that a particular document or information file may be changed) must be established and enforced.

[0004] Simply indexing the content of an information file is not adequate. Semantic analysis, key word tagging and classification categorization (taxonomic analysis) should be conducted. Information tools and computer programs and methods should enable the users (an enterprise) to analyze, categorize and actively manage unstructured information in the enterprise. Theft or accidental loss of customers' information or employees' private information is no longer acceptable. Laws mandate disclosure of such data loses. Further, the public is acutely aware of identity theft and privacy issues. Also, theft or misuse of intellectual property and other confidential business information is a mission critical aspect of many enterprises. To complicate matters, businesses and enterprises do not operate closed informational systems but often times invite partners, customers, vendors, contractors and other third parties to participate in the enterprise informational system. Therefore, most enterprises operate an open ecosystem informational structure. Employees of the enterprise in addition to its partners, customers, vendors and others participate in the production of information and the use/and consumption of information.

[0005] In the past, most commentators believed that a breach of confidentiality and misuse of data was a result of an outsider "hacker" getting access to enterprise information by penetrating a

Best Available Copy

firewall or other security system. However, there is a greater risk to the enterprise that rogue insider will take the information and misuse that information. Further, trusted employees constantly copy files onto laptops, PDAs (personal data assistance), USB thumb drives (flash memory), CDs and other removable media in order to fulfill the tasks assigned by the enterprise. This portable data can be easily misplaced or stolen. Further, the ease of electronic transfer of documents to others exposes the information contained therein to mis-directed addresses, as well as unauthorized theft and misuse.

[0006] Therefore, enterprises and organizations must manage the access to the information and use of the information and monitor such access and use and audit (log) these activities. A system of information rights management should concern the following aspects: (1) control who can open, print or edit a document or information file; (2) prevent or limit distribution of the information file; (3) enable the enterprise to change permissions on an information file per enterprise policy or changed conditions; and (4) maintain an audit trail of such activity.

[0007] The information infrastructure tools also must accommodate the changeable value of information over the life cycle of the information file. Although the information and content of the file does not change, the value of the information changes due to contextual events and external environmental conditions. External events which change the value of information and information file can be summarized as follows: (a) creation and modification and use of the information in the information file as it is processed by the enterprise; (b) completion of the project or event associated with the information file; (c) government regulations imposed on the information; (d) legal actions associated with the content of the information file and financial audits related to such information; (e) distributing the information in the information file to individuals with different access or security clearances or rights; (f) distribution of the information, both physically and electronically, to others who are not in the

Best Available Copy

enterprise; (g) the types and operation of storage devices and the media which retain the information files; (h) the discovery of new uses for the information content in the file; (i) new software applications which replace obsolete applications; (k) personnel changes within the enterprise and associated changes and permissions to access, control and modify the information file; (l) expansion and contraction of the enterprise itself (organizational changes) and the changeable nature of business relationships with partners, customers and suppliers.

[0008] In view of the foregoing, it is necessary to have an information classification system or engine which automatically categorizes information in unstructured information files and labels the same and permits the enterprise to establish a policy for that information and implement the policy recommendation with specific actions to ensure that the information is properly handled, distributed, retained, deleted (document retention) and otherwise managed.

Objects of the Invention

[0009] It is an object of the present invention to provide a system operator or CIO with data processing tools to manage and organize data processed by an enterprise.

[0010] It is a further object of the present invention to provide tools for securing secret or security sensitive sec-con data in the enterprise computer system and to locate, identify and secure select content SC which may be of interest or importance to the enterprise.

[0011] It is an additional object of the present invention to employ a dynamic, adaptive filter to enhance select content (SC) collection, to employ classification systems to organize such SC, and, in some cases, to do so continuously.

[0012] It is a further object of the present invention to assist in data processing or manipulation including processes such as coping, data extraction, data archiving, data distribution, data destruction

Best Available Copy

(a document retention process), inter-enterprise and intra-enterprise data distribution, data access request processing, and data distribution denials.

Summary

[0013] The present invention is a method of organizing and processing data in a distributed computing system. The invention is also implemented as a computer program on a computer medium and as a distributed computer system. Software modules can be configured as hardware. The method and system organizes select content which is important to an enterprise operating said distributed computing system. The select content is represented by one or more predetermined words, characters, images, data elements or data objects. The computing system has a plurality of select content data stores for respective ones of a plurality of enterprise designated categorical filters which include content-based filters, contextual filters and taxonomic classification filters, all operatively coupled over a communications network. A data input is processed through at least one activated categorical filter to obtain select content, and contextually associated select content and taxonomically associated select content as aggregated select content. The aggregated select content is stored in the corresponding select content data store. A data process from the group of data processes including a copy process, a data extract process, a data archive process, a data distribution process and a data destruction process is associated with the activated categorical filter and the method and system applies the associated data process to a further data input based upon a result of that further data being processed by the activated categorical filter utilizing the aggregated select content data. In effect, the system and process translates the sec-con or SC data and then stores the same in certain locations or secure stores. In this manner, a transformation of the data is achieved. By transposing the sec-con or SC data, higher levels of organization are realized or higher levels of security are implemented.

Best Available Copy

[0014] The enterprise designated filters screen data for enterprise policies such as a level of service policy, customer privacy policy, supplier privacy policy, enterprise human resource privacy policy, financial data handling policy, public company reporting policy, health care regulation policy, technical innovation policy, competitive reporting policy and document or data retention policy.

[0015] In another embodiment, the method and system extracts security sensitive content from a data input to obtain extracted security sensitive data for a corresponding security level and remainder data. This extracted security sensitive data is stored in a respective security sensitive extract store. Remainder data is stored in the distributed computer system. Distribution of the data input with respect to said distributed computing system is limited based upon throughput from at least one of the enterprise designated filters. The controlled release of corresponding extracted security sensitive data from the respective extract stores with the associated security clearances for corresponding security levels is permitted by the system. In this manner, security sensitive content is separately stored apart form the select content, which is stored in select content data stores.

Table of Contents

AAA.1.0.0 DigitalDoors Capabilities

AAA.1.1.0 Architecture Specifications

AAA.1.2.0 Scalability

AAA.1.3.0 Product and Process Overview

AAA.1.4.0 DigitalDoors implementation options:

AAA.1.5.0 The Key for Net-Centric security

AAA.1.6.0 Active defense and counterattacks versus common passive defense

AAA.1.7.0 Technical Basis for DigitalDoors Technology

Best Available Copy

AAA.1.8.0 Granular data control

AAA.1.9.0 Collaboration will consist of:

AAA.1.10.0 Granular tagging with metadata

AAA.1.11.0 Coalition And Multinational Cross Domain Collaboration

AAA.1.12.0 Incorporation into a trusted OS AAA.1.13.0 Survivable infrastructure

AAA.1.14.0 Net Centric capabilities

AAA.1.15.0 DigitalDoors Services overview

AAA.1.16.0 DigitalDoors implementation options:

AAA.1.17.0 defenses against outs

AAA.1.18.0 collaboration elements:

AAA.1.19.0 DigitalDoors cornerstones

AAA.1.20.0 Granular content control

AAA.1.20.1 Cross domain

AAA.1.21.0 Security services

AAA.1.22.0 solutions to vulnerabilities

AAA.1.23.0 Replication for perpetual data availability

AAA.1.24.0 Solving net centric needs

AAA.1.25.0 Secured Discovery and publishing services

AAA.1.26.0 Publishing;

AAA.1.27.0 Meta data Posting

AAA.1.28.0 Secured Mediation services for information sharing

AAA.1.29.0 Other services

Best Available Copy

AAA.1.30.0 Secured perpetual storage services

AAA.1.30.1 Meta data posting

AAA.1.31.0 Secured enterprise management services

AAA.1.32.0 Access enforcement:

AAA.1.33.0 Real time granular audit of all critical content in network

AAA.1.34.0 Applications for operating an assured environment services-

AAA.1.35.0 Processing

AAA.1.36.0 Providing scalability, reliability, resilience, fault tolerance, speed

AAA.1.37.0 Enabling generic and standardized data distribution services

AAA.1.38.0 Storage

AAA.1.39.0 Metrics

AAA.1.40.0 Technical specifications, operating parameters, and network design specifications

AAA.1.41.0 DigitalDoors Operating Parameters:

AAA.1.42.0 DigitalDoors Network Design Specifications:

AA.1.0.0 Background: Formlessness for Security

AA.1.1.0 Introduction

AA.1.2.0 Definition

AA.1.3.0 Formlessness

AA.1.3.1 How DigitalDoors Applies Formlessness

AA.1.4.0 Formlessness as Security for Maps.

AA.1.4.1 History and Concept

Best Available Copy

AA.1.5.0 Security through: dispersion, formlessness, reversible chaos

AA.1.6.0 Formlessness to secure maps

AA.1.7.0 Formlessness by breaking context - hiding is not enough

AA.1.8.0 Differentiation between Obscurity and Formlessness

AA.1.8.1 Obscurity

AA.1.8.2 Formlessness without context

AA.1.8.3 Proverbial versus Actual Context Change

AA.1.8.4 Obscurity and Formless Differentiation

AA.1.8.5 Recovering from Obscurity and Formlessness

AA.1.8.6 Formlessness through Microperimeterization

AA.1.8.7 Formlessness the difference between Dispersion and Hiding

AA.1.1.8 Dispersion as Asymmetric warfare

AA.1.9.0 Overview of Formlessness:

AA.1.9.1 The DigitalDoors Formlessness Doctrine

AA.1.9.2 Establishing Formlessness:

AA.1.9.3 Defending intellectual property and trade secrets:

AA.1.9.4 Defending information in web sites:

AA.1.9.5 Legal defenses with formlessness:

AA.1.9.6 Measurement of the information inventory for risk:

AA.1.9.7 Monetizing the information inventory:

AA.1.9.8 The Global Information war

AA.1.10.0There is a need for a technological paradigm change.

Best Available Copy

AA.1.11.0 DigitalDoors Hierarchical Security Sanitization and Reconstitution

AA.1.12.0 Methodology Overview:

AA.1.13.0 A new IT paradigm

AA.1.14.0 Taking advantage of a major economic imbalance

AA.1.15.0 Monetizing the organization's information assets:

AA.1.16.0 Information disassembly and Reassembly

AA.1.16.1 Disassemblying Information

AA.1.16.2 The reasons for disassembly

AA.1.16.3 Finding Value within Information

AA1.16.4 TheDoctrine behind DigitalDoors

AA.1.16.5 Reconstituting Information

AA.1.16.6 Directory Files

AA.1.16.7 A compound data stream that can be stripped into granular items

AA.1.16.8. Data Unification

AA.1.16.9 Conclusion

AA.1.17.0 A New Information Security Paradigm: Neutralize the Loss of Perimeter

AA.1.18.0 Why security is ignored

AA.1.19.0 Security integration and the lack of attention to it

AA.1.20.0 Security as the future foundation for information

AA.2.0.0Design and implementation

AA.2.1.0 Implementation of Granular data control in data streams

AA.2.2.0 Brief Functional Overview

EX. 1002, Page 72

Best Available Copy

AA.2.2.1 Brief Structural Overview

AA.2.2.2 Remote Storage Overview

AA.2.2.3 Operating Parameters

AA.2.2.4 Network Design Specifications

AA.2.2.5 Operating Parameters

AA.2.2.6 Network Design Specifications

AA.2.2.7 Remote Storage Infrastructure

AA.2.2.8 Brief Structural Overview

AA.2.2.9 Remote Storage Overview

AA.2.3.0 Upfront Processing

AA.2.4.0 Proxy architecture

AA.2.5.0 Reclassification- Reconstitution

AA.2.5.1 ReClassification

AA.2.5.2 Recovery Information

AA.2.5.3 Processing with DeClassified Data

AA.2.5.4 Content Filtering

AA.2.5.5 Channel Transmission Protection

AA.2.6.0 The Recovery map

AA.2.6.1 Map Structure Variants and Granularity

AA.2.7.0 Concurrency in granular data control

AA.2.8.0 DeClassifying and ReClassifying E-Mail

AA.2.8.1 Generation and Distribution

Best Available Copy

AA.2.8.2 Receiving and ReClassification

AA.2.8.3 Forwarding and Redelivery of E-Mail

AA.2.8.4 Content Filtering

AA.2.8.5 Processing E-Mail

AA.2.9.0 Declassifying and ReClassifying while Browsing the web

AA.2.9.1Generation and Distribution:

AA.2.9.2 Receiving and ReClassification

AA.2.10.0 Creating multiple levels of categorization

AA.2.11.0 Out-of-Channel Multitier Tagging

AA.2.12.0 Multitier Hierarchical Overlapping Tag Structure

AA.2.13.0 Dictionary Encoding AA.2.14.0 Declassification and scrubbing of maps

AA.2.14.1 Channel Transmission Protection:

AA.2.15.0 Distribution of MS word

AA.2.16.0 Filters, sieves and capabilities

AA.2.16.1 The DigitalDoors classification metadata can be transitory.

AA.2.16.2 Applicability of Sieves and Filters

AA.2.16.3 Manual, Assisted, an Automated Modes

AA.2.16.4 Dictionaries

AA.2.16.5 My Word List

AA.2.16.6 My Group List

AA.2.16.7 My Category List

AA.2.16.8 My Policy List

Best Available Copy

AA.2.16.9 AutoSEC

AA.2.16.10 OPSEC

AA.2.16.11 COMSEC

AA.2.16.12 Manual Selection

AA.2.16.13 Manual Mouseover Selection

AA.2.16.14 Tearline Creation

AA.2.16.15 Other configurations

AA.2.16.16 Accuracy and Granularity

AA.2.17.0 Automating filters and sieves

AA.2.17.1 Classification Conflict Resolution

AA.2.18.0 Processing sets

AA.2.18.1Collections

AA.2.18.2 Sets

AA.2.19.0 Failure of best of breed security

AA.2.20.0 Integrated Single Files, Attachments, and Links

AA.2.20.1 Stream Outputs

AA.2.21.0 Hierarchical Security Sanitization and Reconstitution

AA.2.22.0 Software implementation

AA.2.23.0 Software operations: Sensitivity Level Hierarchy and tear line

AA.2.23.1 Secure Editor application

AA.2.23.2 DeClassifying and ReClassifying while Browsing the web

AA.2.23.3 Generation and Distribution:

Best Available Copy

AA.2.23.4 Receiving and ReClassification

AA.2.24.0 DeClassification Mapping - Channels mapping

AA.2.24.1 Channel Transmission Protection:

AA.2.25.0 Security by dispersion – the cloud

AA.2.26.0 Semantics dictionaries AA.3.0.0 Privacy and security:

AA.3.1.0 Context: The difference between Security and Privacy

AA.3.2.0 Security and privacy: protecting the crown jewels

AA.3.3.0 Surrogate Keywords as unique identifiers for privacy and security

AA.3.4.0 The Metrics of Privacy: Identification, De-Identification, and Permitted versus Forced Re-Identification

AA.3.4.1 Privacy overview

AA.3.4.2 Data Ownership

AA.3.3.3 Definition of Identity

AA.3.4.4 Uniqueness

AA.3.4.5 Sufficiency

AA.3.4.6 Direct and Indirect Identification

AA.3.4.7 Unique Identification and Subset Identification

AA.3.4.8 Indirect Identity Information Sources

AA.3.4.9 Definition of Privacy

AA.3.4.10 Implementing Privacy Policy

AA.3.4.11 Privacy Measurements

AA.3.4.12.Measuring Privacy

Best Available Copy

AA.3.4.13 Measuring Uniqueness of Identification

AA.3.4.14 Density of Identification

AA.3.4.15 Inference

AA.3.4.16 Conclusion

AA.3.5.0 Personally identifiable information: Meta PII

AA.3.6.0 Categories of Personally Identifiable Information (PII)

AA.3.6.1 Technical Identifiers

AA.3.7.0 Toxic spill of personally identifiable information (PII)

AA.3.8.0 From aggregation, inference, reidentification to Anonymity

AA.39.0 Control personally identifiable information (PII) through granular data control

AA.4.0.0 Classification and categorization:

AA.4.1.0 Multiple Hierarchical Tagging and Data Extraction

AA 4.2.0 Commonality of Data: categorization, a multitier hierarchical overlapping tag Structure

AA.4.3.0 Categorization structure tagging, risk measurement and context

AA.4.4.0 Data and Metadata Granularity for decomposition to break context

AA.4.5.0 In-Channel and Out-of-Channel Tags

AA.4.6.0 Capitalization is content-, context-, and concept-driven

AA.4.6.1 Categorization Specialization

AA.4.6.2 Semantic Parsing Rules

AA.4.6.3 Process

AA.4.6.4 Conclusion

Best Available Copy

AA.4.7.0 Data classification – the leakage threat

AA.4.8.0 Semantics dictionaries

AA.4.9.0 Difference between Content, Context, and Concept

AA.4.10.0 Classification as Proxy for Scarcity

AA.4.11.0Threats of Meta Data

AA.4.12.0 Degrees of data structure

AA.5.0.0 Defining Information

AA.5.1.0 The meaning of information and related context risk

AA.5.2.0 Content Concept and Context - Correcting the Security Gap for Data Object Models

AA.5.3.0 The Structure of Information

AA.5.3.1 Information Structure

AA.5.3.2 DNA versus RNA

AA.5.3.3 Granular Content

AA.5.4.0 Implementing of MS Office

AA.5.4.1 Microsoft Office Suite and Applications

AA.5.4.2 Repress Thoughts of Banishing MS Office

AA.5.4.3 Office Versions, Releases, and the Data Object Models

AA.5.4.4 Content Security

AA.5. 4.5 Implementing Document Protection

AA.5.4.6 Conclusion

AA.5.5.0 The contextual information matrix

Best Available Copy

AA.5.6.0 Future direction for information awareness

AA.5.7.0 Unstructured data

AA.5.8.0 Information measurement and management

AA.5.9.0 MS Office Document Structure: Protection and Processing

AA.5.9.1 The MS Office Object Model

AA.5.10.0 The criticality of right context

AA.6.0.0 Search and data mining

AA.6.1.0 Meta Data Hound search - Adjacency with Permutations and Combinations

AA.6.2.0 Search Convergence and Search Divergence

AA.6.3.0 Convergent search

AA.6.3.1 Divergent Search Paradigm

AA.6.4.0 Data mining and inference attacks

AA.6.5.0 Metaseaerch data mining

AA.6.5.1 Description of the DigitalDoors Metasearch Process

AA.6.5.2 Combinations and Permutations

AA.6.5.3 Automated Contextual Pruning

AA.6.6.0 Knowledge Discovery: Search Convergence and Divergence

AA.6.6.1 Initial Process

AA.6.6.2 DigitalDoors MetaSearch

AA.6.6.3 MLS-Compliant Search and Distributions

AA.7.0.0 Security:

AA.7.1.0 Security Keystone overview

Best Available Copy

AA.7.1.1 Security

AA.7.2.0 Malware attack the tentacles of Insecurity

AA.7.3.0 Security is broken

AA.7.4.0 Meta data dangerous double hedge sword

AA.7.5.0 The transition content security via Data Leakage Prevention

AA.7.6.0 The failure of best breed security

AA.7.7.0 Why Security of information

AA.7.8.0 Security as an open set

AA.7.9.0 Dangers in information interactions, the need for granular content isolation

AA.7.10.0 Prevention of leakage through context and range control

AA.7.11.0The vulnerability of perimeter security and the need for granular data control

AA.7.12.0 Leakage from Data combination

AA.7.13.0 The Risks from XML

AA.7.14.0 The risks of new information infrastructures

AA.7.15.0 Granular content dispersion as security

AA.7.16.0 Security by the One-Time Pad

AA.7.16.1 Perfect Secrecy

AA.7.17.0 Risk in repurposing of information

AA.7.18.0 The ineffectiveness of Perimeter security

AA.7.19.0 Deficiencies in perimeter security and encryption

AA.7.20.0 The breakdown of perimeter security and the need for Dispersion

AA.7.21.0 Asymmetric Security

Best Available Copy

AA.7.22.0 Dispersal security versus perimeter security

AA.7.23.0 Metadata risks, the need for a lockdown

AA.7.24.0 Competitive Analysis of Security Methods

AA.7.24.1 security Technology

AA.7.24.2 Encryption and Security History

AA.7.24.3 Cipher Science

AA.7.24.4 integrating disciplines

AA.7.24.5.Unique Architecture

AA.7.24.6 DigitalDoors versus Other Technologies

AA.7.25.0 Security by dispersion to the cloud

AA.7.26.0 Defenses in Depth: layered security

AA.7.26.1 Why layered security works

AA.2.27.0 Perimeter security, the failure of the Sandbox

AA.2.28.0 The critical danger of Commercial off the shelf software

AA.2.29.0 COTS Implementation

AA.7.29.1 Unfolding Catastrophe

AA.7.30.0 DigitalDoors Access and Rights

AA.7.30.1 Authorization

AA.7.30.2 Authentication Database

AA.7.30.3 Registration/ Rights database

AA.7.30.4 Location Map Database(s)

AA.7.30.5 Architecture Overview

Best Available Copy

AA.7.30.6 Primary Benefits

AA.7.31.0 Securing the Internet

AA.7.32.0 Spread of viruses within meta data

AA.7.33.0 Touchpoints and gateways as back doors

AA.8.1.0 Multi level Security Information Sharing:

AA.8.1.1 Information sharing overview

AA.8.2.0 Multi Level Security: new concepts for information sharing

AA.8.3.0 Information sharing through granular content control

AA.8.3.1 Gaps in Controls

AA.8.3.2 Risks from Distributed Stovepipes and Silos

AA.8.4.0 The Failure of Tag Reflexivity for Information Sharing

AA.8.4.1 Tag Reflexivity

AA.8.5.0 Multi Level Security needs to adopt granular data control to achieve its objective

AA.8.6.0 Accommodating the Bell LaPadula model

AA.8.6.1.0 Maps and Recovery Information

AA.8.7.0 Challenges of Multi level security information sharing

AA.8.8.0 Implementing Cross-Domain MLS Information Flow

AA.8.9.0 Multi level Security for the Internet

AA .8.9.1 Cultural and Technical Constraints

AA.8.10.0 A Paradigm Shift

AA.8.11.0 Implementing Cross-Domain Information Flow - Multi level security

EX. 1002, Page 82

Best Available Copy

AA.8.12.0 Negotiated multi level information sharing

AA.8.13.0 Security through information Distribution

AA.8.14.0 Implementation of information sharing with Rigorous Protections

AA.8.14.1 General Concept

AA.8.14.2 Implementation map

AA.8.14.3 Technical Implementation

AA.8.14.4 A Tool in a Toolbox

AA.8.14.5 Maturity of the Paradigm Shift

AA.9.0.0 Intangibles:

AA.9.1.0 Creation and production of intangibles

AA.9.2.0 Intangible data assetization

AA.9.2.1 Assetization for Asset Monetization

AA.9.2.2 Differential Valuation

AA.9.2.3 Conclusion

AA.9.3.0 The Intangible assets production model

AA.9.4.0 Monetary information valuation

AA.9.4.1 Monetary Valuation Methodology

AA.9.4.2 Weighting Content, Context, and Concept

AA.9.4.3 Discovering and Monetizing the Intangibles

AA.9.4.4 Static and Dynamic

AA.9.4.5 Conclusion

AA.9.5.0 Assetitization of intangible content

Best Available Copy

AA.9.5.1 Definitions

AA.9.5.2 Conclusion

AA.9.6.0 Attribution of monetary value to information

AA.9.6.1 Monetization

AA.9.6.2 Contextual Conflicts

AA.9.6.3 DigitalDoors Methodology

AA.9.6.4 Weights

AA.9.6.5 Hierarcharchial Assignments and Enhancements

AA.9.6.6 Conclusion

AA.9.7.0 Information Lifecycle Valuation

AA.9.8.0 Security Gaps: Binding Infrastructure with Process with User with Data .The need for a new intangibles production model

AA.9.9.0 Intangible Trade secrets

AA.9.9.1 Background

AA.10.0.0 Introduction to the Digitaldoors Dictionary

AA.10.1.0 DigitalDoors Dictionary

A.1 Creating An Information Infrastructure

A.2 Sampling and Creation of a Simple Info Infrastructure

A.3 System Tools and Processing Methods for an Information Infrastructure

A.4 Information Dispersal - Basic

A.5 Alternative to Information Dispersal Algorithm

A.6 Adding Prior Art Information Dispersal Algorithms to Granular Data Dispersion

Best Available Copy

A.7 Automatic Dispersal of Data to Remote Storage Locations

A.8 Dispersal - Breaking Down the Context of Data Streams while Maintaining Access to the Content of its Parts

A.9 Granular Tagging - Basic

A.10 Automated Granular Data Level Manipulation

A.11 Process for Creation of Automated Granular Tagging and Labeling

A.12 Process for Creation of Automated Tagged Tear Lines

A.13 Category Tags as Placeholders in Databases

A.14 Controlled Release - Basic

A.15 Layered Build-up of Data Streams, Layer upon Layer

A.16 Granular Data Release from Third Party Repository

A.17 Creation of Modular Data Components from Granular Data Streams

A.18 "Rolling" Exposure of Granular Data Streams

A.19 Controlled Release of Extracts after Verification

A.20 Multi Purpose Usage of Filtering and Extraction Process

A.21 Filtering and Copying "Selected Data"

A.22 Combined Process of Extracting Part of the " Selected Data" and Copying Parts of the "Selected Data"

A.23 Data Mining Search Results Basic - Distribution as Multi Level Security Versions

A.24 Data Mining - A Search Creating Automatically Key Words

A.25 Data Mining - Continuous Data Mining with Query Refinement

A.26 Automatic Continuous Search – with Divergence and Convergence

Best Available Copy

A.27 Distributed Data Mining

A.28 Content Inference Engine - Basic

A.29 An Inference Engine - Searching for the Unknowns

A.30 Securing Information Sharing with an Inference Engine

A.31 Automatic Web Data Inference Crawler Robots

A.32 Changeable Data Classifications Based on Content Inference Threats

A.33 Search Operations Creating New Continuously Growing Digital Data Streams

A.34 Distributed Searches from Different Storage Nodes

A.35 Storage of Separated Data Pieces in One Location

A.36 Storage of Granular Data Streams - Basic

A.37 Configuration of Granular Data Streams Dispersal to Distributed Storage Stores

A.38 Granular Data Stream Transport

A.39 Shuttling Data and Granular Data Streams Between Distributed Storage Stores

A.40 Replication of Granular Data Streams with Minimal Security Risk

A.41 Anonymizing Classified and Unclassified Storage Stores

A.42 Controlled Release of Granular Data Streams - Granular Data Manipulation and Control

A.43 Sanitization Process with Inference Filtering - Basic

A.44 Process for Sanitizing Data on the Network for Privacy Security and Survivability

A.45 Automated Sanitization of Important Data

A.46 Sanitization of Non-Important Data

A.47 Release of Granular Data Streams in With Increasing Levels of Access Identifiers

A.48 Release of Granular Data Streams After Obtaining Consideration

Best Available Copy

A.49 Secured Release of Granular Data Streams into Multiple Screens

A.50 Ecosystem Based on "Granular Data Control" - Basic

A.51 Eco System for Information Superiority and Survivability Based on "Granular Data Control"

A.52 System and Method for Minimizing Data Exposure

A.53 Creating "Digital Doors" for Access and Controlled Release in a Data Stream

A.54 General Aspects - Basic

A.55 Automatic Creation of a User or Object's Identification Profile

A.56 Copying Filtered Granular Data and Dispersing

A.57 Combining Extraction and Copying for Filtering of Filtered Granular Data

B.1 Basic Operational Theory of Secure or Select Data Storage

B.2 General Operation

B.3 System Configurations

B.4 Input or Initial Processing Considerations (Basic Filter Modules)

B.5 Storage

B.6 Extraction and Storage Enhancements

B.7 Basic Reconstruction

B.8 Reconstruction Techniques

C.0 Testing the Security System

C.1 The Etiology of Information

C.2 Metasearch Engine for Knowledge Discovery: Search Convergence and Divergence

C.3 Initial Process - Metasearch Engine and Knowledge Expander

C.4 Simple Metasearch Engine

Best Available Copy

C.5 Automatic Metatagging

C.6 Automatic secondary tagging of a "range"

C.7 MLS, Multi-level Security-Compliant Search and Distributions

C.8 Benefits of the Metasearch - Knowledge Expander

C.9 Information Life Cycle Engine

C.10 Information Life Cycle Searches

C.11 Search With Prioritized Classifications

C.12 Simple Filters

C.13 Complex Filters

C.14 The Editor As An Information Expander

C.15 Document Object Model (DOM) - Protection and Processing

C.16 Recognizing The Document Object Model (DOM)

C.17 Content

C.18 DOM Process Editor

C.19 Five Informational Attributes

C.20 Select Content Introns and Exons

C.21 Hierarchical Taxonomic Class Examples

C.22 Knowledge Expander (KE) Basic Program

C.23 Secured Metasearch Engine Based on Granular Data Control

D.1 Secure Editor, DOM Editor (General) and DOM Editor Applied as Extractor Modules

D.2 Applications of the Secure and SC Data Systems

D.3 Automatic Features

Best Available Copy

D.4 Multiple Security Levels

D.5 Sharing Data with Different Security Levels - Data Mining

D.6 Multilevel Security through Sanitization with Reconstruction of Sanitized Content

D.7 Multiple Independent Levels of Security (MILS)

D.8 MLS Server and Workstation Components

D.9 Flexibility of the Present MLS Approach

D.10 Multiple Extraction Filters and Application Outline

D.11 Multiple Security Features for Data

D.12 Adaptive Data Security -Event Responsive System

D.13 Email, Web-based and Other Types of Applications

D.14 Portable Computing Device Environment

E.1 General System Comments

[page blank]

Brief Description of the Drawings

[0016] Further objects and advantages of the present invention can be found in the detailed description of the preferred embodiments when taken in conjunction with the accompanying drawings

Best Available Copy

in which:

[0017] FIGs. A-1a and A-1b diagrammatically illustrates a basic system diagram, method and flow chart for an information infrastructure.

[0018] FIG. A-2 diagrammatically illustrates a basic information life cycle process or system wherein secure content (sec-con) and select content (SC) is separately stored per categorical filters operable by the enterprise.

[0019] FIG. A-3 diagrammatically illustrates a simplistic dynamic filter and select content (SC) storage coupled to a data processing event.

[0020] FIG. B-1A diagrammatically illustrates a basic system diagram showing filtering and storing extracted data and remainder or common data and, in an enhanced embodiment, generating and storing a map.

[0021] FIG. B-1B diagrammatically illustrates a system diagram showing reconstruction of the data, various security clearances and both electronic reconstruction and visual reconstruction.

[0022] FIG. B-2 diagrammatically illustrates a system showing major components of a single personal computer (PC) system, a networked system with several PCs (a LAN or WAN) and the network coupled to a telecommunications system and the Internet and shows the interconnection with a server and multiple, Internet-connected memory units.

[0023] FIG. B-3 diagrammatically illustrates a basic flowchart showing reconstruction for various security levels.

[0024] FIG. B-3A diagrammatically illustrates interleaving distinct data into different memory locations in a video memory.

[0025] FIG. B-4 diagrammatically illustrates a flowchart showing one embodiment of the

Best Available Copy

principal portions of the data security program.

[0026] FIG. B-5 diagrammatically illustrates a flowchart showing the basic elements of the reconstruction process for the data security program.

[0027] FIG. C-1 diagrammatically illustrates deconstruction of an input document into segments.

[0028] FIG. C-2 diagrammatically illustrates the information life cycle knowledge expander (KE) search engine process.

[0029] FIG. C-3 diagrammatically illustrates the KE process of using search terms, data collection and depth-of-search control.

[0030] FIG. C-4 diagrammatically illustrates the KE process for the primary search results.

[0031] FIG. C-5 diagrammatically illustrates the KE process for the output representation of search term data and search result data of supplemental documents.

[0032] FIG. C-6 diagrammatically illustrates a filter program.

[0033] FIG. C-7 diagrammatically illustrates an editor program.

[0034] FIGs. C-8A, C-8B. C-8C and C-8D diagrammatically illustrate extraction of select content (SC) and a hierarchical taxonomic system based upon TS, T, C and UC security code classifications.

[0035] FIG. C-9 diagrammatically illustrates a data object tree or map for a document model (DOM).

[0036] FIG. C-10 diagrammatically illustrates varies metadata levels in the DOM.

[0037] FIG. C-11 diagrammatically illustrates a DOM editor process.

[0038] FIG. C-12A diagrammatically illustrates another DOM editor process.

EX. 1002, Page 91

Best Available Copy

[0039] FIGS. C-12B, C-13C and C-12D show examples of categorical identifiers, ranked id categories and categorical ids for a business.

[0040] FIG. C-13 diagrammatically illustrates a basic KE search process.

[0041] FIG. C-14 diagrammatically illustrates an expanded KE search process.

[0042] FIG. C-15 diagrammatically illustrates KE search process options.

[0043] FIG. C-16 diagrammatically illustrates a hardware implementation for the KE search system.

[0044] FIGs. D-1, D-2, D-3, D-4 and D-5 shown in the drawings as omitted.

[0045] FIG. D-6 is a computer network diagram showing various user communities.

[0046] FIG. D-7a diagrammatically illustrates a flowchart showing the key component steps for the multiple layer security program for the community of users.

[0047] FIG. D-7b diagrammatically illustrates a multiple level security system accessed by users having different security clearances (which also represents a data mining system and operation).

[0048] FIG. D-8 diagrammatically illustrates a flowchart showing the key components of an adaptive security program adaptable to various levels of electronic attacks, hacker or hack attacks.

[0049] FIGs. D-8A to D-8F show tabular examples of attack and response levels for Standard Automatic Defenses Matrix; Optional Automatic Defenses Matrix; Security Meter Module; Normal Work Mode; Threat Made Table; and, Attack Mode.

[0050] FIG. D-9 diagrammatically illustrates a flowchart showing the key components of a multiple encryption program using multiple types of encryption in one document or data object. Encryption may include data compression and/or expansion or parsing techniques.

[0051] FIG. D-10 diagrammatically illustrates a chart showing the key components of the

Best Available Copy

parsing, dispersion, multiple storage and reconstruction (under security clearance) of data.

[0052] FIGS. D-11A and D-11B diagrammatically illustrate a flowchart showing the key components of one embodiment of the e-mail security system (jump points D-11-A and D-11-B link the flow charts).

[0053] FIGS. D-12A and D-12B diagrammatically illustrate a flowchart showing the key components of one embodiment of the invention implements the security system on a web browser (jump point 12-A links the flow charts).

[0054] FIG. D-13 diagrammatically shows several revenue systems which may be employed with the data security systems described herein.

[0055] FIG. D-14 diagrammatically illustrates a portable computing device (or the critical parts thereof, see FIG. B-2 for further details) associated with the portable data security locator system and method.

[0056] FIG. D-15 diagrammatically illustrates a basic flow chart for the portable security system program in accordance with the basic principles of the present invention.

[0057] FIG. D-16 diagrammatically illustrates a MILS or multiple independent levels of security computer network with three (3) sub-networks at security levels top secret TS, secret S, and unclassified U.

[0058] FIG. D-17 diagrammatically illustrates a security guard computer useful in the network of FIG. 16.

[0059] FIG. D-18 diagrammatically illustrates a workstation useful in the MILS network.

[0060] FIG. D-19 diagrammatically illustrates a filter program flow chart.

[0061] FIG. D-20 diagrammatically illustrates a basic security editor program.

FIGs. D-21, D-22, D-23 shown in the drawings as omitted.

[0062] FIG. D-24 diagrammatically illustrates a general flow chart for a comprehensive filter operable on a document DOM.

[0063] FIG. D-25 diagrammatically illustrates a basic application of the comprehensive DOM editor.

[page left blank]

Best Available Copy

Detailed Description of the Preferred Embodiments

[0064] The present invention relates to an information infrastructure with many different features and characteristics. The infrastructure is a system, a methodology and a process implemented on a personal computer (PC) system, on a computer network (LAN or WAN) and over the Internet and computer programs and computer modules and an information processing system to accomplish the functions described herein. An Abbreviations Table is provided at the end of this specification. Also, the General System Comments section E.1 describes certain aspects relative to the implementation of this invention. In the drawings, and sometimes in the specification, reference is made to certain abbreviations. The Abbreviations Table (section E.1) near the end of the specification and the Digital Doors Dictionary at section AA.10.1.0 provide some correspondence between the abbreviations and the item or feature.

AAA.1.0.0 DigitalDoors Capabilities

[0065] DigitalDoors delivers security solutions which conform to the secured information sharing requirements as stated in the Presidential Executive Order 13356 of 8.27.04 and in the Senate's National Intelligence Reform Act passed in September 2005.

[0066] DigitalDoors delivers information assurance through security processes that secure automatically (or manually) the data content elements. This simultaneously enables data security, data privacy protection, secured sharing and collaboration, and survivability within a unified infrastructure.

[0067] DigitalDoors technology provides security against insiders and augments traditional perimeter-based security methods. It represents a major step toward granular data content security and secured leveled information sharing necessary for unified national security and protection of privacy. It introduces granular content security and privacy protection, while enabling data sharing without

Best Available Copy

exposures of source, methods, and activities.

[0068] DigitalDoors is inherently platform-independent and process-independent; it is a single enterprise-wide solution for all types of platforms, network infrastructures, cross-platform distribution, extended information workflows, and traditional data processing operations. The DigitalDoors technology is fully flexible and compliant with existing and developing standards for databases, XML, and hierarchical metadata, as specified in this document.

[0069] Given the practical implementation of data security at a granular data level and the modular, multi-platform initiative, the core taxonomic technology enables anonymous information sharing so that organizations can share information to connect the dots in a practical, sustainable, and effective process.

AAA.1.1.0 Architecture Specifications

[0070] DigitalDoors configuration is robust enough to support security while maintaining data structures consistent with legacy data operations, traditional desktop applications, and most structured and formatted data types. Support for HTML and XML is available at this time. Specifically, DigitalDoors will operate with SOAP, WSDL, UDDI, WS-Interoperability, XACML, PKI, XKMS, X-KRSS, biometric, SAML, and most other web-enabled architectures, trusted platforms, Unix, Linux, as well as host-based legacy environments. The technology is feasible within or across domains, and can function within multiple level trusted domains.

[0071] In DigitalDoors existing network, system, data processing, ID assurance, encryption technology, Internet-based transaction processing, messaging, and database activities coexist with limited workflow complications. DigitalDoors reviews data sensitivity level at the payload and messaging level; as a result, it can interoperate with the GIG transport services, with core enterprise

Best Available Copy

services, and any community of interest.

AAA.1.2.0 Scalability

[0072] DigitalDoors is scalable both vertically and horizontally, or any mix thereof. The functions can be run centralized or dispersed, or both. It can run as hardware or software. As software, it can run on low-end user desktops, servers, appliances, or massively-parallel hosts. Complex processes (or sense CPU-bound processes) can be offloaded to coprocessors or grid-based resources. Demonstrations show functionality on past generation PC hardware without any peculiar software or hardware, thereby confirming a simple scalability scenario. DigitalDoors is also designed from the ground up with networking awareness and sensitivity level awareness to support single user environments, networked client-server infrastructures,and complex distributed GIG-type scenarios. Location references support local and fully-qualified paths, X-paths, X-pointers, and other long addressing methods.

AAA.1.3.0 Product and Process Overview

[0073] DigitalDoors provides Defense in Depth security capabilities which ensure the Integrity and continuity of the information environment and the information it stores, processes, maintains, uses, shares, disseminates, disposes, displays, or transmits.

[0074] DigitalDoors presents solutions to the following vulnerabilities in networks, computing infrastructure, core services and discreet systems:

Creation of many single points of failure in a net centric environment

Survivability and lack of constant data availability

Lack of resiliency against attacks including denial of service attacks

Unsecured collaboration and cross domain data exchange leakage

EX. 1002, Page 97

Best Available Copy

Protection against insider attacks,

Defense against malicious traffic

Defense against war operations including spy ware

AAA.1.4.0 DigitalDoors implementation options:

[0075] It is an add security capability which can be overlaid and integrated into existing security management systems. It can function as a stand alone security service.

AAA.1.4.0 Perpetual Data Availability

[0076] DigitalDoors delivers basic security Access Management capabilities, however it goes far beyond that and delivers the capability for perpetual data availability. Perpetual data availability enables access to data even when different storage servers are attacked or compromised. Single points of failure, are eliminated by dispersing the data into different storage servers. To eliminate the risk in replication of the data the documents /data streams are split into granular data segments (both data and metadata).

[0077] Compromise of an encrypted granular data segment on a compromised server can't compromise the entire document or data stream.

AAA.1.5.0 The Key for Net-Centric security

[0078] Net-Centric security will be based on granular content control. DigitalDoors products are based on breakthrough patented technologies, which deliver comprehensive solutions to critical net centric needs. The products deliver maximum security, survivability and collaboration capabilities through the introduction of security at the granular data level within documents and data streams. The process can be outlined as follows:

Best Available Copy

[0079] DigitalDoors creates automatically or manually a sanitized version of the original document and/or a data stream, which contains no sensitive data.

[0080] The system in essence creates a "lowest common denominator" a non classified, sanitized document.

[0081] That document version can be shared with many users including those with low or no clearances.

[0082] The critical granular content within the document or data stream is tagged, encrypted, extracted and dispersed to distributed storage.

[0083] The result: a remainder unclassified sanitized document/ data stream with its critical /classified components stored in distributed storage.

[0084] An attack will not cause damage or only very limited damage since the critical classified granular content is not in the original document.

[0085] The sanitized document will be upgraded for each specific user according to his security level and role at wire speed. Sanitized, declassified documents and/or data streams will be converted back into classified documents, at wire speed, when a user presents his identification.

[0086] The reconstitution will take place based on the clearance levels and roles of each user.In essence different versions of the original document/ data stream will be reconstructed according to the rights of each different user.

[0087] Users who present their credentials will see instantly those missing parts of the document that they are allowed to see. As an example a mid level intelligence analyst will see the most of the elements reconstituted back into the document while a local commanding officer will see a reconstituted version with all the original elements.

Best Available Copy

[0088] The remainder document is a critical element presented by DigitalDoors. The remainder document enables not only wide secured sharing but also a centralized management of documents and/or data streams while the secure elements are decentralized to avoid single points of failure. This includes solutions against malicious traffic:

AAA.1.6.0 Active defense and counterattacks versus common passive defense

[0089] Critical content is not there, creation of air gaps. Work within all 7 OSI layers with filters will process every header and payload packet.

[0090] Firewalls and intrusion detection appliances are not successful in detouring major malicious attacks. There are different systems analyzing packets headers and payloads for malicious data blocking. DigitalDoors takes a different approach. DigitalDoors sees those data blocking efforts as passive defenses. Passive defenses have been proven to be much less useful than active defenses and counterattacks. In active defenses, the defender takes advantage of the terrain and other elements.

[0091] DigitalDoors was designed to operate as an active defense and as counterattack. It assumes that even the best of breed systems will not be able to block thoroughly the growing number and sophistication of malicious traffic.

[0092] The basic assumption is that the attacker will always be a step ahead of the defender. DigitalDoors therefore creates the ultimate last line of defense by having the critical elements of content extracted to dispersed secure storage locations.

[0093] In essence DigitalDoors creates air gaps. Malicious attacks might hurt the non-critical remainder content but the critical content is kept away in highly secure distributed storage locations unknown to the attackers.

[0094] Controlled release of critical content from storage locations will take place only after

Best Available Copy

extensive scans for malicious attacks.

[0095] DigitalDoors standard functionality is focused at ISO layer 7 and above; specifically, it is intended to function within the presentation and application layers of the ISO model and protect data within infinitely granular hierarchies. As more traditional networking filter implementation, it functions and can be separately implemented at layers 1 through 6. With the filters, each bit it comes into contact with, header or payload is analyzed. Malicious elements are filtered out at wire speed. DigitalDoors will add on the best of brand malicious filtering capabilities after their certification by DoD. The DigitalDoors new version will enable malicious attacks to trigger DigitalDoors defenses-in-depth's last line of defense. Intrusion detection alerts will trigger a double-tiered defense:

[0096] Substantial extractions of content from documents, data streams will take place. The extracted critical content will be transported to distributed storage.

[0097] The release of content from storage for reconstitution in documents and data streams will be stopped or minimized.

AAA.1.7.0 Technical Basis for DigitalDoors Technology

[0098] DigitalDoors is a novel but practical combination of well-known computing disciplines, including secret sharing, information dispersal algorithms, distributed computing and best of breed security methodologies. In essence, the system creates virtual "digital doors" in documents and data streams. Those digital doors establish security and enable secured collaboration through defenses-in-depth capabilities to ensure end to end integrity, continuity of the data in storage and in transfer. DigitalDoors offers distinct core solutions and advantages that are not solved by DOD initiatives and products like CIBIS, CASPER, cross domain guards, and best of breed commercial products.

Best Available Copy

[0099] DigitalDoors contributes to the security by implementation of granular data control. The DigitalDoors solutions include the following cornerstones:

[0100] Net centricity creates major benefits but also creates dangerous risks.

[0101] The key to manage those risks is through granular control of the critical elements within the networks.

[0102] The critical element within the networks is data, which needs to be controlled granularly at the granular data level.

[0103] Therefore documents / data streams need to be parsed manually or automatically into smaller granular content parts, in order to enable granular control.

[0104] These granular content parts need to be tagged and classified per security level and role.

[0105] Users will have access to those granular tagged parts according to the rights they were assigned in their user profile.

[0106] In essence there is a need in net centric environments for recreation of various digital versions, on the fly which are reconstructed according to different users rights and roles.

[0107] Creations of on the fly versions save costs both in bandwidth storage and speed. There is no need to keep in storage many different versions and there is no need to send full documents multiple times across networks.

[0108] To counter the major security risks of net centricity the granular content parts should be tagged, encrypted, extracted, and dispersed to distributed storage locations.

[0109] Providing perpetual data availability and data survivability by replication of the different granular parts and storing the replicated copies on many servers.

Best Available Copy

[0110] Replication on many servers is enabled without security risk because the replicated items are very granular and even if they are compromised they are out of any context and do not compromise the bigger document or data stream.

AAA.1.8.0 Granular data control

[0111] There is a need to create "digital doors" in documents and data streams to enable granular control and achieve maximum collaboration, security, survivability, confidentiality, and privacy.

[0112] Digitaldoors Creates Granular Digital Rights Management for Collaboration. The core enabler of the DigitalDoors collaboration product is the creation and control of granular data segments within a document and/or a data stream. The parsing of granular content segments in a documents and/or data stream, can be done manually, automatically, or a combination of both. Metatags are assigned manually and also automatically to the newly created granular data segments. The Metatags include user roles, and define who and what processes are allowed to access each segment. When a user requests data through the discovery services, an automatic matching process will take place, matching user's profile with the metatags. The control of granular parts of content within data streams enables sharing of content with many users where each can see what he is allowed to see in a document/ data stream in accordance with his personal profile and the policy embedded into the metatags within the different content segments.

[0113] Users will be able to access the data according to their roles. For example a user with an intelligence role will be able to view the whole reconstructed content while a user with a logistics role will be able to view only a limited reconstructed version of the original content.

AAA.1.9.0 Collaboration will consist of:

[0114] User's presentation of his identification

Best Available Copy

[0115] Automatic matching of user's profile with Metatags of the different tagged segments of the document/data stream.

[0116] Wire speed reconstruction of a version of a document/ data stream according to users profile.

AAA.1.10.0 Granular tagging with metadata

[0117] Label data (and other resource) at content and context level for sensitivity, compartment, information, categorization, and other arbitrary designations

[0118] User and process clearances (inclusion and/or exclusionary) based on sensitivity, compartment, information, categorization, and other arbitrary designations

Granular Digital Rights management:

Support revocation of rights in real time

Support downgrading in real time

Support upgrading in real time

Controlled release of granular data:

Enable controlled release of data at all user , process and application levels

Enable distribution of core services and rapid information deployment to the edge

AAA.1.11.0 Coalition And Multinational Cross Domain Collaboration

[0119] DigitalDoors enables secured cross domain sharing as urgently needed in coalition and multinational cross domain collaboration. Real time ad hoc interaction between different trusted and untrusted participants is enabled in the very difficult cross domain environment. A document/ data stream is divided to parts, each part is tagged according to the security level of the content (TS, S, C, U), it can also be tagged for the roles of users that should be allowed to access it (TS a, TS b, and Sa).

AAA.1.12.0 Incorporation into a trusted OS

[0120] Incorporated with a trusted OS (for example a certified Trusted Solaris and a TCS guard) DigitalDoors enables a user to divide a document into granular content segments, tagging and encrypting each of the newly parsed parts according to the classification level of its content (TS, S, C, U) and dispersing the tagged parts to different distributed (TS, S, C, U) storage locations throughout the enterprise. The different segments can also be tagged for users roles. (TS a, TS b, S a, etc.).Users will be able to access a document or data stream they request which will be reconstituted at wire speed according to their role and clearance level.

[0121] Cross domain collaboration is enabled with connectivity of DigitalDoors products to a certified guard.

Cross domain sharing in both directions High to Low, Low to High

Secured Cross Domain sharing for US, and coalition forces

Advanced metatags for security clearness and roles

Organizational and user defined filters

AAA.1.13.0 Survivable infrastructure

[0122] DigitalDoors Creates a secured survivable information security services infrastructure that lowers costs and maximizes secure collaboration for real time situational awareness.

[0123] It Secures data at point of origination, in rest and in transit based on methods of content granularity.

[0124] Provides data confidentiality, availability, integrity, and non-repudiation to users with override functionality and lowest common denominator as failsafe

Best Available Copy

[0125] Cross-platform, cross-network, and integrates with existing identification, authentication, and authorization systems

[0126] Integrates with Cross Domain guards with EAL-4 or higher certifications

[0127] Support platform, guard, network, workflow, gateway data, and metadata interoperability for allies, partners, coalition forces

AAA.1.14.0 Net Centric capabilities

[0128] DigitalDoors suite of net centric products and services delivers the following sought solutions to critical challenges of net centric organizations:

1. Security
2. Survivability
3. Constant data availability
3. Secured collaboration/ coalition cross domain sharing
4. Automatic posting and alerting
5. Specialized discovery/data mining

[0129] The DigitalDoors evolutionary services can enable users to get the right information, at the right time, in the right format, as needed to meet operational tactical and support missions.

AAA.1.15.0 DigitalDoors Services overview

[0130] DigitalDoors is a net centric core technology which delivers among others, the following services:

1. Secured collaboration services

[0131] "The existing Collaboration capabilities are technically outdated, expensive, have serious

Best Available Copy

security flaws and do not scale well. There is the need to continually look at collaboration technology trends and potential redesigns and/or enhancements"

[0132] Global Information Grid Core Enterprise Services Strategy- Draft version 1.1a 7/9/03

[0133] Office of the Assistant Secretary of defense for Networks and information integration

[0134] The DigitalDoors product addresses each of the above limitations taking secure collaboration from an "elusive goal" to an achievable goal.

[0135] DigitalDoors products deliver an evolutionary technology, which enables secured seamless collaboration .It provides comprehensive access to information from anywhere, at any time on any device and application.

AAA.1.16.0 DigitalDoors implementation options:

[0136] It is an add on collaboration capability which can be overlaid and integrated into existing shared workspaces, whiteboards and applications etc.

[0137] Or it can function as a stand alone collaboration service.

AAA.1.17.0 defenses against outsuiders and insiders

[0138] DigitalDoors defends against possible killer collaboration traps , security killer gaps as described in the IA/ Security services segment. Possible collaboration killer gaps can be:

War operations on collaboration sessions including substitution of misinformation in sessions by outsiders or insiders

Destruction of storage and backup storage

Unavailability of data due to denial of services attacks

[0139] DigitalDoors contributes breakthrough security added value to the following

AAA.1.18.0 collaboration elements:

Session management

Text Collaboration

Audio and Video over IP Network

White boarding and annotation

Application Sharing

Application Broadcasting

Virtual spaces

[0140] DigitalDoors contributes to the security of collaboration by implementation of granular data control.

AAA.1.19.0 DigitalDoors cornerstones

[0141] The DigitalDoors solutions are based on the following cornerstones:

[0142] Net centricity creates major benefits but also creates dangerous risks.

[0143] The key to manage those risks is through granular control of the critical elements within the networks.

[0144] The critical element within the networks is data, which needs to be controlled granularly at the granular data level.

[0145] Therefore documents / data streams need to be parsed manually or automatically into smaller granular content parts, in order to enable granular control

[0146] These granular content parts need to be tagged and classified per security level and role.

[0147] Users will have access to those granular tagged parts according to the rights they were assigned in their user profile.

[0148] In essence there is a need in net centric environments for recreation of various digital

Best Available Copy

versions, on the fly which are reconstructed according to different users rights and roles.Creations of on the fly versions save costs both in bandwidth storage and speed. There is no need to keep in storage many different versions and there is no need to send full documents multiple times across networks.

[0149] To counter the major security risks of net centricity the granular content parts should be tagged, encrypted, extracted, and dispersed to distributed storage locations.

[0150] Providing perpetual data availability and data survivability by replication of the different granular parts and storing the replicated copies on many servers.

[0151] Replication on many servers is enabled without security risk because the replicated items are very granular and even if they are compromised they are out of any context and do not compromise the bigger document or data stream.

AAA.1.20.0 Granular content control

[0152] DigitalDoors creates Granular Digital Rights management for collaboration:

[0153] The core enabler of the DigitalDoors collaboration product is the creation and control of granular data segments within a documents and/or datastream. The parsing of granular content segments in a documents and/or datastream, can be done manually, automatically or a combination of both. Metatags are assigned manually / automatically to the newly created granular data segments. These metatags include also which user roles are allowed to access each segment. When a user requests data through the discovery services, an automatic matching process will take place, matching user's profile with the metatags. The control of granular parts of content within data streams enables sharing of content with many users where each can see what he is allowed to see in a document/ data stream in accordance with his personal profile and the policy embedded into the Metatags of the different content segments.

[0154] Users will be able to access the data according to their roles. For example a user with an

Best Available Copy

intelligence role will be able to view the whole reconstructed content while a user with a logistics role will be able to view only a limited reconstructed version of the original content. Collaboration will consist of:

1. User's presentation of his identification

2. Automatic matching of user's profile with Metatags of the different tagged segments of the document/data stream.

3. Wire speed reconstruction of a version of a document/ data stream according to users profile.

[0155] Granular tagging with metadata:

1. Label data (and other resource) at content and context level for sensitivity, compartment, information, categorization, and other arbitrary designations

2. Assign users and processes clearances (inclusion and/or exclusionary) based on sensitivity, compartment, information, categorization, and other arbitrary designations

[0156] Granular Digital Rights management:

1. Support revocation of rights in real time

2. Support downgrading in real time

3. Support upgrading in real time

[0157] Controlled release of granular data:

1. Enable controlled release of data at all user , process and application levels

2. Enable distribution of core services and rapid information deployment to the edge

[0158] DigitalDoors enables secured cross domain sharing as urgently needed in coalition/ multinational cross domain collaboration. Real time ad hoc interaction between different trusted and

untrusted participants is enabled in the very difficult cross domain environment. A document or data stream is divided to parts, each part is tagged according to the security level of the content (TS, S, C, U), it can also be tagged for the roles of users that should be allowed to access it (TS a, TS b, Sa). Incorporated with a certified OS (like A DIA certified Trusted Solaris and a TCS guard) DigitalDoors enables a user to divide a document into granular content segments, tagging and encrypting each of the newly parsed parts according to the classification level of its content (TS, S, C, U) and dispersing the tagged parts to different distributed (TS, S, C,U) storage locations throughout the enterprise. The different segments can also be tagged for users roles. (TS a, TS b, S a etc.).Users will be able to access a document or data stream they request which will be reconstituted at wire speed according to their role and clearance level.

AAA.1.20.1 Cross domain

[0159] Cross domain collaboration is enabled with connectivity of DigitalDoors products to a certified guard.

[0160] Cross domain sharing in both directions High to Low, Low to High

[0161] Secured Cross Domain sharing for US, and coalition forces

[0162] Advanced Meta tagging for security clearness and roles

[0163] Organizational and user defined filters

AAA.1.21.0 Security services

[0164] DigitalDoors provides Defense in Depth capabilities which are far beyond the current state of the art. Those defenses in depth ensure:

[0165] The integrity and continuity of the information environment and the information it stores, processes, maintains, uses, shares, disseminates, disposes, displays, or transmits.

AAA.1.22.0 solutions to vulnerabilities:

[0166] DigitalDoors presents solutions to the following vulnerabilities in networks, computing infrastructure, core services and discreet systems:

1. Creation of many single points of failure in a net centric environment

2. Survivability and lack of constant data availability

2. Lack of resiliency against attacks including denial of service attacks

2. Unsecured collaboration and cross domain data exchange leakage

3. Protection against insider attacks,

4. Defense against malicious traffic

5. Defense against war operations including spy ware

AAA.1.23.0 Replication for perpetual data availability

[0167] DigitalDoors delivers basic security Access Management capabilities, however it goes far beyond that and delivers the capability for perpetual data availability:

[0168] Perpetual data availability enables access to data even when different storage servers are attacked or compromised.

[0169] Single points of failure, are eliminated by dispersing the data into different storage servers.

[0170] To eliminate the risk in replication of the data the documents /data streams are split into granular data segments (both data and Meta data).

[0171] Compromise of an encrypted granular data segment on a compromised server can't compromise the entire document or data stream.

AAA.1.24.0 Solving net centric needs

[0172] DigitalDoors products are based on breakthrough patented technologies, which deliver comprehensive solutions to critical net centric needs. The products deliver maximum security, survivability and collaboration capabilities through the introduction of security at the granular data level within documents and data streams.

[0173] Firewalls and intrusion detection appliances are not successful in detouring major malicious attacks. There are different systems analyzing packets headers and payloads for malicious data blocking. DigitalDoors takes a different approach. DigitalDoors sees those data blocking efforts as passive defenses. Passive defenses proved to be much less useful then active defenses and counterattacks. In active defenses the defender takes advantage of the terrain and other elements.

[0174] DigitalDoors was designed to operate as an active defense and as counterattack. It assumes that even the best of breed systems will not be able to block thoroughly the growing number and sophistication of malicious traffic.

[0175] The basic assumption is that the attacker will always be a step ahead of the defender. DigitalDoors therefore creates the ultimate last line of defense by having the critical elements of content extracted to dispersed secure storage locations.

[0176] In essence DigitalDoors creates air gaps. Malicious attacks might hurt the non-critical remainder content but the critical content is kept away in highly secure distributed storage locations unknown to the attackers.

[0177] Controlled release of critical content from storage locations will take place only after extensive scans for malicious attacks.

AAA.1.25.0 Secured Discovery and publishing services

[0178] Provide edge users with the ability to search, mine, and pull data based on Meta data

created automatically and or manually

[0179] Providing content and context security while maintaining catalog, search, and mining services regardless of sensitivity levels

[0180] Designed to support common delivery and posting of data in real-time

[0181] The challenges:

[0182] Enable users, down to the last tactical mile, with the capability to pull whatever they want, from wherever they want, from wherever they are.

[0183] There is a need to publish content as well as post Meta data which will enable to locate the published material.

[0184] DigitalDoors discovery services enable different users across the GIG to locate relevant

[0185] Files, data bases, services, directories, web pages, data streams, by using Meta data descriptions stored in various directories, registries, catalogs, repositories.

[0186] DigitalDoors is a security based Discovery service. Discovery services can create major vulnerabilities to the GIG; an attacker who accesses successfully a Discovery service must not be permitted to access sensitive data.

[0187] The following service are available for Discovery and Publishing services:

AAA.1.26.0 Publishing;

[0188] DigitalDoors enables publishing and posting data to common storage spaces as early as possible.

[0189] The system posts sanitized versions of the documents and/or data streams (full documents if needed.)

[0190] Documents are sanitized manually or automatically and then posted automatically.

[0191] The posting can be done automatically as part of an organizational policy. The advantage of posting sanitized versions of a document is that the user maintains control of the document's critical granular content.

[0192] The viewers who are interested in a published, posted sanitized document can request from the owner or a mediating authority the missing content. The full document can be released for consideration including exchange of sought information or payment etc.

[0193] The sanitized platform document can be automatically posted within the community of interest or other places so other users can locate it easily. The posting is done automatically or manually.

AAA.1.27.0 Meta data Posting

[0194] Meta data posted in directories, registries catalogs etc. should only enable a user access to a sanitized version of the document/ data stream.

[0195] Users who are interested to see the full version of a document/ data stream will need to present identification.

[0196] The presentation of an ID will transform the sanitized, declassified documents and/or data streams back into a fuller version, at wire speed, based on their clearance levels and roles. Different versions of the original document/ data stream will be reconstructed according to the rights of each different user

[0197] Meta data is created automatically/ manually and may be posted in Meta data directories in Communities of interest or other locations.

[0198] Edge users pull capabilities at last mile.

[0199] In addition to labeling and tagging content segments, DigitalDoors scans and categorizes all content in a document/ data stream. A user can search for information based on Meta data in tags and

also by key word searching and category searching. The search results will be in delivery of a sanitized results .With presentation of identification a tailor made version will be created for each user based on his profile.

AAA.1.28.0 Secured Mediation services for information sharing

[0200] In order to enable any substantial aggregation, fusion, translation and integration of data for delivery of situational awareness support, there is a need to motivate different people in the network to share their data.

[0201] How can data owners be motivated to share with others?

[0202] DigitalDoors enables users negotiate the exchange of data, services or products in a net centric environment.

[0203] DigitalDoors Mediation services creates a framework for negotiations between parties that might not have any former relationship or trust.

[0204] The basis of all negotiations and brokering is giving users a motivation to share their data.

[0205] The consideration for data can be information, services, money etc.

[0206] A true market place, for arms length negotiations is enabled by the ability to release granular content in a controlled way following receiving of agreed upon consideration.

[0207] Data can be released in parts and in stages based on the response of the other parties. Such mediation is automated or can be done also manually.

AAA.1.29.0 Other services

[0208] Other Mediation services provided by DigitalDoors Are:

Full auditing is enabled.

Auctioning is enabled, highest bidder will get all or parts of released of data

Discovery Metadata dictionaries

Taxonomies and Ontologies.

AAA.1.30.0 Secured perpetual storage services

[0209] A net centric environment demands secure perpetual data availability. There is a need to create a survivable storage net centric environment, avoiding single points of failure.

[0210] Replication of data on distributed servers defeats single points of failure, but it creates security risks based on over exposure data. DigitalDoors overcomes the replication risks be overcome and delivers a perpetual data availability service.

[0211] DigitalDoors delivers basic security Access Management capabilities, however it goes far beyond that and delivers the capability for perpetual data availability. Perpetual data availability enables access to data even when different storage servers are attacked or compromised. Single points of failure, are eliminated by dispersing the data into different storage servers. To eliminate the risk in replication of the data the documents /data streams are split into granular data segments (both data and Meta data).Compromise of an encrypted granular data segment on a compromised server can't compromise the entire document or data stream.

AAA.1.30.1 Meta data posting

[0212] DigitalDoors enables posting data to common storage spaces as early as possible.

[0213] The system posts:

Meta data in directories, registries and

Posts / publishes sanitized documents (full documents if needed.) in public spaces in the Communities of interest.

[0214] Documents are sanitized manually or automatically and then posted automatically.

The advantage of publishing sanitized versions of a document is that the user maintains control of the document's critical granular content. The viewers who are interested in a published sanitized document can request from the owner or a mediating authority the missing content. The full document can be released for consideration including exchange of sought information or payment etc.

[0215] DigitalDoors enables real time updates to each extracted content segment in storage. As events evolve content can be changed, updated, erased etc.

AAA.1.31.0 Secured enterprise management services

[0216] DigitalDoors delivers highly granular Digital Rights Management across the enterprise. Delivering paradigm-changing technology that solves critical long-term challenges in security and sharing Security by screening, capture extraction tagging and reconstitution of critical granular content.

[0217] Creation of content real time awareness throughout the networks

[0218] Data is audited, replicated, and transported to distributed storage

[0219] An enterprise wide management service can control highly compartmentalized access to Communities of Interest, and to compartments within the Communities of interest.

[0220] All data will be sanitized. Users within a community will have access to Meta data and sanitized versions of the documents, e mail or other data streams they are seeking. Upon providing an additional identification the sanitized documents will be reconstituted according to the user's role and clearance level at wire speed. The service will support Public key as well as biometric appliances.

AAA.1.32.0 Access enforcement:

[0221] DigitalDoors has the capability to enable access on a very granular basis uncommon today.

[0222] DigitalDoors delivers a technical proprietary capability for granular Digital Rights

management which will fundamentally change the way sensitive and private information is managed. DigitalDoors has developed its own access management basic capabilities: Identity management, Authentication, Authorization and Access enforcement. DigitalDoors was designed to integrate within large organizations and therefore it can incorporate or integrate with other such capabilities.

AAA.1.33.0 Real time granular audit of all critical content in network

[0223] DigitalDoors delivers unprecedented logging and audit capabilities. Enabling detecting of attacks by usage patterns.

[0224] The granular classification of data and the granular control of access to that data enables real time granular audit of each user in the network. The auditing entity can zoom into the documents and the parts of the document a user has accessed. This granular capability is a forceful tool in locating insider attacks. An automated analysis of the logs alerts to threatening usage patterns. The system can locate anomalies in times of access and usage, in location, and anomalies in the type of content accessed within documents or data streams.

[0225] Time stamping and notarization with digital certificates will be offered .

AAA.1.33.0 Secured Messaging services

[0226] The Granular content control introduced by DigitalDoors is an enabler for a wide range of messaging services supporting delivery of information and changes in real time with alerting. DigitalDoors off the shelf Outlook e mail is an example of integrating the DigitalDoors technology within messaging services.

AAA.1.34.0 Applications for operating an assured environment services-

[0227] According to DISA Establishing a "cloud" a "network" of enterprise computing service providers, delivering operational hosting environments consisting of common hardware platforms,

operating systems and core applications is efficient and cost effective. However this construct creates tempting challenges for attackers. Reverse engineering or penetration of one system can lead to a mass effort to penetrate the other similar systems.

[0228] The proposed centralized manage network will possibly create single points of failure, which are also identical technologically, standards wise and operationally.

[0229] DigitalDoors delivers a security solution that enables the said planned centralization yet avoids single point of failure by decentralizing critical granular data within the network.

[0230] DigitalDoors delivers perpetual data availability and survivability capabilities

[0231] Best defense against denial of service attack

[0232] Resiliency via granular replication with minimal security risk

[0233] - Highly secure content storage and delivery within survivable distributed networks

[0234] - Encryption and decryption of data streams including granular segments

[0235] DigitalDoors delivers perpetually available and secured data, by a process which divides and parses data streams (documents, e mails, etc).

[0236] The encrypted content pieces are spread in different storage servers. Because of their granularity, there is no security risk in replicating those pieces in many servers. The replication and storage of those pieces on different storage servers delivers resiliency against attacks including denial of service attacks. Attack or compromise on a few storage servers will not cause harm to the integrity of the whole document or data stream.

AAA.1.35.0 Processing

[0237] Maximize data throughput, minimize data overheads

[0238] Support core legacy operations

[0239] Support above the OSI Application Layer in terms of data, metadata, and usage biases for contextual security

[0240] Support for any interpretable data stream (ASCII, signal, image, etc.)

[0241] The network – centric benefits:

1. Provides users down to the last tactical mile, with pull capabilities whatever they want, whenever they want from wherever they are.
2. Enable data posting to COIs and other common storage spaces instantly
3. Deliver cross domain interoperability among allies, coalition and multinational partners
4. Deliver real time information and information changes to the battlefield and edge users.

5. Deliver highest level information assurance / security

[0242] DigitalDoors supports multiple heterogeneous sources of data and

[0243] Multiple data formats. (It conforms to many of DISA's requirements: Multiple platforms. Any type of data file or stream. Multiple languages. ASCII,EBCDIC,UNICODE. Supports high-transaction-rate databases, legacy host mainframe applications, inventory, payroll, and financial applications, MAN/MAN, resource planning, Internet storage, Web-based dissemination with push and pull, and other COTS software and activities. Support and augment COTS desktop applications such as Exchange, Outlook, and Office (such as Word, Excel, PowerPoint, Project, Explorer.)

AAA.1.36.0 Providing scalability, reliability, resilience, fault tolerance, speed

[0244] DigitalDoors is designed to deliver perpetual data availability. It is a distributed granular information system that will provide availability reliability resiliency and fault tolerance. Data is split into small tagged segments, encrypted, replicated and dispersed.

[0245] The reliability resiliency and survivability of data is enabled by eliminating single points of failure. Sensitive data is split and cryptographically distributed across various servers. Compromise of one server or a few servers will cause no damage or a minimal damage.

[0246] Data bases, encryption keys and other core capabilities are also dispersed to avoid single points of failure. The risk management capabilities of the system enable users and administrators to gage and control the level of risk by setting the level of granularity. The design for all process, all inputs, all outputs, and storage is granular. This addresses the core issues of scalability, reliability, resilience, survivability, fault tolerance, speed, mobility, reconfiguration, and flexibility.

[0247] Implementation and infrastructure is designed with the assumption that systems, servers, network, user devices, and channels will be disabled by natural disasters, management and user accidents, willful intent, sabotage, and direct and asymmetric warfare. As a result, DigitalDoors can be partitioned in as many granular components and backups as desired, not only for scalability but also for resiliency and survivability.

[0248] Most DigitalDoors processes are neither processor- bound, disk-bound, or limited by network performance even on low-end PCs. Therefore, CPU performance is rarely an issue. However, when vast legacy archives of documents need to be assessed, downgraded, and distributed, this processing can be distributed since each thread is independent. This means that the DigitalDoors process is inherently scalable subject to the linear nature of standard textual documents. The legacy conversion process would be best distributed on a per document basis in order to process content and metadata within context.

[0249] DigitalDoors is designed as highly scalable system, that enables publish & subscribe communications.

Best Available Copy

[0250] DigitalDoors enables different methods of publication.Automatic and or manual publication of files, and/ or sanitized files. Meta data publication and/ or publication of sanitized of Meta data (including Meta data directories, catalogues)

[0251] Subscription communications are enabled with unique granular capabilities.

AAA.1.37.0 Enabling generic and standardized data distribution services

[0252] DigitalDoors enables generic and standardized data distribution services. The system delivers major advantages by implementing granularity into its distribution services. Granular distribution enables sharing information based on a match between the labels and the security profile of a user. Granularity is supported to the sensitivity/security, compartment, information, and category levels. Generally, DigitalDoors supports incoming and outgoing data flows for ID, authorization, and integrity checks and configure selection, extraction, and dispersion to support each data and format type, and any metadata or catalog security issues.

[0253] Relates to architectural initiatives (e.g., shared software services, Web, XML, etc.\ Inherent to the open and granular design. DigitalDoors has demonstrated applications with HTML, XML, SOAP, and various COTS applications.

[0254] Operates in a network-centric architecture

[0255] DigitalDoors design for data granularity and process resiliency and fail over presumes functionality with generic user interfaces, black-box processing, web interfaces, integration with COTS products, and integration with legacy processes. DigitalDoors coexists within an extended network, with mixed platforms, and various architectures, from appliances through mainframes.

[0256] Supports event modes

[0257] DigitalDoors has built-in triggers to initiate data protection or terminate data access based

on day or week, time of day, load levels, thresholds, alerts, events, security breaches, and user-defined situations. Access to information is proscribed to a lowest common denominator, or multiple levels simultaneously. This could preclude access--except to a template, some minimal partial display, or full display, with anything configurable in between. Upgrading of information is granted by network-centric or local ID, authorization, and other superior officer/manager overrides. This is granular control at each of the sensitivity, the compartment, the information label or tag, and categorization levels.

[0258] DigitalDoors has added support from its initial plain-text to include encryption, HTML, XML, metadata, dispersion, multi-level display, biometrics, and data mining. R&D drives technical enhancements, but user requirements and understanding how to implement user needs drives design changes and innovation.

[0259] Future DigitalDoors applications will be supplemented by tested cutting edge technologies, which are highly important for collaboration and security. Such technologies may include categorization technologies, semiotic inferences, list processing etc.

AAA.1.38.0 Storage

[0260] Storage overhead is minimal for DigitalDoors. The storage requirements are limited at the upper end by the distribution of the core application library. The upper bound is probably a standard CD. However, the databases for inclusionary and exclusionary categorization technologies is driven by the number of languages, image types, audio samples, and complexity of inferences needed for the user environment. For example, American English requires roughly 60MB.

[0261] The recovery information and downgraded document is generally stored just once for all levels, compartments, labels, and categories. Sensitivity levels, compartments, information labels (tags), and categorization is very efficient, thus yielding a raw storage requirement of about 106% of the

original legacy data stores. Encryption and compression can alter this substantially downward, to an average of 30 to 50% depending on the compressibility of the raw data.

[0262] On the other hand, storage is a direct factor of the number of sensitivity levels, compartments, and labels when multilevel storage is maintained. Backups, dispersion, and redundancy is a direct factor as well. The granularity is critical to issues of failover, local performance, resiliency, survivability, and fault tolerance.

AAA.1.39.0 Metrics

[0263] The DigitalDoors process is analogous to string processing. Performance is comparable to opening a text file, manually editing a text file, and saving that file to a remote secure storage. This is typical and traditional mainframe, server, and desktop overhead, When advanced techniques are applied for inclusionary and exclusionary categorization of data, metadata, and complex hierarchical data structures, more CPU, disk, and network overhead is required. However, when the task becomes bound, it can be pipelines, partitioned, run in parallel, or accelerated by hardware.

[0264] COTS Benchmarks of the DigitalDoors process shows performance that ranges from 34,000 to 360,000 instructions per recognizable element in the data stream. This ranges from 56 to 783 ?sec on IBM Thinkpad laptop yielding a Winstone value of 23. This would correspond to sub-millisecond response times on a IBM 3700 and like hardware.

[0265] When security is established by dispersion, distribution, and replication over a wide-area network as proposed by the GIG-BE, the single most significant component in response time will be signal response time. Signal response time in an Internet-like grid infrastructure is purely a factor of signal propagation time, gateway/router hops delays, firewall lags, and multiple server processing queu(e)ing waits. Satellite response times of 1.5 seconds, and router hops with 10 or more reroutes at

0.300 ?sec drives the GIOG metrics of multi-second response times. DigitalDoors application performance times are presumably immaterial in contrast.

AAA.1.40.0 Technical specifications, operating parameters, and network design specifications

[0266] DigitalDoors is a series of differential process step that can occur on any platform, ranging from a mainframe to a low-end laptop. It enables content-level security in any type of data stream on any type of platform. It is inherently scalable from single to unlimited users. Content-level security can be effected at any layer of the ISO OSI model, from the bottom physical layer (level 1) to the top application layer (level 7). In addition, DigitalDoors uniquely functions within an application layer data stream to analysis content, concept, and context for security sensitivity to protect again brute force inference with statistical methods and data mining, assemblage of information, and human intelligence methods. In other words, DigitalDoors is processing any type of data stream for potential security lapses.

[0267] Most deployments are likely to require enterprise operation support with mixed client/server architecture. Automated OPSEC and COMSEC processing can occur as background processes on virtually any computing, Web/Internet, and networked platform. The DigitalDoors architecture was designed with run anywhere mentality, including hardware chipsets. However, in most cases, manual tagging presupposes a graphic user interface for selecting and identifying data stream content. As such, the core technologies require only visual element outlining, automated content identification, list lookup and matching, and basic data processing. Processing can be offloaded to the client or performed on the server. Dispersion is a function of a secure networking infrastructure. Most operations occur at network wire speed, this being the primary performance limitation.

[0268] DigitalDoors has installed and demonstrated its effectiveness on currently deployed

COTS environments. This includes a wireless laptop network with a wireless SAN, a 10BaseT Intel client-server architecture, a 100BaseT and Gigabit Intel client-server architecture (Win95 through Win 2003 AS and W2K EAL-4), several MS Terminal Server environments, and integrated with Solaris v8 and v9 and Trusted Solaris v8 servers and clients. We have tried the technology on Palm IV; we assume that DigitalDoors is portable to PDAs, handheld devices, cellular phones, and other hardware devices subject to their memory, connectivity, and CPU prerequisites. For point of reference, we have integrated and demonstrated DigitalDoors with MS Windows, as add-ins for MS Outlook and MS Exchange, and presume simple functional integration for any MS Office, database, data workflow, signal transmission, or COTS products. We have demonstrated DigitalDoors with ASCII files, MS Word documens, HTML, and XML metadata.

[0269] The client requires only a basic GUI interface (at a minimum) with NetBEUI, NetBIOS, or IP networking over any network transmission medium. The server components require no GUI and can run as a background process with NetBEUI, NetBios, or IP networking over any inter-compatible network transmission media. Interprocess communications is enabled through client/server channels, OS messaging, and RPC. Support over LAN, WAN, Internet, VPN, NAS, SAN, with remote commercial storage services has been tested. DigitalDoors has also been successfully tested with encrypting network hardware (Intel) and native P2P encrypting protocols.

Table: The Data Declassification process is defined by these steps:

Authorization of security levels

Sensitivity

Compartment

Information

Categorization

Authorization of security roles

Sensitivity

Compartment

Information

Categorization

Identification of a data stream

Identification of the format for the data stream

Configuration of intended security levels

Sensitivity

Compartment

Information

Categorization

Categorization of data within the stream (e.g., tagging)

Selection of data within the stream for extraction

Extraction of data within the stream

Downgrading of data stream

Creation of upgrading recovery data files

Data backup

Data replication

Data dispersion

The Data Reclassification process is defined by these steps:

Authorization of security levels

Sensitivity

Compartment

Information

Categorization

Authorization of security roles

Sensitivity

Compartment

Information

Categorization

Access to downgraded data stream

Upgrading of data stream

[0270] Access security is established through a network ID and authorization process, such as Kerberos and enhanced with the compartmentalized trusted operating systems, such TSOL. Access security can be enabled for most networked devices.

[0271] Downgrading the data stream is manual or automatic, or assisted. Automatic OPSEC and COMSEC rules can be enforced prior or post manual selection (tagging of the sensitivity levels, etc.) The processes are asymmetric by design. Downgrading can require substantial time, but upgrading is limited only by the network and the time lags to deliver a signal cross-country through multi-hop switches and routers that are part of the commercial public infrastructure. Realize that firewall, guard, filter, and router hops frequently entail 300 ?sec unidirectional delays; this being the primary performance delays. DigitalDoors has designed to the architecture and the process flows such that steps

are optimized to each side of the firewall, guard, filter, router, and storage server mesh, including even optimization at the client for the extensive high-level content, concept, and context identification and dispersion. Manual tagging can be performed on any computer that supports a GUI interface. Processing is bound to disk or network performance limitations. Tagging information can be processed locally or delivered to a server for service-side processing. The user's ability to locate and highlight sensitive material is the functional performance limitation.

[0272] The data stream can be any of type of data. This includes signals, files, data streams, interprocess messaging in a workflow or high-volume transaction processing environment, email, instant messaging, logical information, and COTS data files. Data streams can be free form or structured, such as a COBOL data set, a RDBMS database, an MS Word document, or metadata, such as SGML, HTML, or XML. The DigitalDoors philosophy is to enhance security while maintaining traditional process workflows. Note that encrypted files and data streams require discrete and complete decryption before they can enter the transactional processing. Failure to decrypt such files abends systems and workflow because the format and content is unexpected. In addition, while encrypted XML and metadata is secure within that form, once decrypted is usually carries far more security-sensitive content than realized.

[0273] Manual processing requires trivial overhead. CPU requirements are minimal for the client and server components. However, OPSEC, COMSEC, and enhanced automatic and assisted data selection requires additional CPU resources and approximately 400MB of storage. Even laptops are suitable. Automatically or assisted tagging information can be processed locally or delivered to a server for secure processing. Processing is bound to roundtrip network performance limitations, and database lookup seeks times. As the complexity is increased to automatically search data, metadata, content, context, and concept for sensitivity within the data stream, performs slows. Performance is affected by

disk seek times and the depth of the searching and DigitalDoors categorizing process. Performance ranges from 34,000 to 360,000 instructions per recognizable element in the data stream. This ranges from 56 to 783 ?sec on IBM Thinkpad laptop yielding a Winstone value of 23. DigitalDoors processes a document of 140 words in 70 seconds on that laptop, but within 19 seconds on Dual Intel CPU Server with SCSI drives (Winstone value of 89). A production SQL database server provides several orders of magnitude better lookup performance than a client database on a laptop; it also scales better. Experience with Sequoia or comparable scalable and fault-tolerant architectures indicate single digit ?sec response times.

[0274] Dispersion performance is a function of network channel responsiveness and network storage performance characteristics. Wireless Ethernet 11Mbits/s down to 2.2 Mbits/s is sufficiently faster than database seek times and storage, but nonetheless effective in demonstrations.

[0275] ReClassification performance occurs at wire speed subject to the performance limitations of network infrastructure and remote network storage performance. We have recovered 2 MB documents within 3201 ?sec using a multiple-hop router network and 10Mbits/s Ethernet hubs to Intel desktops. This is equivalent to raw file display on the same networks and systems.

[0276] The footprint is purposefully small. Storage requirements include application storage, databases, temporary cache, and secure and dispersed network storage. Distribution is typically provided on a standard CD (640 MB) or business-card CD (120MB). User storage per user is as little as 15KB for a configuration profile. Application storage is approximately 30MB for clients, although this can be shared as a single store on a server or terminal server. Database storage requires approximately 231MB. This includes all language dictionaries, categorization elements, synonyms, and semiotic and grammatical inferences. Transactional processing, rollbacks, and security logs are optional, but on the

same order of magnitude. All storage can use SSD, USB, PCMCIA, or flash RAM or networked devices. Performance is enhanced about 45 times with SSD. Flash RAM is marginally faster than mechanical hard drives due to lack of device driver optimization.

[0277] Data storage for user documents, data sets, data streams, metadata, and the like will be comparable to the size of the existing data store. Note that a design criteria is that all extra-security information, recovery files, auditing, and logs are stored in separate channels both for security reasons and format preservation. Expect additional overheads of only 6% for dispersion of recovery files. However, realize that minimum block sizes could increase that. Optimization for recovery files, which tend to be small, is possible by setting block sizes to 1KB or smaller on a special recovery file storage site.

[0278] When sources are dispersed and replicated for multiple location storage, or creation for individual TS, S, and C archives with support for compartments, information labels, and categorization, then each replication obviously increases the storage requirements by about 100%. Proxy caching is effective when documents are often requested; however, for security purposes, caching of recovery files is insecure. It its also ineffective unless the recovery files are bound to the same sensitivity/compartment/category user because the storage server generates and only delivers at the level dominated by the user clearance.

[0279] DigitalDoors assumes that the native OS encryption will be applied or that a certified encryption system is available within the network infrastructure. DigitalDoors will integrate with that facility. Encryption overheads are comparable to what exists now. When unclassified or downgraded documents and the recovery files are stored black and delivered black, the encryption overhead will be

approximately 106% of what is now. In spite of that low overhead, this provides infinitely granular data storage and infinitely granular upgrading.

[0280] Interprocess communication and messaging for client/server is optimized to minimize traffic loads and point-to-point encryption overheads. This limited performance overhead is most evident when a terminal server environment is taxed by many users, who nonetheless see performance equivalent to that performed on the client desktop. User or process identification and authorization is interspersed throughout the DigitalDoors process—it is not a one time when the user logs in—so there is some ongoing validation chatter. This rate is granular and configurable by the security administrator on a range from seconds to hours.

[0281] There are true performance and storage benefits when downgraded documents are shortened and mechanically compressed. However, for functional reasons, a core design of DigitalDoors is that the data stream format is typically preserved intact (less extractions or substitutions) to allow for workflow processing without abends, interrupts, or crashes. Even HTML, XML, and other metadata are typically preserved to retain the original format so that DigitalDoors is providing security without compromising functionality.

AAA.1.41.0 DigitalDoors Operating Parameters:

[0282] Any CPU platform with a development platform and as little as 64KB of RAM. Any platform with support for either C, Java, or VB is suitable. DigitalDoors can run as a hardware process, an applet, an add-in, process filter, a dynamic link library, as an NT or RPC service, as a standalone client, as an in-process or out-of-process server, and as an enhancement to a firewall, IDS, IPS, guard, or operating system. DigitalDoors interacts with Kerberos, Active Directory, smart cards, biometrics, encrypting file systems, trusted operating systems, and many NAS, SAN, and other remote filing

systems.

AA.11.42.0 DigitalDoors Network Design Specifications:

[0283] Because the core concept of DigitalDoors is to support collaboration, information sharing, dissemination, Internet and Web services, data mining, and data dispersion to remote storage, the client/server infrastructure requires only a core network. Integration preference is for a COTS network environment, such as Internet, Intranet, LAN, MAN, WAN, NetWare, NFS, or Microsoft networking, Network protocols can include NetBEUI, NetBIOS, and IP on a physical channel of 1Mbits/s wireless (Zenith), infrared, Bluetooth, 802.11a or 802.11b, or 802.11g, Ethernet, FDDI, iSCSI, frame relay, ISDN, DSL, T1, T3, OC3, and SONET

AAA.1.42.0 Configuration options available for each product/s

[0284] DigitalDoors enables a wide variety of configurations according to the needs of different networks. Since the system introduces granular marking and labeling and securing of granular data segments the spectrum of configurations options is big. Those configurations can be set per organization policy at set up minimizing future administration work.

[0285] Following are some possible stand alone configurations:

Table of Configurations:

Digital content rights management- granular content control

Downgrading documents through manual, manual and automatic corrections, and automatic modes for sharing, coalition interoperability etc.

Cross Domain sharing high to low, low to high when connected to DTW certified guard

Real time total content analysis and categorization (headers and payloads) all OSI layers

Data activities: Processes, creates, extracts, encrypts, disperses, reconstitutes, inserts, duplicates, erases and audits

User created filters

Search for any characters

Manual marking, labeling and securing

Manual marking and labeling with automatic COMSEC override

Automatic marking labeling and securing

Manual posting and broadcasting

Automating posting and broadcasting

Encryption of content segments

Multi Encryption of content segments (different types of encryption within a document)

Information operations- misinformation

Passive auditing

Data replication

Data rerouting

Data compression

COMSEC

OPSEC

With add-ons- Intrusion detection

With add-ons firewall

[0286] Many other different granular defenses-in-depth configurations are available. Those include granular segments replication and their dispersal to many servers to defend against denial of

service attacks.

AA.1.0.0 Background: Formlessness for Security

AA.1.1.0 Introduction

[0287] Formlessness is the old, new, and renewed paradigm for security. Formlessness is part of ancient military strategy. Of old, it provided invisibility, fear, stealth, cunning, capability, and power. Of new, formlessness defeats ID theft, espionage, diversion of data streams, and the unfettered distribution of sensitive data now overwhelming every aspect of society. Security is badly-served by existing systems and methods, in particular, by monolithic perimeter barriers or by accretion of point solutions. Formlessness is the instrumentation to renew security and also competitiveness. Formlessness enables the architecture for dispersion, which provides for arbitrary levels of semantic and granular content control.

[0288] There is little substantive in literature about formlessness. There is even less about its relevance to security. The few western citations on formlessness include Sun Tzu, Robert Greene, and references by others to Herodotus of Halicarnassus and the generals Xenophon, Julius Caesar, Vegetius, and Von Clausewitz. Mao Zedong wrote on the board game of weiqi (aka Go) and its relevance to military strategy. Miyamoto Musashi wrote A Book of Five Rings (Go Rin No Sho) in the 1600s. This is a Japanese extrapolation on the Chinese The Art of War (Sun Tzu) to Samurai and the different culture of business from the view of a person experienced in the cultural and military upheavals of the time. See: References to weiqi as an <> to contemplate formlessness are referenced by:

Himly 1896 : Karl Himly, "Die Abteilung der Spiele im `Spiegel der Mandschu-Sprache'" [The section on games in the 'Encyclopaedia of the Manchu language'], T'oung Pao, Series II, Vol. 7, pp. 135-146 (E. J. Brill, Leiden, 1896).

Karlgren 1951 : Bernhard Karlgren, "Excursions in Chinese Grammar", Bulletin of the Museum of Far Eastern Antiquities, No. 23, pp. 107-133 (Stockholm, 1951).

Koizumi 1991 : Koizumi Shingo, "Chugoku Kodai no Yugi Ryupo ni tsuite" [On the game liubo in ancient China], Yugi-shi Kenkyu No. 3, Oct. 1991 (Osaka).

Kotaka 1943 : Kotaka Yoshisaburo, "Nihon no Yugi" [Games of Japan] (Hata Shoten, Tokyo, 1943).

Lau 1965 : D. C. Lau, "Some Notes on the Sun-tzu", Bulletin of the School of Oriental and African Studies Vol. 28, pp. 317-335 (London, 1965).

Li 1980a : Li Songfu, "Han-Wei de Weiqi" [Go in the Han and Wei periods], Wei Qi No. 99 pp. 27-29, No. 100 pp. 27-29, No. 101 pp. 18-19 (Shanghai, Jan.-Mar. 1980).

Li 1980b : Li Songfu, "Sun-Lü Yipu de Zhenwei" [Is the Sun-Lü game genuine?], Wei Qi No. 103, pp. 51-53 (Shanghai, May 1980).

Liu 1985 : Liu Shancheng, ed. "Zhongguo Weiqi" [Go in China], 2 vols. (Shu Rong Qiyi Chubanshe, Chengdu 1985).

Masukawa 1983 : Masukawa Koichi, "Tobaku III" [Gambling, Vol. III] (Hosei Daigaku Shuppankyoku, Tokyo, 1983).

Masukawa 1987 : Masukawa Koichi, "Go" [Go] (Hosei Daigaku Shuppankyoku, Tokyo, 1987).

Murray 1952 : H.J.R. Murray, "A History of Board-Games other than Chess" (Clarendon Press, Oxford, 1952).

Needham 1962 : Joseph Needham, "Science and Civilisation in China", Vol. 4 Part I, pp. 314-334 (Cambridge University Press, 1962).

Ogawa 1932 and 1933 : Ogawa Takuji, "Shina ni okeru Iki no Kigen to Hattatsu" [Origin and development of go in China], Shinagaku, Vol. VI, No. III (July, 1932) pp. 57-76, and Vol. VII, No. I

(May 1933) pp. 79-90 (Kobundo Shobo, Kyoto).

Potter 1985-86 : Donald L. Potter, "Go in the Classics (ii): the Tso-chuan", Go World, No. 42, Winter 1985-86, pp. 19-21 (Ishi Press, Tokyo).

Shen [undated] : Shen Zicheng, "Weiqi" [Go] (Guoguang Shuju, Hong Kong, undated).

Sawyer 1993 : Ralph D. Sawyer, "The Seven Military Classics of Ancient China", (Westview Press, Oxford, 1993).

Shirakawa 1993 : Shirakawa Masayoshi, "Igo no Genryu wo Tazunete" [Seeking the origins of go], Igo Kurabu, No. 10 (October 1993), 136-138 (Nihon Kiin, Tokyo).

[0289] More is accessible to scholars in original languages. Nevertheless, there is a western reference to Kant's Esthetic Theory, in relationship to Air Force pilot John Boyd when describing his OODA loop, or academic discussion of 4th and 5th generation warfare (4GW and 5GW) strategies. In this country, some of these oriental techniques are applied to modern risks. LAX airport has implemented security that applies the theories of formlessness through the randomized distribution and rounds of security guards, although this thesis refers more specifically to the Bayesian-Stackelberg Theory instead of forecasting randomness or asserting randomness to confuse and mask. For the most part, formlessness is a concept that gets rare mention but represents a paradigm shift necessary to implement security, privacy, confidentiality, and survivability within our information ecosystem, the competitive economic marketplace, and warfare.

AA.1.2.0 Definition

[0290] The common definition of formlessness is the state of being amorphous, without shape, or having an indefinite shape. Think of liquid within a container as characterizing formlessness, assuming whatever shape the container compels it to be. Formlessness is usually a philosophical or

religious term linked to the ultimate, infinity, indescribability, and transcendence. That also is without a real relevance to warfare, security, or business competitiveness unless formlessness is given utility.

[0291] The definition, of relevance here, is that formlessness is the condition of a resource (and more specifically, information as a resource) without context or utility. Consider formlessness as stealth, ambiguity, and surprise when applied to functional security. Formlessness is not chaos. Formlessness is not randomness. Formlessness is not emptiness or nothingness. Formlessness is not a vacuum. Formlessness is not order or the sum of everything. Formlessness embodies all those elements. Yet, it is not that; it none of those things, but all of those things. Emptiness and nothingness are based on a context, such that a room with emptiness and nothingness could have alternative value. It might not even be empty, merely have appearance of that.

[0292] Formlessness suggests form, or the expectation of form. Formlessness does not have meaning without the contradiction of form, structure, expectation, utility, usage, value, or purpose. Perhaps the real state of formlessness is the reversibility from order and complexity into chaos and simplicity, thus given rise to a two-way production model that converts raw materials into wealth and the deconstruction of wealth back into raw materials, primary of which, is energy. Albert Einstein has shown us that energy and matter are interchangeable. As such, formlessness could reflect the degree of available contextual information embodied by either energy or matter. Without form there is limited recognition of formlessness. Degree of formlessness might best be described as the amount of potential in contrast to the level of complexity.

[0293] Formlessness is a contradiction. It is nothing, yet it is everything. Formlessness is content without definitive context or concept. The content alone can have value, but not necessarily. The content requires context in meaning, a concept of purpose or utility to define process or attribute intention.

Formlessness is a definition without a meaning, a broken tool, or a power leak. Formlessness is a coherent concept with a very long history of practical value. In more practical terms, formlessness is management of dispersed resources, power, assets, people, information, and elements of data streams so that the elements lack value without context.

[0294] The very existence of formlessness is predicated on the context or the concept, or the lack thereof. When semantic or semiotic clues are lacking, the situation takes on a chaotic aspect. When patterns can be perceived within this cloud, the formlessness resolves into a purpose, a process, a resource, a workflow. Areas within the cloud might still be perceived as chaotic, might not be important to the recognized patterns, or might still be unresolved. With the right control, these too will resolve, or remain formless as a protective barrier to market entry or a buffer zone against adversaries. Formlessness is a utilitarian tool for both defensive and offensive tactics, and a source to instill confusion when its patterns are lost or used against us. Such a tool has most value when efficiently used and applied correctly. That requires recognition for its applicability and experience for its proper (and evolving) usage.

[0295] This concept is antithetical to western philosophy. It just does not represent a useful concept for most people. It is even scary for some people. Others, just shrug it off, and ignore the idea for lack of getting the concept. This concept, however, is important, practical, and very much hard to grasp. It is very much a part of nature. It is the motion of fish in schools. It is the internal rhythm within herds on the African plains. It is the creation of chemical molecules that self-organize as DNA molecules or RNA processes. It is the construction of a beautiful picture from pieces of ugly fractal squiggles. A clash of cultures and embarrassment in verbalizing formlessness inhibits the understanding of formlessness. As soon as the concept registers, it sticks with a person. Everyone gets it. If you feel lost,

you will get it too.

[0296] Formlessness is the separation of meaning into chaos. Formlessness is the correlation of seemingly random parts into a coherent whole. Formlessness is the time before the Big Bang and formlessness is the time after the big bang, which is the state of the universe in which we currently live. While there was a form and order to the universe before the big bang, there is a different form and order now. The contexts are merely different, creating different opportunities and existences, and with a relationship between them. Formlessness is a perception of context, of other contexts, and the application of the confusion of a context.

[0297] Formlessness unravels the value of information or resource to a lower level, an application of the 4th law of the Theory of Thermodynamics, in that everything runs down, has no heat, and lacks differentiation. Formlessness is useless unless it can be reversed. Think of it as reversible entropy, or as some call it, reversible redaction when applied to information. If it can be reversed, then it can be used for security, privacy, confidentiality, and survivability. Reversal is reconstruction with integrity, reliability, survivability, security, authentication, authorization, trust, and polling against the efforts to undermine the dispersion of formlessness. The value of a resource is degraded by:

disassembling its whole into parts that have no value except with the other parts

distributing those parts to prevent reassembly

Intermingling parts with other parts

lack of content, context, and concept

[0298] This embodies that lack of differentiation in thermodynamics (2nd law). When parts are interchangeable, security is only achieved by protecting the parts in lockdown. Information is usually not interchangeable; it is usually specific to a context. If information is interchangeable, lockdown of

the whole or the parts through encryption works and is the simplest solution to date. When process complexities require matrix-level access controls or uncontrollable downstream distribution, DigitalDoors' granular content control is simpler and more flexible. It resolves the inherent conflict between security and functionality.

[0299] Under the concepts of Claude E. Shannon, information theory regards information as only those symbols that are certain to the receiver.

[0300] A unifying theory known as information theory was developed and became the subject of intensive research. Claude E. Shannon, whose initial ideas appeared in the article "The Mathematical Theory of Communication" in the Bell System Technical Journal (1948) defined information to include the messages occurring in any of the standard communications media, such as telegraphy, radio, or television, and the signals involved in electronic computers, servomechanism systems, and other data-processing devices. The theory is even applied to the signals appearing in the nerve networks of humans and other animals. The signals or messages do not have to be meaningful in any ordinary sense.

[0301] The chief concern of information theory is to discover mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store, and otherwise process information. Some of the problems treated are related to finding the best methods of using various available communication systems and the best methods for separating the wanted information, or signal, from the extraneous information, or noise. Another problem is the setting of upper bounds on what it is possible to achieve with a given information-carrying medium (often called an information channel). While the central results are chiefly of interest to communication engineers, some of the concepts have been adopted and found useful in such fields as psychology and linguistics. The boundaries of information

theory are quite vague. The theory overlaps heavily with communication theory but is more oriented toward the fundamental limitations on the processing and communication of information and less oriented toward the detailed operation of the devices employed.

[0302] The Shannon information theory descries information as symbols within a stream and self-obvious to the receiver. No external information is presupposed or allowed, other than -- and Shannon does make assumptions -- a format, a coding method, and a meaning structure, just not a knowledge system. He excludes the inference of information, relationships to other information and symbolic expansion, or an extended philosophy with pragmatic, functional, intellectual, cultural, developmental, or conflicting bases. However, this is not enough on which to base information measurement. This is specifically relevant within this discussion of formlessness. Just because the symbols are uncertain to the receiver does not mean that sender is in any way uncertain about the information, or that there is another intended receiver different from the actual receiver for whom this message is actually directed. This disparity opens up sending covert messages in the clear, meaning hidden in plan sight, or information with greater value than obvious. Security that does not consider this broader extension of meaning is security with leaks. This specifically is relevant to issues of aggregation, inference, and interaction, how the receivers interpret the context of formlessness, and also in the assessment of structured data and unstructured data because structure is context as well.

[0303] Although the concept of formlessness might be uncommon within western literature, real world examples abound. Any business with mobile service vehicles, like HVAC or plumbing repairs, exemplifies master control over swarming technical people. Both server and storage virtualization is the reorganization and management of formlessness. The logistics of shipping, fulfilled by FedEx and UPS for example, detail a pragmatic result with an obvious level of transitional process confusion; this too

is a practical application of formlessness. Additionally, current agile software development techniques stress aspects of formlessness, including the flexibility of code/design development cycles, distributed work efforts, and reuse of code.

AA.1.3.0 Formlessness

[0304] Only formlessness that obscures content, context, and concept is sufficient to provide effective security. This is a direct contradiction to security by obscurity; obscurity is a subset of formlessness that does not inherently provide effective, long-term, and pragmatic security. Obscurity by knowledge restriction is weak because knowledge has a way of leaking out. Ignorance is not a quality we would realistically associate with a competitor. Security by obscurity is based on ignorance, and that is not a viable assumption.

[0305] Dispersing information to a masked, contracted, truncated, or hidden location is a form of security by obscurity. It is a case of concealing information to prevent its use or reuse. Of course, it works to a point, but once the location is reverse-mapped, the camouflage is no longer effective. Often methods are invented to defeat this camouflage, or things that seem to work become institutionalized, like IR film or putting color-blind spotters in airplanes. It is only a matter of time, real time and not non-linear exponential overhead, posed by encryption and/or access control systems

Furthermore, the obscured file frequently is matched to attributes for ownership, date, and usage and these expose any basic attempts to mask its value, intent, purpose, and security risk. Additionally, the obscured file, whether structured or unstructured, typically contains within various forms of metadata that expose intent, purpose, and ownership, thereby disclosing the very information supposedly obscured. Obscurity is a subset of formlessness, but not an inclusive one. The file obscured still retains integrity, context, utility, and reversible attributes that allow for reconstruction of the obscured location and the

intended utility without special tools. Not only is the supposedly-obscured file in plain view, but its key to recovering the obscured location is also in plain view. The content and form is intact, and contains enough sensitive information and context to reestablish is rightful location and utility.

[0306] Unfortunately, most operating systems contain kernel-level tools or are enhanced with all manner of search utilities that map the file system. These tools are designed to restore file losses and restore order. These range from more esoteric system administration lost block and segment tools, defragmentation, to disk repair tools. User-level tools walk the directory hierarchy, finding all files, the path to all files, disk space utilization, quota management, to reporting tools that map usage by location, creation user, user of last date, date of creation, name, type, and so forth. The bottom line is that such tools expose any obscurity or intent to affect security by obscurity.

[0307] Dispersing intact information or a file across multiple systems, attached storage systems, remote storage, Internet storage, and information fabrics although complicating the level of obscurity in reality does not alter the basic problem. The hiding of information in plain site or mixed up with other like files does not obscure the basic attributes stored by the file system. It can make it difficult for the owner (likeliest user) to recover a file, but it will not in any deter a motivated system administrator or hacker from remapping and classifying the type and value of information contained on one system or a network of such systems.

[0308] However, if information within a file is categorized, selected for PII or security sensitivity, and extracted as elements for dispersion, thereafter those elements lack context and recovery without a detailed map. Since this breaks the basic or extended attributes of most file systems, dispersion works if and only if access to a map is controlled, access to dispersed information is controlled, and the dispersion is sufficiently random to preclude forced re-identification of parts with the source or parts

with parts. Granular access control predicates effective dispersion security; it is not knowledge-driven or security by obscurity.

[0309] In conclusion, obscurity is a subset of formlessness that does not inherently provide effective, long-term, and pragmatic security. Obscurity by knowledge restriction is weak because that knowledge itself has a way of leaking out. Only formlessness that obscures content, context, and concept is sufficient to provide effective and pragmatic security.

AA.1.3.1 How DigitalDoors Applies Formlessness

[0310] DigitalDoors enables information formlessness with reversibility, extraction of critical elements with reconstitution (or reconstruction) to the original value. This provides formlessness without any utility. Reduced utility is a practical matter in controlling the degree of formlessness at any value less than total; that value being a measure of control over granularity. True formlessness requires the complete separation of form from function, the reduction of meaning from context and purpose with what DigitalDoors calls dispersion, and the complete erasure of conceptual meaning that is less than arbitrary or indiscriminate. Utility of formlessness requires that even such arbitrary parts retain some designation, such designation remaining arbitrary and indiscriminate, to affect what DigitalDoors calls reconstitution, and others such terms as reassembly, recomposition, reconstruction, inline translation, or reversible redaction.

[0311] Dispersion increases the scarcity of information by creating formlessness, thus increasing its security and its relative value through scarcity. The information no longer exists, although it is known to exist or once existed, and this formlessness is reversible. Scarcity is also another important factor both in security and in the economics of wealth production models, and scarcity is manipulated effectively through dispersion.

[0312] The other application of formlessness is the recognition that dependencies apply to destabilize the survivability and security of the wealth production methods and the efforts to decouple these relationships. The relationships exist as contexts and concepts. When they are decoupled, we reduce the inherent fragility of the overall ecosystem. Specifically, DigitalDoors applications do not require specific platforms or a defined network infrastructure. Appliances that provide processing and appliances in the cloud for storage are sufficient. Most production processes are complex enough as it is, so the recoupling of rigorous and specific infrastructure requirements from process and information (through dispersion) increases overall reliability, survivability, and security.

[0313] When formlessness creates parts that do not contain within themselves the context for reassembly, the parts can be stored together. When formlessness contains assembly clues or instructions, the parts must be scattered or individually protected. When structure is removed from information, the interpretation of the information is uncertain or unlikely, thereby imparting formlessness. When formlessness is combined with randomness and a loss of differentiation through large number of like parts, parts can be stored together if there remains a way to differentiate individual parts uniquely. As such, the nature of the parts determines the level of individual granularity, and whether parts must distributed uniquely or can be saved together in the same storage.

If the purpose, strength, and intent is hidden in the multitudes of options within formlessness, the defense and the offense is masked until the actual attack. Consider the text from Punctuation--Context alters Concept stripped of its structure, format, capitalization, and punctuation.

[0314] Other formats, like the grid or a database, show different geometric contexts. Could this text have a vertical or diagonal message too? Could other forms preclude some words over a pattern or filtered window that hides some words like red letters masked through a red filter? Could the message

be read through an algorithmic pattern, skip one word, next word, two words, next word, skip three words, next word, and so forth? Formlessness enables dispersion, sets a pattern for breaking contexts and assigning new ones, and allows dispersion to secure meaning.

[0315] The hallmark of formlessness is the uncertainty or ambiguity within a data stream processed by DigitalDoors Whether an extraction is masked without the use of a placeholder or a placeholder marks by the location of the redaction, ambiguity is key. When an extraction creates a disjointed flow, structure, or semantics then it is obvious something is missing. This can be just as indicative as when a placeholder substitution replaces the extraction. It is important that any placeholder convey as little additional information as possible, and if at all possible, any choice for that placeholder provide a statistically insignificant clue. This parallels the utility of the Security by the One-Time Pad and why the one-time pad still remains one of the most secure encryption key methods.

[0316] Formlessness can be applied to text, documents, images, audio, video, structured data, hierarchical data formats, and anything else represented by computer files of which we are aware. This includes specialized metadata formats and hierarchical document object models too. All forms of data, such as audio, image, video, database, compiled software applications, even CPU machine code are data streams that are generic until a purpose or utility is applied to them as imbedded metadata or external directives.

[0317] For example, an audio file displayed in MS Paintbrush is unlikely to have a semantic value nor will a database streamed as a video. When a correct guess is made, even when only a partial placement of audio is shown as a video, contextual clues establish a semantic neighborhood that can be explored until an exacting match is found. Hiding the purpose, file extension, or assignment of the appropriate application is an example of security by obscurity that does not withstand much attack. It

is a matter of time until patterns within the data stream or specific markers establish the link. Random trials can also define the type and context for the data stream. Note that a spreadsheet, a database, or a recovery map are all just specific instances of a generic data stream, as detailed in AA.1.4.0 Formlessness as Security for Maps.

AA.1.4.1 History and Concept

[0318] Formlessness is an oriental (primarily Buddhist) concept underscoring philosophy and warfare for thousands of years. It is so prevalent and ingrained within the basic culture that it transcends definition. Because formlessness defined part of the social fabric, it didn't really need substantiation and direct reference within the related literature. It even has its own Chinese symbols. Formlessness is a cumulative concept based on other perceptions. Similarly, many western concepts like freedom, democracy, and independence are cumulative and require a foundation for utility as well. These are not standalone ideas and definitely not separable from the culture. They do not export well to other societies and cultures until the foundation is in place. This is as true for formlessness as it is for our essential western concepts, and thus an understanding of formlessness is as part of a social culture as an understanding of freedom is part of a western heritage and social and intellectual development.

[0319] Formlessness represents the highest level of enlightenment. It is foreign to western literature and philosophy because of the different focus in science, reality, legitimacy, and central material and corporeal value systems. Formlessness, in the guise of zero, infinity, and quantum bits that both exist and do not exist at the same time, and multiple universes postulated by string theory, are all certainly the new areas in science. These are new areas for exploration, statistics, randomness, transitions, and so forth. The occidental view of science is different in some ways, but it does exist and even predates most western civilizations. It is symbiotic with the complexities of formlessness. Writings

on Tao and Zen often invoke formlessness. You might recognize formlessness in symbolic representation as yin and yang. Yin is formlessness; yang is form. Yin has no meaning without yang. Yang has no meaning without yin. The symbols have no independent meaning, value, or translation when they are apart. Together, yin and yang form the Toa, a philosophy, a way of living, and a method for progress. The Tao is usually represented by this symbol, called the Tai Ji Tu, a social concept.

[0320] The circle contains the basic universal polarity of yin (dark area) and yang (light area) and shows the continual motion of change. As a side note, tractates also discuss the reversal of the color scheme and the interchangeability of meaning. Yin can also be can form and light and everything. Yang can be formlessness and dark. This self-contradictory and circular self-reference is a fundamental part of the oriental concept. Western biblical literature brings for the light and a form that is good from the darkness or nothingness. The science of formlessness is not well developed, if really at all, and these concepts represent a belief structure when even modern physics fails to provide a finished theory for everything.

[0321] The beginner's Taoist formula of creation is this: The Tao is also the Wu Ji, or ultimate nothingness/formlessness, symbolized by the empty circle: The Wu Ji contains within it the Tai Ji, or supreme ultimate: The Tai Ji, being the source of time and space is the source of the two universal forces, yin and yang. As the Tai Ji separates into yin and yang then all of creation is formed from their interaction. Another description of the formula is that the Tao produces the One, the One gives birth to the Two (yin and yang), the Two interact giving birth to the ten thousand things (all reality and phenomena).

[0322] The concept of yin and yang is about understanding polarity, dualism and complementarity. In western philosophy, as currently practiced, there is only a duality of good and evil,

or more to the point, security versus functionality, or as we profess, security versus efficiency where functionality is replaced or measured in terms of efficiency. Ying and yang is absolute and well-known. The Chinese concept (and other eastern teachings) states that duality is only one aspect of the polarity of all life. While different cultures have legal and moral codes that stipulate good and bad, on a deeper universal level, the balance of duality is always in flux.

[0323] The principle of yin and yang is that everything is comprised of both aspects in varying proportions. Nothing is solely a yin or yang thing, it is always some relative combination. Every thing or activity has an opposite polarity, every yin has its yang. Yin and yang are opposites, but yin and yang are identical. The images form contrasting mirrors. Yin is Yang, although it is not. This manifest contradiction underscores the subtlety of occidental concepts and the depth in the meaning of formlessness.

[0324] Yin and yang eventually and always change into one another, just as life continually changes around us. This happens in the physical universe, the natural world, the social world, and inside our body (as noted and treated by Chinese Medicine). Instead of being defined states, the two forces are complementary and will always find legitimate expression. In physics, this is known (only in part) as: "each action has an equal and opposite reaction." The goal of life is to balance our inner selves, our way of daily life and relationship to the Tao by balancing the polar forces of yin and yang in all aspects of our being.

[0325] Through study, meditation, conscious movement (Tai Ji Chuan and Qi Gong) and experience with teachers, the inner world becomes simpler and the outer world more calm. The 10,000 thoughts and phenomena go back to their source as the Tao is found in daily awareness. Formlessness is not a tool that can be added to a toolbox or bolted on as an aftermarket solution. If anything,

formlessness is an approach to applying existing tools.

[0326] This discussion previously asserted that without form there is limited recognition of formlessness. This is the essence of yin and yang. Logicians and mathematicians call this circular definition. It is not without some cultural tension that western logic meets oriental wisdom. Indeed, the very symbol of formlessness is the mirror image of a circular pattern. Formlessness is and it isn't. The essence of formlessness is form, and lack of form, and form itself presupposes formlessness from which to assemble form. Form and formlessness are production processes interrelated and interdependent on each other and from context.

[0327] Sun Tzu in the Art of War specifically connects successful techniques of war to formlessness as a concept, but assumed his occidental audience understood exactly what formlessness is. This concept was basic to the culture and metaphysics. However, as related above, this is an unfamiliar concept to western civilization, and part of the reason that it is appears so infrequently and only in reference to military strategy and Su Tzu. The concepts are being meticulously implemented by terrorists, such as al-Qaeda, to enable asymmetric warfare. But formless asymmetric conflict requires a higher degree of management and control than typically applied in western-style organizations and is frequently underestimated by western-style organizations. So in many ways, formlessness is not less, inferior, or subservient to organized warfare.

[0328] In fact, western military thinkers classify basic conflict with available endemic items as weapons as 1st generation warfare (1GW), massed armies squared off against each other as 2nd generation warfare, modern war with mass destruction and genocide as 3rd generation warfare, 4th generation warfare as that wage by mobility, whereas 5th generation seems destined to be about asymmetry and disruption, perhaps a war without kinetic weapons at all, waged only with the value of information at once formless

and patterned.

[0329] This is described later in greater detail for those interested in the connections between economics, warfare, strategy, information, and asymmetry. These ideas derive from seminal strategic thinkers, like Su Tzu. Sun Tzu's observations specifically on formlessness include:

[0330] "Be extremely subtle, even to the point of formlessness. Be extremely mysterious, even to the point of soundlessness. Thereby you can be the director of the opponent's fate."

[0331] "Victorious warriors win first and then go to war, while defeated warriors go to war first and then seek to win."

[0332] "Therefore when you induce others to construct a formation while you yourself are formless, then you are concentrated while the opponent is divided...Therefore the consummation of forming an army is to arrive at formlessness. When you have no form, undercover espionage cannot find out anything intelligent to form a strategy."

[0333] Although there are no direct connections in his writings between formlessness and deception, the context is obvious in occidental culture. Trotsky used these techniques, and referenced them in his journal, during the Russian Revolution to overcome the greater forces of the Tsar and monarchy. It is appropriate to security in that information retains its utility and value only within context as soon as information decays, is dispersed, and relationships to meaning and culture are lost, its value is lost too.

[0334] Another common aphorism, "Plan for peace but prepare for war" also cites the duality between these two and the need to create deception of intent (or concept). Resources applied for peace need to have value in war, and war materials need to be transformable into peace returns. Formlessness also applies to your own resources for you might not know how they can be better organized or even

turned against you. Peace and war are opposites, yet sometimes they occur at the same time and within the same context. The world has experienced cold war and cold peace, neither peaces nor wars, but very clearly coexisting conditions of cooperation and disagreement. Formlessness in war and peace, are yin and yang, complementary but contradictory, and timeless as war becomes peace, and peace becomes war again.

[0335] Formlessness plays significantly in warfare. Conspiracy theories abound that the assassination of the Archduke Ferdinand was planned by the Russian Special Conference and the Russian Military General staff as early as 1913. There is some clear evidence they trained and funded the Black Hand and the actual Serbian co-conspirators. Recent publication of the Russian attaché's journal (2005) reinforces that claim. Yet nearly a hundred years later, we cannot tell whether this material is authentic. The Zimmerman Telegram accurately shows that control over information and maintaining secrecy through formlessness was the valuable instrument to drag America into the WWI on the side of the British. Cyber attacks on Estonia and Georgia during 2007 and 2008 were thought to be the work of Russian military intelligence to effect democratic destabilization and as a prelude to invasion, but the trail of evidence has been carefully protected to date. Attacks, defenses, and subversive activities without attribution tend to be more frightening and effective than complete knowledge. Formlessness takes many forms.

[0336] Formlessness in war and business competition is also about classification bias. Information can be applied to an intended activity but repurposed by an adversary to something else. The benefits of information could seem obvious, such as a deployment of forces or supply reinforcement. However, could work against that information to choose different battlefield that confuses that planned deployment or makes resupply impossible. Al-Qaeda 9/11 attacks demonstrated that airplanes seen only

as transport vehicles and previous targets for opportunity of hijacking could become flying bombs. Effective usage of the classification bias includes the recognition of alternatives outside-the-box and the further protection of the assets to preclude repurposing. You do not want to presume tactical information has value only within your own contexts and a few obvious attack scenarios. Formlessness is enabled by thinking out-of-the-box. Security and functionality are also contradictory, often in direct inverse proportion to the other. This is typical with information sharing, where information needs to be shared, but the more information that is distributed in order to allow analysis of risks and assess potential for threats, the more information distribution actually occurs that has the tendency to leak secrets, expose secret plans, and undermine the very security sought. However, if security is amended as a Yin and Yang concept, security can be reinterpreted as part of functionality, as collaboration for functionality, rather than as a direct detraction to it. When security is perceived as the granular control of information and its distribution, functionality is not directly affected, but supported.

[0337] Yin and yang plays a crucial role as well. Just as it is necessary to centralize management in order to direct operations, it also concurrently necessary to decentralize the resources to establish security, power, survivability, continuity, prevention of repurposing resources by adversaries, and to enable rapid deployment. Catastrophes, natural or human, are the primary annihilation of single-point resources. It is as true with military shock forces as it is for information security. This is both the substance of guerilla combat, asymmetric warfare, and also information survivability. Command plans the strategy but implements with small and dispersed forces that are rapidly deployed to address a requirement. Sometimes that command is centralized, sometimes command is delegated to situational-awareness leadership, and other times, command and control (C^3) is self-sufficient and is self-evident within each force.

[0338] The value of formlessness is expressed in the oriental marshal arts of Jui Jitsu as Ju No Ri. Formlessness can be applied to attack and defense such that resources and energy can be deflected and so reapplied against the opponent. This is a method of converting the opponent's resources against them, a method of warfare first described by General Ulysses S. Grant in his Civil War march on Atlanta and now described as asymmetric and next generation warfare. See the white paper on Ju No Ri.

[0339] The success of attacks on small outlying resources is driving many to consolidate or outsource. Consolidation also concentrates resources that have no value apart, but assembly to create bulk, barriers, new knowledge, access to others, and the power to act. While many organizations are rediscovering fortress-like perimeter protection in order to consolidate the footprint and resources necessary to provide security and achieve economies-of-scale, they are also correspondingly seeing a rise in successful and focused attacks as a direct consequence of this unification.

[0340] Centralization is both a risk and a benefit. Decentralization makes the problem more complicated but provides the benefits in the distribution of this security risk. The combination of centralization of management but distribution of risk through decentralization of storage is the combination of yin and yang. This is the DigitalDoors implementation of formlessness, the centralization information, the dispersion of parts, and the reconstitution of the whole.

[0341] Protection technologies of firewalls, IDS, IPS, code verification, access controls, and other perimeters represent barrier-based protection. They do not work ipso facto because we are experiencing a security meltdown, but in conjunction these very barriers are creating isolation and disharmony as well. The existing "solutions" are so rigid they might be creating problems worse than those by our adversaries.

[0342] In conclusion, success is about security, power, and action with inaction. It is about the

formats and formlessness. Formlessness enables the architecture for dispersion, which provides for arbitrary levels of granularity and semantic content control. It is about illusion, self-illusion, but self-knowledge. The application of formlessness in nature or by intelligence is to create opportunities and confusion. For one side it is Yin, the other Yang. Out of nothingness, comes everything, a pattern, an opportunity, or the confusion of the adversary.

AA.1.5.0 Security through: dispersion, formlessness, reversible chaos

[0343] Dispersion as a security method is an implementation of formlessness. It only works when the structure, format, and content are disrupted sufficiently to preclude even an knowledge-based or statistically-driven forced reconstitution of the pieces. DigitalDoors granular content control presumes a level of statistical formlessness with independence between the resulting redaction and the maps, the resulting redaction and the extracts, and even independence and randomness among the storage locations and within the extracts themselves. Collections of related extracts paint an undesirable picture of purpose, utility, meaning, source, and risk.

[0344] Dispersion effected through even the random sequential partitioning of a data stream fails to statistically enhance security. There are more ways to do this poorly than correctly. An example of poor dispersion is the separation of a data stream into seven or twenty three consecutive divisions. The failure of this concept is that these divisions are not independent of each other either at the context, context, or concept levels. Words likely fit together likes pieces of jigsaw puzzle.

[0345] Sentences or paragraphs might reference or flow into each other, or unique structures link together without intellectual doubt. Realize that a mosaic (or collage) reconstructed from subsets or overlapping sets does not have to have to replicate the original exactly as a forced facsimile erodes efforts to provide security, in particular when each piece retains a contextual standalone value. Fraud

and counterfeits expose a like but different problem. Authentication and validation provide added-value to verisimilitude or facsimile, whatever is attained, and are critical to the information life-cycle management of dispersive-based security.

[0346] Multiple overlapping divisions provide even more meaning for unintentional reconstitution because of the clues in the parts that overlap. This overlap is common when a RAID-like feature is added to allow for complete reconstitution even when one of the divisions is permanently lost because the overlaps provides clues for aggregating the like pieces into a set and restitching them correctly back together. However, continuity and survivability with RAID-like duplication is fundamentally more secure when dispersions are duplicated, but these dispersions are independent of each other.

[0347] Dispersion when combined with encryption is still consistent mathematically with the same level of strength as the underlying encryption foundation. It really does not add anything to functionally enhance security by any orders of magnitude. In contrast, dispersion as a security method is predicated on the level of randomness achieved, where randomness is judged on the disruption of the content, context, and concept basis for the redaction and extractions. This includes data stream formats, encoding, symbolization, metadata, structure, and hierarchy, which are a part of context and concept. In other words, a database secured through redaction is likely to retain contextual clues about content, purpose, the intended workflow, and the processing application unless it is significantly desiccated.

[0348] Therefore, the effective of dispersion as a security method is a based on the implementation of formlessness as level of achieving a reversible chaos. Randomness and independence are key measures of dispersion security, just as they are in establishing the strength of encryption, PKI, and one-time pads, and similar statistical methods are pertinent for establishing the levels of security

achieved. Randomness in content, context, and concept is a more difficult mathematical problem because semiotic and syntactical analysis is an open-ended universe driven at best by approximations, the current state of wisdom (and interpretation of meaning), and existing technology.

AA.1.6.0 Formlessness to secure maps

[0349] Reconstitution is the restoration of redacted content, context, or concept. This process uses a map to restore information exactly like (or similar to) the original source data stream before the DigitalDoors security redaction. DigitalDoors creates maps to reverse the formlessness securing the redacted source data stream and/or enabling granular control over information sharing and discovery.

[0350] The map is metadata -- a higher-order representation of information -- represented either as a set of instructions, blueprints, positions and sequences, or overlaps used to reassemble brushstrokes, pixels, bits, characters, rescaled or reformatted pieces recreating semblance to the original source data stream.

[0351] A map is no different from other raw data stream, source code, applications, workflow, or formatted and structured information. A map is information confined by a context and concept for its presumed utility. It is a specialized data stream with a purpose-driven format and specific context like any other. It is no different from data in any other file format or structure secured by DigitalDoors

[0352] The map itself can be dispersed and/or encrypted just like the source data stream it represents, as another level of security formlessness. This formlessness of a map is its security as with any other data stream.

[0353] Integrity for the map is assured using the exact technology as with any other data stream. A map can be stored to encrypting file storage systems, or stored in the cloud hashed, encrypted, digitally signed, or dispersed, much like any other data element or stream. Versions, replications, and

configurable dispersion options for the map assure survivability and security even within the most dysfunctional environments. Combinations of these tools are also supported where performance, configuration management, and overhead are the only practical limiting factors.

AA.1.7.0 Formlessness by breaking context - hiding is not enough

[0354] True security requires the splitting of content, context, and concept and then dispersion and hiding of those split pieces. This processing of the very asset that is in need of protection transforms the asset into a formless asset that has no value and no reuse value. However, formlessness when reversible, reconstitutes that value as needed, for as long as needed, without providing porosity or reuse value to the adversary. Dispersion of the data, dispersion of the metadata, and the disruptions of these references effects absolute sustainable security.

[0355] Hiding by itself is not enough. Hiding is not perimeter-based security. That's a good thing because security at boundaries and the edge is at best a solution that is not absolute. Every functional perimeter is ultimately porous, ambiguous, and diminishing as a protection method. Perimeter-based security is also an infrastructure that does not protect assets directly. Hiding alters attributes of assets to mask content, context, and concept.

[0356] A security method that is different promises a security solution. However, it is an unfulfilled promise. Hiding alone does not work. It is ultimately unsustainable. Hiding is a security method based on the assumption that you are smart while your adversaries are not. It works to a point, and then diminishes just like a perimeter. You are creating 'something' to protect the assets but not protecting the assets directly. That 'something' is just a misrepresentation of the assets. Adversaries find the assets nonetheless. Adversaries are intelligent.

[0357] You cloak a fighter plane as an empty field, a battle tank as a spoil pile, an aircraft carrier

as a tug boat, a vault as a picture on a wall, a significant data asset as something not material or even as waste, or displace the position or availability of a special operations group. That is hiding, obfuscation, obscurity, masking, deception, or camouflage. Hiding is synonymous with displacement, reformation, transformation, mistagging, disinformation, and misinformation of assets. This does not secure the assets, but misrepresents the value, position, accessibility, and/or purpose of the assets. This hides references to assets.

[0358] Secrets are hard to keep. Operations tend to expose value, position, accessibility, and purpose of the assets over time. Adversaries understand that something is being hidden to disadvantage them. Secrets tend to have a context and a similarity that leak through aggregation, inference, and interaction. It works to a practical and functional degree, but is not absolute. The failure of this method is the false expectation that because it works sometimes, it works all the time and everywhere. It is only a porous solution, just like perimeter-based security. It is a false security.

[0359] Adversaries are smart, focused, flexible, determined, and motivated to find weaknesses across a broader asymmetric attack universe than possible for defense. DigitalDoors refers frequently to the security risks from aggregation, inference, and interaction because traditional security fails within these venues. Information compilations and similarities expose secrets. Inference is useful to make intelligent guesses or mathematical assessments of them. Workflow, sharing, and information distribution creates more points for leakage, and statistically over time one or more fail. Hiding is a doorway with porosity, no different that a perimeter, but it does not really exist even though it is positioned as though it exists. Hiding is a falsified perimeter, one that does not even exist. The danger is that many security methods are forms of this.

[0360] Security is limited by obfuscation. Security is limited by obscurity. Security is limited by

hiding. Security is limited by masking. Security is limited by deception. Security is limited by camouflage. These are all one and the same. It is a security that does not even exist, except as a deception. Security by splitting and dispersion of content, context disruption, and concept randomization is different. It represents formlessness. Hiding by itself is not strictly formlessness; it is about hiding the information about the assets but not really protecting the assets directly. Hiding is an aspect of formlessness only in regard alone to the dispersion of the attributes about the asset, which is dispersion of the metadata.

AA.1.8.0 Differentiation between Obscurity and Formlessness

[0361] It is useful to explore examples of obscurity and formlessness using Where's Waldo as a metaphor. A presentation of images shows traditional and non-traditional efforts to hide, obscure, or redact and disperse the pictorial character of Waldo. Obscurity is not as effective as formlessness.

AA.1.8.1 Obscurity

[0362] Obscurity is the state of being inconspicuous, garbled, camouflaged, or indecisive. Hiding things in plain sight, distributing valuables around, and dispersing resources are classic examples of obscurity.

[0363] Obscurity is the example of hiding Waldo within a mesh of similar resources. It is still in plain site, at least the average site, but it is obscured through limitations of normal visual acuity. Waldo will be discovered, whether quickly or eventually, depending on the quality of the tools employed. Some people are better at pattern matching than others because they can see exact colors or outlines embedded within camouflage. There will be someone with above-average skills to see through the fog or others who create specialized tools to circumvent the expectations of the mean time required to break what is essential a cipher based on normal human limitations. It is not a question of if, but of

when, and what new technologies can enhance solution. Waldo is still Waldo and cannot be anything contextually or conceptually different. The presumption and limitation of obscurity is that the fundamental context and concept of Waldo must remain intact as a definitive iconic entity with a known presentation and semantic value. Waldo is always Waldo.

AA.1.8.2 Formlessness without context

[0364] Formlessness is the state of information or resource without context or concept. Obscurity is an inferior subset of formlessness. Formlessness is the example of disassociating the image of Waldo into separate brush strokes or pixels, rescaling the pieces, altering the color, blending the pieces into other frameworks such as other games, works of art, or advertisements. Formlessness robs information of its inherent semantic value. The context of the Where's Waldo game no longer exists and the concept of Waldo as the focus of interest is buried within many other processes and requirements. Waldo is no longer an identity but a series of information elements equally likely to represent a number, an image of almost anything, a letter in a word, a recipe for food, or a workflow formula. The presumed utility is deconstructed into formlessness. Waldo is a binary data stream, a collection of alternative representations in a new basis, a conversion into a new encoding, or a disassociation of parts from the whole. Waldo is something else.

AA.1.8.3 Proverbial versus Actual Context Change

[0365] Obscurity is the proverbial needle-in-a-haystack. A small element is lost within a large amorphous blob. The needle is still and always a needle within pile of fresh or dried grass. Solutions include the manual effort to extract this value from the chaff, or better methods to enhance the effort. A magnet is more effective because it lowers the precision and time requirements. A huge electro-magnetic renders the problem instantly moot. The issues of obscurity can be rendered instantly moot

with technical solutions. Formlessness is an abstraction of the needle into something else within another abstraction of the background chaff. That abstraction is conversion of information into other formats, disassociated forms, or representations and transformations without contextual or conceptual overlap with the original situation. Proverbial obscurity constrains the problem to a contextual basis, whereas formlessness is requires a contextual set change.

AA.1.8.4 Obscurity and Formless Differentiation

[0366] The differentiation between obscurity and formlessness is not altered by the format, structure, intended utility, metadata, accessibility, or ownership of the information. Obscurity is still a subset of formlessness. Obscurity is camouflage but not the semantic disassociation enabled by formlessness. Information as data, process, or maps -- and any other conceivable form -- is bound within the same ecosystem. Differences between information types are arbitrary at best. Database structure, metadata, hierarchical document object modeling (DOM), and maps are consistent with any other data format.

[0367] Generally, each is transformable into any other format. Information is one and the same, as are all subsequent expressions of these basic forms. As such, the logic of obscurity and formlessness applies equally to raw data streams, organization documents, SQL databases, macro code, computer applications, maps to information elements, and complex human workflows. The greatest difficulty might be in the concise organization of these and in creating methods to convert bloated and verbose formats into a minimalist form.

AA.1.8.5 Recovering from Obscurity and Formlessness

[0368] Breaking through obscurity is an 'aha' moment when the camouflage becomes invisible. The information was there, always there, just not discernable without that recognition or help from

advanced tools. Formlessness is a contextual transformation and dispersion of information that is not reversible without a precise process to restore the information elements back into a form and structure consistent with the intent of the original message. The iconic representation is not necessarily reconstituted exactly but at least contextually similarly. There is enormous flexibility in range of what is functional and acceptable, thereby allowing a downstream utility not constrained by the traditional dichotomy between efficiency and security.

[0369] Reconstitution is the partial or complete restoration of the content, context, or concept through a process defined and described as a map creates maps to reverse the formlessness imparted to dispersed information. A map is metadata, a higher-order representation of information, either as a set of instructions, blueprints, positions and sequences, or overlaps used to reassemble brushstrokes, pixels, bits, characters, rescaled or reformatted pieces to recreate a semblance of the original source data stream. A map is no different from raw data stream, source code, applications, workflow, or formatted and structured information because it is still information confined by a context and concept for its presumed utility. Formlessness is superior to obscurity in disassociating the utility, potential, or purpose of information to apply security to a cloud ecosystem and enable granular content control.

AA.1.8.6 Formlessness through Microperimeterization

[0370] Microperimeterization is the creation of a closed set within an open set for the purpose of advancing security. This is the process sandbox. Microperimeterizing does not create more security than perimeterization, only complications because creating open subsets does not create closed sets. They are still open. If additional security were imposed through access controls on some or all of those open sets, they would have limited access and exist as controlled closed sets.

[0371] When DigitalDoors implements microperimeterization through selection, extraction, and

dispersion, it is also creating open sets with limited access control. However, DigitalDoors goes a step further by breaking the context and concept within the content so as to add formlessness into this security mix. The open sets contain information with limited intent, purpose, and utility. It is statistically no different from randomly created chaff. These resulting open sets have none or negligible value and are indistinguishable from random sets.

[0372] The DigitalDoors recovery files exhibit both microperimeterization and formlessness, which together impart a flexible improvement to security. Dispersion of content break the context; these dispersions impart classification bias that establishes the essence of formlessness because the is no clear meaning for the dispersed content, no applicable consolidating context, and the context for the dispersions is lacking so meaning is randomized. Meaning is not completely random, as the examples with the definition of "ball" detail, but the meaning has such a wide range as to preclude useful analysis. Microperimeterizing with granular content control goes a long way to overcome the clear and present security limitations. Formlessness creates a venue for assets without known value and risk basis.

AA.1.8.7 Formlessness the difference between Dispersion and Hiding

[0373] Security is enabled by various means, such as perimeter protection, encryption, hiding, and dispersion. Perimeter security is the predominant method for infrastructure, while encryption has defined data security during the recent communications and information revolutions.

[0374] Perimeter security is under constant attack and is failing to reflect the virtualization of the zone or the cloud that require protection. Encryption is the obfuscation process to transform data using a mechanical, mathematical, or other statistical process rendering it unrecognizable and useless to all but intended parties. Encryption disrupts operations and increases complexity to the breaking point. Dispersion and hiding, like encryption, are forms of data obfuscation.

[0375] Hiding is the process of obscuring the location, utility, or purpose of data -- but not the data itself which is left intact and fully-functional -- thus rendering it difficult to use or find to all but intended parties. Hiding is obfuscation of location without content or control alteration. Hiding is a camouflage that presumes the inability to find, catalog, interpret, or utilize the data.

[0376] Hiding alone is not sufficient because it presumes that the hiding places will not be (ever or soon) discovered. Hidden data retains integrity and can be explored, mined, searched, and inventoried to retrieve definitive meaning. Many common tools reveal hidden information. Hiding presumes more difficulty in reverse-engineering location than is really the true situation. Sophisticated tools and workflows provide an enormous range of emergent reverse engineering capability to unravel hiding as a security method. The protection is based on the obscurity of the actual hiding places or access control to the hiding places. Computer operations, human intelligence, and the weaknesses in access-based security make breaking hiding relatively trivial.

[0377] One of the primary problems of hiding is that the location is often forgotten. The hiding locations are forgotten over time, with changes of personnel, attrition, end of tours of duty, and advancements. We are still finding caches of ammunition from all recent wars as far back as the French Revolution (1989-1993). Archeologists find other types of weapons dating back many millennia. People are like squirrels that forget where they hide their nuts.

[0378] Masking is also a type of hiding, implemented as redaction, extraction, reformatting, subset queries, and information partitioning. This is implemented as standard redaction with a black marker, as displaying a partial credit card or surrogate account number, or by information aggregation as practiced in medical journal reporting where PII is only shown in composite.

[0379] Dispersion is the data security obfuscation process to transform data by disrupting the

utility, meaning (semantic and semiotic), and purpose of data and by rendering it unrecognizable and useless to all but intended parties. It recognizes the granular structure of data and provides security based on meaning. This transcends implied or expected usage with definable crypto analytical difficulty. Dispersion can be controlled from none, some (partial), or all, with unlimited granularity. Dispersion increases the scarcity of information by creating formlessness. The information no longer exists to provide the intended message, although it is known to exist or once existed, so this formlessness is reversible.

[0380] There is a functional difference between dispersion and hiding. Dispersion is content annihilation. Dispersion utilizes the distribution of resources without regard to utility, practicality, or basis and the complexity to interpret the results. Dispersion does not require hiding. Data extractions and dispersal results can be placed into plain sight because security is provided by the break in content, context, and concept. Dispersed information lacks meaningful content integrity.

[0381] Perhaps, statistical assemblage could be used to separate and coordinate dispersed information, but the density and complexity makes this unlikely. The continued and ongoing incremental usage of dispersion increases the factorial complexity to separate and coordinate extractions.

[0382] Several companies (RSA, Cleversafe, RevStor) are providing dispersion based on block-level data stream division. In other words, the file is divided into 9 or eleven pieces, which are encrypted and replicated to multiple storage locations. This type of security is predicated on the inability to overcome the encryption infrastructure and stitch the separate pieces back together. As the number of replications and opportunities for aggregating and aligning the pieces increase, statistical security decreases. Since the edges of these arbitrary divisions are often syntactically unique enough, alignment is often trivial because it is computationally linear and solution symmetry simplifies the problem as each

puzzle is reversed. Dispersion requires an additional complexity because concept and context are also disrupted so that content is randomized for effective security in a non-linear space.

[0383] Dispersed information lacks the effective context needed for later usage. This is the same security enabled by the one-time pad encryption of a message. Each random reassemblage of the dispersed information is as equally likely as any other. Some might incorrectly suggest that inherent structure or syntax can help reassemblage of information, but if context and concept are extracted along with the content, purposeful aggregation, inference, and interaction is prevented.

[0384] In effect, there is uncertainty whether enough of the dispersed pieces are aggregated, that the correct dispersed pieces are included in each set, or that partial reassemblage is sufficient to provide non-trivial accuracy of the content, context, and concept. As an example, to understand the various meanings of "ball. " The example sentence "The pitcher has thrown the ball, an inside curve ball, to number 27, Red Ball, and the umpire has called it a ball" shows how partial message recovery creates uncertainty and as many likely misinterpretations as the number of extractions and dispersions.

[0385] In other words, when a document stream is parsed, crown jewels are independently extracted and these crown jewels are individually dispersed, any validated connection between the redacted source document and the extractions and even among the extractions themselves is difficult to establish. Meaning is arbitrary without the clues provided by structure and order.

[0386] Furthermore, each dispersed crown jewel is also independent of others, thereby preventing aggregation or inference attacks. Mathematics provides added security in that as the size, distribution, and overall randomness of the extraction storage increases, forced connections cannot be correctly established. Overall randomness can be further enhanced by the generation of extraction and dispersion chaff. The density of the chaff could range from 0 to nearly 100%, clouding what value might

exist.

[0387] Dispersion represents a puzzle with puzzle pieces that are independent from each other (e.g. no jigsaw clues with color, content, connection, or placement hints), an overwhelming selection of pieces to false and legitimate puzzles, duplicate pieces that could imply a content continuity or merely a red herring, any lack of structure or order for reassemblage, and missing (withheld) pieces through the granular content control process. By the way, the metaphor of a puzzle used here presumes a two-dimensional reassemblage space, but that is by means even certain. Many data structures are multi-dimensional or hierarchical, so reassemblage complexity is an unknown z-space. This establishes unlimited dimensions of uncertainty, thus cryptographic security.

[0388] There is also no certainty that all dispersed extracts are available from the same storage location. These independent pieces could be dispersed to different hidden storages, and while the pieces might be discovered in time and the hiding places exposed, the context among any of the storages and pieces within remains tenuous at best. Depending on the degree of granularity allowed in the redacted message, several different and equally meaningful sentences could be reconstructed, much like the equally-likely decryption of a one-time pad message with trial keys.

[0389] Additionally, it is a false presumption that the source and thus the reconstructed messages are syntactically correct, do not purposefully distort grammar, or merely provide symbols into another coded medium. The structure itself is important to provide context. The file format could also be a database, an email, or merely bits encoding in a different basis, and the lack of foreknowledge as to the utility disrupts any possible concept. Perfect secrecy is a strong notion of crypto analytical difficulty that is neither knowledge-based nor obscured in false presumption of sufficient complexity to prevent immediate and widespread disclosure.

AA..1.8.8 Dispersion as Asymmetric warfare

[0390] Every protection measure instigates countermeasures to defeat it. Countermeasures are procedural or technical. This is the nature of an arms race. Better weapons beget better countermeasures. Better countermeasures beget better weapons. Arms races are counterproductive to progress and often reverse it.

[0391] Security has a history of measure versus countermeasure. Physical barriers and point solutions have proven ineffective, and aggregations of these methods also fail. Such efforts tend to divert attacks to softer targets, provoke new attack vectors, or instigate a bewildering array of new technological complexities to circumvent the security while achieving the desired attack goals.

[0392] Aggregation of physical barriers and point solutions often create new problems that reverse the benefits of security methods by undermining core process; this often describes the situation of security versus functionality.

[0393] Asymmetric warfare is one of the few methods to bypass this arms race. It is an indirect attack. Asymmetric warfare is applied against non-military and civilian targets, affected through misdirection, third parties, and covert activities. Although there are some direct countermeasures to asymmetric warfare, indirect methods generally are superior.

[0394] Often asymmetric warfare masked through third parties raises the cost of retaliation so high as to be politically reprehensible. This relies on the concept of plausible deniability and the use of the human shield. Few clear retaliation opportunities are available against the Russian cyber attack on Estonia in 2007 because it is unclear exactly how the attack was perpetrated, what resources where used, and who sponsored the attack. Hoaxes can also be as disruptive as real attacks, leaving fewer telltale markers and no actionable responsibility. Lashing out at shadowy enemies is at best inadvisable, could

damage uninvolved pawns creating new enemies, creates more collateral damage, and is usually futile.

[0395] One of the reasons information security is so difficult is that attacks are asymmetric and covert. Hardening the infrastructure and securing the process with validated code is a countermeasure against only clear and present dangers, not against aggregation, inference, and interaction. In time, hard targets will give way to soft ones, and there are generally more soft targets and more avenues for access to these soft targets.

[0396] In 5th Generation Warfare theory, countermeasures to asymmetric warfare are diplomacy and winning the hearts and minds supporting the insurgency. Withdrawal is effective but is often perceived as yielding to the demands and goals of a terrorist agenda that if successful emboldens other like asymmetric efforts. Dispersion often slows and weakens the control and speed at which countermeasures can be deployed, but it is one of the few methods successful to infiltrate an asymmetric enemy and retain a level playing field.

[0397] Dispersion of assets and soft targets is the method to thwart asymmetric warfare. It is particularly relevant to digital warfare because targets can be replicated endlessly, easily dispersed (even intertwined within enemy resources), and quickly reformed. Dispersion is an old and practiced example of formlessness. DigitalDoors demonstrates the technology of dispersion and utility of formlessness through granular content control and unlimited levels of reconstruction.

[0398] Reconstruction enables guerilla-like utility through the ad-hoc reversibility of dispersed data. This measure protects the raw data through dispersion and limits the extent of possible damage while retaining the effectiveness of the measures and the speed at which they can be deployed. Because these measures are asymmetric, countermeasures are at best inadvisable and self-destructive. In fact, systemic countermeasure attacks against dispersed infrastructure and dispersed data are likely to create

significant collateral damage to the attacker's own infrastructure, leading to scenarios of mutually-assured destruction.

[0399] Dispersion relies on the creation of intertwined dependencies that become inadvisable targets. DigitalDoors has yet to discover a countermeasure to the dispersion measure that does not raise the costs and ramification of attack and retaliation beyond what is sustainable. Dispersion is the child of formlessness, a concept of warfare and strategy that has withstood the tests of time.

Intangible value hides below the surface. That intangible value escalates when its apparent scarcity is engineered, maintained, and protected. Nearly 95% of any corporation, even old-line agricultural and industrial ones, are represented by property assets not booked as balance sheet entries. Many assets are scattered around or perceived as corporate property or some sort of ambience without an immediate need for recognition, inventory efforts, formal monetary valuation, and trade secret protection.

AA.1.9.0 Overview of Formlessness:

AA.1.9.1 The DigitalDoors Formlessness Doctrine

[0400] The Information revolution has created extremely powerful technological tools, which can give superior competitive capabilities to organizations that leverage them. Using advanced information search and data mining capabilities combined with information inference tools can give an organization a major advantage over its competitors and in some instances may bring to the devastation of the competition.

[0401] The existence of such powerful tools and the easy access to them creates a new environment in which competitors and opponents may piece together the most important secrets of an organization, and thus try to undermine its mission, its market and possibly its reason for economic existence.

[0402] Combination of data mining tools, inference tools with cyber attack tools like Trojans, worms, etc creates a dangerous combination which can be lethal to the survivability of the target organization.

[0403] Such an environment is in essence an information battlefield and there is a need to adopt methodologies from the battlefield to guarantee minimum losses and organizational survivability.

[0404] The Formlessness Doctrine calls for adopting of different methods which will make the organization's information assets unknown. Only that information that needs to be released will be released in a controlled granular manner, the other information will be dispersed in secure distributed storage locations making the organization information assets be untouchables.

[0405] The Doctrine of formlessness calls for creation of intellectual property portfolios, which are dispersed assemblages of tools like patent portfolios, trade marks, copyrights trade secrets and which are not apparent and therefore not easy for attack.

AA.1.9.2 Establishing Formlessness:

[0406] The key to formlessness is organized order. Formlessness is in essence organized chaos. There is a need to establish at first order and organization of the information assets before creating the organized chaos.

[0407] All information assets of the organization must be located, and inventoried. Inventorying all assets calls for classification of all information assets and their tagging.

[0408] This requires inventorying structured data stored in databases, semi structured data like e-mails, and unstructured data.

[0409] After classification and tagging of the information assets dispersal process needs to take place, dispersing the assets to different distributed stores. All dispersal will be mapped in maps, which

will be vaulted in safe locations.

[0410] When there is a need for information asset, it will be released in a controlled manner after fulfillment of release conditions. For example a release may require answers for the following questions; (the 5 Ws) what is been asked for, by whom, for what reason, at what time and where.

AA.1.9.3 Defending intellectual property and trade secrets:

[0411] Trade secrets are sensitive information of the organization that can not be protected by legal means like patents, trademarks and copyrights. Such Information assets may include customer's lists, scientific formulas, sensitive studies, PII personal information etc.

[0412] Such information is usually kept in databases throughout the organization. The Formlessness Doctrine calls for dispersal of such granular information assets to different distributed storage stores. As described above.

AA.1.9.4 Defending information in web sites:

[0413] The Formlessness Doctrine calls for minimization of exposure of sensitive information on web sites. Sensitive information will be released from a web site based on identification of a viewer.

[0414] In essence web sites need to be transformed into Multi Level Security systems where different people will get different access to different parts of the content based on who they are and what are their roles.

[0415] The Formlessness doctrine calls also for release of information assets from such web sites for monitory consideration. As described below.

AA.1.9.5 Legal defenses with formlessness:

[0416] The state of formlessness needs to apply to all types of information assets including those

that are highly sensitive and private and those information assets that are in the public domain like information in web sites. Furthermore securing information assets by applying formlessness must include securing ideas and concepts that are important to the organizations. Such ideas may be kept as trade secrets in secure vaults, but when the concepts are protected by legal means such as in the case of patents or trademarks and copyrights there is also a need to protect those legal instruments so they will not be undermined and voided by the usage of advanced data mining tools. For example a patent may be attacked by intensive data mining efforts to discover prior art that may make the patent protection be worthless.

AA.1.9.6 Measurement of the information inventory for risk:

[0417] All information assets will be measured for their risk value. An information piece will be evaluated for the risk it may cause when released. The risk will be measured with different tests including advanced inference and aggregation tests.

AA.1.9.7 Monetizing the information inventory:

[0418] To achieve a serious priority for the information assets inventory there is a need to treat the information assets as economic assets with a specific monitory value. Each information piece will have its monitory value. People moving pieces of information will see those as monitory transactions having all the ramifications of dealing with economic assets. Ant waste of negligence in handling an information asset will be treated as negligent handling of money within the organization.

AA.1.9.8 The Global Information war

[0419] In the past few years major technological changes have taken place in the digital world that are creating fundamental changes in the way we interact and communicate. The extreme interconnectivity between networks and people in combination with the new powerful information

searching capabilities create extremely powerful capabilities. In essence a new information paradigm can be used both for good and bad to our society.

[0420] The advanced technology capabilities enable organizations to search and find critical information about their competitors. Searches on the Internet, combined with advanced data inference tools can yield for organization highly valuable information on the competitors. The intelligence may include: research and development results, names of customer, new marketing plans etc.

[0421] Advanced searches may also yield prior art results that were hard to locate in the past, which may serve to attack and void the patents of competitors.

[0422] A planned focused intelligence gathering operation may lay the basis for eroding the base of the competitors.

[0423] To defend against such search and inference attacks an organization must control tightly all its information assets both the internal and external ones.

[0424] Information assets on the organizations web sites must be controlled allowing release of information to parties based on their identity. The public will view very general information while selected people will have access to more information based on their security clearance levels.

[0425] Information within the company must be inventoried, classified and tagged. Sensitive data, the crown jewels needs to be located and transferred to safe vaults. Release will take place in a controlled manner.

[0426] The acceptance of such precautions is a challenge to the culture of an organization.

[0427] Currently there are no technological solutions to stop the Tsunami of different cyber attacks. The different types of attacks are spreading like cancer. There is a need to fight spreading cancer with surgery, radiation and chemotherapy. Vitamins or aspirins are not the solution. The technological

solutions currently are like vitamins and aspirins. If the technological solutions were good the cyber cancer would have been stopped.

[0428] There is a need for true strong leadership, at the highest level that will study and understand, and do something about it. There is a need for a national agenda that will relate not only to government, but to businesses and citizens. There is a need for a united front. This issue should be the top priority because the US. is losing its superiority as a superpower without even understanding that it is happening.

AA.1.10.0There is a need for a technological paradigm change.

[0429] This is a real emergency which calls for new technological directions. And Allocation of monies to development of the right technologies like DigitalDoors software and patent portfolio, rather than doing more of the same technologies.

[0430] Because of our openness as a society other countries have access to our public information on the Internet they then attack our networks and get the needed sensitive information to complete the puzzle.

[0431] In essence the US is being robbed out of its biggest asset, the intellectual property of the American people. Trillions of dollars in Intellectual property are removed yearly out of the US yearly without our understanding of what is going on. People are failing to understand the criticality of the problem because it's very hard for anyone to comprehend the aggregation of risks and threats created by the information revolution.

AA.1.11.0 DigitalDoors Hierarchical Security Sanitization and Reconstitution

[0432] The sanitization and processing of a data stream into non-sensitive and sensitive items is unlimited in that the original data stream can be decomposed into an arbitrary and unlimited number

of subordinated representational streams. DigitalDoors can create an unlimited number of transformations of the original data stream and also an unlimited number of aggregations of sensitive items. Security is created through the process and method of dispersing content and granularizing items that define context so as to preclude information knowledge, aggregation, and inference.

[0433] Maps and metadata for mapping, reconstitution, and (none-to-partial-to-complete) recovery of the original data stream can also be decomposed into an arbitrary and unlimited number of streams and dispersed to an arbitrary and unlimited number of local or remote storage locations. DigitalDoors can create an unlimited number of transformations of the original data stream. The purpose for the DigitalDoors data stream transformation is: to secure content; to secure context; to prevent data aggregation and inference; and to preclude data repurposing by insiders and attackers.

[0434] DigitalDoors can create an unlimited number of aggregations of sensitive items for knowledge management. Reconstitution, therefore, is controllable from none-to-partial-to-complete contingent upon a role-based multi-level security identification and authorization hierarchy. As a result, security is created by minimizing the access to the content in the data stream and thereafter enabling hierarchical reconstitution of granular pieces into the data stream in a controlled environment.

[0435] Vetting through identification and authorization enables (none-to-partial-to-complete) recovery of the original data stream or a step-wise defense-in-depth reconstitution in one, two, or simultaneous multiple levels of hierarchy. In other words, reconstitution with semblance to the source data stream can be recomposited in situ with the redacted sensitive items, misinformation, or other tags, tokens, and labels. Filters enable this multi-dimensional simultaneous hierarchical redaction and reconstitution into one, two, or many simultaneous data streams with controlled multi-level versions, or the exact reconstitution of the original data stream.

[0436] A 'sensitive' item is any type of bit pattern within any type of original data stream which is identified by the DigitalDoors processes and methods. A sensitive item can include characters, words, audio segments, image ranges, synchronization symbols, phrases, paragraphs, markers, structures, references, metadata, formats, indexes, tags, labels, tokens, URLs, embedded constructs, metadata, or pointers to external structures.

[0437] Where a sensitive item creates a dissonance in the functional flow or contextual meaning, the range for the sensitive item can be extended to include punctuation, markers, additional words, and linguistic or data processing formats to preclude inference to this sensitive item. A sensitive item is identified at any arbitrary level or simultaneous multiple levels of hierarchy, such as sensitivity level or levels, compartment or compartments, parts of speech, potential overlaps into various purposes, etc. Filters enable this multi-dimensional simultaneous hierarchical characterization of the item.

[0438] Substitution is used to replace the sensitive item in the original data stream.

[0439] The substitution is usually represented by a null set (that is, nothing) to preclude inference of the missing sensitive information. The substitution can also include a non-null set, represented by metadata tags, standard DOD labels, referential linkages, or placeholders sufficient to maintain functional computerized processing. Placeholders can include tokens, misinformation, or other hierarchical structures that are arbitrary or unlimited.

[0440] It is critically important to recognize that the fundamental 'replacement' for the sensitive item is out-of-band to the source data stream; in other words, the replacement is the null set but could be a hidden marker, visible marker, or tag presence. Such in-band replacement of the sensitive item with a 'token' is contingent only on conformance with needs to strictly maintain computerized or human activity workflow without disruption. Specifically, many operations necessitate a fixed-sized and precise

field structure, column structure, or hierarchical organization and disruption with high-byte toggling, tokenized tagging, or alteration of the data stream with in-band substitutions will cause operations to crash.

[0441] The sensitive items are discarded, or dispersed in an unlimited number of local or remote storage locations. Items can be saved each within its own unitary streams (item-wise), or aggregated together. Aggregated streams o items can include all items of like sensitivity level or levels, compartment or compartments, parts of speech, or coordinated for a specific purpose or security requirement. In other words, a sensitive item might be stored as a single element, combined with others all conforming to a single sensitivity level, or combined into a hierarchy with a minimum and a maximum sensitivity level.

AA.1.12.0 Methodology Overview:

[0442] DigitalDoors is a methodology for creating and building intangible assets to create economic value. (Like trade secrets).In a global information war competitors try to find out what are your assets and your plans to build new assets. They are using all kinds of means: Trojans, aggregation of information, compromising insiders etc.

[0443] The multiple threats to information are so many and so overwhelming that the best option is to emulate the military strategy of tactical withdrawal when confronted by overwhelming forces. There is a need to protect the information in the information war and by using - scrubbing/ redaction will deliver the military strategy of Tactical withdrawal /retreat.

[0444] When the environment is ready the retreating army regroups and counterattacks. The parallel in the information world is DD reconstruction of information extracts back into the documents/ data streams. The ideal state is when all or most of the content in documents is withdrawn from the

document, achieving a state of formlessness.

Table: Processing Stages at the Semantic Content Level –

Paralleling Military Strategy:

1)Tactical withdrawal/ organized retreat (scrubbing/ redaction).

2) Dispersion - in different locations (storage stores)

3) Optional -Ambush -(Seeding disinformation)

4) Regrouping (reconstruction)

5) counterattack- (searching in storage locations and creating new information assets).

[0445] The retreat strategy has been discussed at length by Carl Von Clazowitz in his book on war.("retreat to the inner part of the country").

[0446] It was practiced by the Mongols (Genghis Khan- Feigned retreat). All this is in granular content and must be operated automatically.

AA.1.13.0 A new IT paradigm

[0447] The flood of different types of dangerous attacks on information call for a realization that there is a need for new world order in IT.

[0448] The first step which is needed is a realization that we are all overwhelmed by attacks, including Trojans, hackers, disgruntled employees, espionage etc.

[0449] We are pawns in a global information war and are attacked constantly and relentlessly. Every week that goes by we are suffering more drastic losses. We are losing peoples trust in our ability to hold their information securely and might be losing their business because of that.

[0450] Realizing that we are confronted by a war and that we are losing the battles we need to withdraw our important asserts to a safe location regroup there and then continue our operations safely.

Like in a battlefield we need to retreat to safe locations in order to regroup. By withdrawing our information assets we are preventing the attackers from seeing our assets and from damaging them. The sensitive information extracts need to be withdrawn to safe and hidden locations. The ideal method is to withdraw the extracts and transport them into a "cloud". The "cloud" consists of many servers and the sensitive extracts may be dispersed all over the cloud to achieve high level security.

[0451] The transport of sensitive granular extracts to the cloud allows operations of the organization continue and are not blocked and stopped as is the case with implementation of Data Leakage Prevention methods which block files with sensitive data from passing through the organizational firewall. Release of extracts from the cloud creates a " Data on Demand " reconstruction in which sensitive extracts will be reconstructed back into the data stream based on identity of the user and other possible criteria like his role, mission, location time etc. The more sensitive information is extracted and transferred to the cloud the more secure the file and the organization.

[0452] A healthy security state for an organization is the state in which its files are redacted/ scrubbed and are in essence frames populated with skeleton content that allow continuation of operations.

[0453] The ideal security scenario for an organization is the state of formlessness where almost all of the information in a file is extracted and the result is a state of formless the form has evaporated. A withdrawal of sensitive extracts calls for a highly organized process. To achieve that every content element needs to be automatically or manually categorized, and inventoried. The categorization of all content elements in the file is the basic and beginning step. Following that extraction of sensitive elements takes place. The extraction can be done manually, automatically of a blend of both manual and automatic.

[0454] Automatic extraction is based on combination of different semantic filters.

AA.1.14.0 Taking advantage of a major economic imbalance

[0455] Society gives lip service to information being a valuable asset but in reality it is mostly considered valueless. As an example the Internet which contains many Billions of documents which are there for any taker for free. Another example is the current accounting methodologies do not value information as an asset on the balance sheet.

[0456] The lack of recognition of the value of information leads to:

[0457] 1) Lack of inventorying of information assets in organizations (60%-85% of the data is unstructured)

[0458] 2) The lack of information inventory precludes from using current information assets as economic assets as well as from building upon those information assets to create more assets. The current environment of valueless information assets precludes us from monetizing those assets. The bottom line is we are losing economically by not being able to inventory our existing assets, and not being able to monetize them.

[0459] The crucial imbalance in the basic recognition of information as valuable assets and thereafter an underutilization of such information assets is an economic flow and an unreasonable aberration. Understanding such an imbalance and taking advantage of it will lead to substantial economic gains.

[0460] Society for 20 years has focused on the perimeter on the tools storing and processing information and not on the substance it self: the information and the content. The focus on the computing and storage tools was spearheaded by vendors who have made hundreds of billions of dollars selling their products. Dealing with the content became the niche of only few small companies. The resulting

situation is that 60% -to 85% of the information of organizations is unstructured, which means the information is un- inventoried and therefore mostly unusable.

[0461] In essence even with trillions of dollars of purchases of IT tools and capabilities over the past 2 decades most organizations can't access process and leverage between 60 to 85% of their information.

[0462] Organizations can't show Internal Rates of Return (IRR) on their information investments because they really never treated the information as an asset class on its own merit with an economic value. There was no inventory, therefore no value could have been attached to those assets and now there is nothing to show for.

AA.1.15.0 Monetizing the organization's information assets:

[0463] To take full advantage of the major economic imbalance in information the following steps are proposed:

1) Inventorying the information assets

[0464] There is a need to establish an information inventorying process in the organization. The same way that retail stores or supermarkets inventory all their merchandise using barcodes, information assets throughout the organization must be inventoried. All information assets must be classified and tagged. This inventorying effort must cover, structured, unstructured and semi structured (e-mail) information.

2) Data mining - enlarging the information asset base

[0465] After assets are accounted for in an inventory, organized efforts must be undertaken to enlarge the base of those assets that are important to the organization.

The subjects of importance will be inputted into a data-mining engine, which will search the Internet and

other sources for relevant data assets.

3) Leveraging automation and legal power – the Innovation Production Line

[0466] After relevant assets are located, they will be analyzed automatically and categorized into assets that are to be treated as trade secrets- (potential contacts, leads etc to be stored in data bases.) and ideas that are of important values and need to be secured by patent protection.

[0467] Innovative ideas will be processed automatically and inserted automatically into templates of Provisional patents. The provisional patents will be sent automatically by e-mail to the patent office with automated payment.

[0468] By inserting legal framework such as patents trademarks etc into the newly formed information products the value of the information product grows substantially. To maximize the value of the created information assets people should be involved in the process to twick the input the system receives for example inputting additional key words.

4) Leveraging web scalability – the Innovation Factory

[0469] To take full advantage of being early in the market to benefit from the big information imbalance, there is a need to leverage the scalability of the web. Specifically to leverage the new social graphs capabilities.

[0470] Combining an automatic information assets production line with interaction of many people across the web to fine tune the production process will create a powerful combination – bringing an automated information creation process with the innovative input of many people.

5) Monetization

[0471] As the organization creates information products, the organization can sell such products to partners as well as to the public over the web.

[0472] Proposed model is using DigitalDoors controlled release process to sell parts of the created assets over the Internet. For example releasing first part for free the other parts for payment.

AA.1.16.0 Information disassembly and Reassembly

AA.1.16.1 Disassemblying Information

[0473] Anyone can easily encrypt, shred, delete, or destroy information so that it no longer retains utility. However, DigitalDoors takes information apart, specifically information represented by any arbitrary data stream, removing any utility, so that it can be reassembled for later use. Information is selected in various ways, selections containing a paragraph, sentence, line, word, or character, and are extracted from the data stream.

[0474] When information is taken apart, the message is changed, obscured, broken, and the entropy of the parts is increased. Entropy is the solution to government-sponsored terrorism, criminal-initiated attacks for financial gain, or preserving the economic stability of entire sectors or even nations. This creates a flexible, enduring, and inherently independent security method. This information disassembly is consistent with the need to implement privacy, confidentiality, resilience, continuity, and survivability in the ecosystem of previously inconsistent goals of mutual security, information sharing, and knowledge discovery.

[0475] Information reassembly ranges from zero to full reconstitution in minute rheostat degrees under granular control. In other words, you can get none, some little, some more, or all of the data stream back. Information is encoded with one or more systems of labels and tags or classifications

[0476] DigitalDoors supports a system of arbitrary, unlimited, overlapping tags so that preplanned, far-ranging, or multipurpose utility is supported, and a utility that is not preconditioned on information source, intended usage, or the application of increasingly sophisticated knowledge

management tools that can rework or expose information secured under existing systems.

[0477] Tagging is supported both as in-channel codes, a metadata, and out-of-channel markup. The out-of-channel is the preferred method because of its enhance security and lack of additional information for inference. Granular control means that arbitrary systems of user access control can be applied to the rheostat. Granular control includes sensitivity level, compartment, word or phrase categorization, role, group membership, syntactical element, mission, location, time of day, payment or trade, and other tags or labels.

[0478] Realize that tags condense and synopsize data streams, usually within a cultural context, and tags represent a denser risk than the initial content. Separation of content, randomization of context, and limited connection to tags are important security features.

[0479] Why disassemble information? DigitalDoors does this to enable security, privacy, confidentiality, functionality, survivability, and compliance, among other reasons.

AA.1.16.2 The reasons for disassembly

[0480] Security establishes a state of freedom from danger, harm, or risk of loss.

[0481] Privacy is instituted by a method, desire, or process to maintain personal, business, or corporate identity without releasing it to others.

[0482] Confidentiality is an expectation and workflow process that ensures that the content of a transmitted message cannot be determined except by the possessor of a key associated with the message, even if the transmitted bits are observed.

[0483] Survivability means that dispersed information can be regrouped to reconstitute the whole.

[0484] Compliance means that the technology fulfills the legal or organizational requirements

to protect and retain confidentiality of data and process protection.

[0485] Information generally exists within a structure dictated by purpose or the software tool used to create, collate, organize, store, distribute, or share it. The structure can be a document format, a file format, a database table or many associated tables, a Microsoft Word document, or an explicit structure (such as a TIF image format).

[0486] More often than not the structure is a metadata hierarchy. All of these structures, whether defined internally or externally or by the application, are ultimately a purpose-specific implementation of metadata. Even where these structures are buried within a code library or specific instructions that shave off pieces of information from an application data file, every organization of any data stream can be converted and represented by a metadata definition, and, is ultimately a metadata hierarchy.

[0487] A metadata hierarchy is a security, privacy, and confidentiality risk combined with but also apart from the data stream. The problem is that metadata is an all-purpose file structure until itself that also defines data streams within, redefines purpose and usage of imbedded data streams, and supports the inclusion of functionality. Metadata is so flexible and generic that infinite layers of purported but obscured utility can mask, distort, or hide data within. Often, the metadata hides more information and risk than the expected data stream within. Metadata in any form, whether data, application, or mixed from both provides powerful clues that enable aggregation, inference, and interaction undermining traditional security methods.

[0488] DigitalDoors removes this metadata and creates raw data streams in primitive forms that can be secured. DigitalDoors can secure the data stream and reintroduce the metadata. However, reintroducing metadata restores the risk from the metadata, maybe not the same metadata, but enough that recreates measurable and significant security risk. However, disassembling the data stream

specifically and reconstituting the information as needed enables a wide range of security, and depending on the substitution techniques applied, also maintains the consistency and functionality of the data for most information processing operations.

[0489] Note that some structured financial data can be secured in this manner -- although the need to maintain posting, trial balances, and consistency between accounts and roll-ups complicates -- but does not preclude -- the process of creating secured but working data streams through selective data extraction and granular content control.

AA.1.16.3 Finding Value within Information

[0490] Value of information is found in the data stream and through its meaning, utility, and purpose. Value is tied to paragraphs, sentences, phrases, words, or characters. However, value is also greater than the sum of the parts, so that aggregation, inference, and interaction both internally and with additional external data sources represents an important security consideration. Value is modified by encoding method and can be covertly enhanced by the encoding method or metadata structure. Every computer file exists in storage in a linear representation of "0s" and "1's" that ultimately is organized or clumped into units of meaning.

[0491] The data stream is an information flow with a starting point and a length. It does not matter how it is encoded or how the data stream represents information. Text, text in any language, text defined by words or larger units of paragraphs, numbers, names, words, images, sound, video, signals, data sets, and so forth are all represented by a data stream. One of the key benefits of the DigitalDoors technology is the recognition of risk in units far less monolithic than an entire document or document package. Granularity is achievable at word, character, symbol, or image bit levels. As such, security can be effected so that the choice is not whether to distribute a document or not, but what parts of the

document can be distributed to what sensitivity levels, to which users, and under what controls. Documents do not have to be rewritten, ghosted, or endlessly tear lined but can be automatically processed once and distributed in Secured form or granular reconstituted streams at any time now or in the future. Recallable granular content control overcomes the time element to existing security paradigms; so in effect you can "put the cat back into the bag."

[0492] DigitalDoors support uncontrolled secondary or tertiary distribution without fear of compromise in sources, methods, security, or repurposing.

[0493] DigitalDoors applies automatic methods, white lists, gray lists, black lists, complex Boolean and placement associations, symbol recognition, categorization to value selection and subsequent information extraction. These lists (actually database entries) need only be consistent with the type of data stream, the language, and the encoding methods. An organization can create multiple databases for different languages, for example, and apply one, some, or many. Database entries to match sound, images, or video overlay apply just the same way as text, although sub-message element selection is specific to channel, encoding, or fuzzy matches in terms of non-linear placement. This is pertinent to mixed media, as in sounds or images.

[0494] It is important to recognize that selection is multitasking, supporting manual, assisted, and automatic features, supporting any order of use or reuse, with manual editing and removal of any and all selections at any time with reapplication of any of the manual, assisted, or automatic methods again. Note, however, that COMSEC recognition finalizes selection that cannot be overridden, for obvious reasons of enforcement. However, new selection methods (not currently imagined) can added into DigitalDoors to augment or enhance the recognition of aggregation, inference, and interaction.

AA1.16.4 The Doctrine behind DigitalDoors

[0495] As explained, finding information in a data stream is necessary and, independent from the DigitalDoors process. DigitalDoors applies automatic methods, white lists, gray lists, black lists, complex Boolean and placement associations, categorization, manual user selection, manual overrides, and can included other methods as they are developed and proven effective. However, after selection takes places, all of these selections must be reconciled for overlap and conflict. We do this by analyzing selections in a table.

[0496] All selections are defined by a starting offset and an ending offset. This technology is consistent for all types of data streams and encoding methods. Extracting the selections is not always simple or straightforward because excising the information is contingent on foolproof reconstitution and the application or MilSpec security techniques embodied by the Bell-Lapadula Model (BLM), Saltzer & Schroeder's Principles of Secure Design, and others.

[0497] It is a significant complication applicable to the architecture of security ecosystem design for the military or for business. DigitalDoors stores selections that are extracted in one, many, or distributed locations. Reconstitution can occur in steps (like defenses-in-depth), up to a predefined sensitivity level, or to full reconstitution. Because of this, starting and ending offsets as defined by the selection process, change as a function of selection priorities and positional precedence. Selections are not independent until DigitalDoors extracts them and creates a recovery file. Selection dispersion creates data independence by virtue of contextual dissociation; this can be further enhanced uniform data stores by creating random chaff. In other words, the position of all selections is a complex simultaneous choreography that depends on the selection and categorization of all prior selections.

[0498] For example, selection priorities and positional precedence becomes important when extracting selections to create directory, recovery, and secured files. This can be a complicate multi-path

choreographed hierarchy for conformance to cross-domain MLS information sharing ecosystems. Specifically, certain security and versioning options requires a multithreaded pass first ordered by position (starting to ending) with each sensitivity level (starting to ending) for the recovery files to conform with BLM. Then certain security and versioning options requires a second multithreaded pass ordered only by position (high to low) to create the secured files since each subsequent selection is dependent on the (missing) positions of any and all prior selections.

[0499] For example, consider this well-known 35 letter pangram (e.g. holoalphabetic sentence uses every letter of the alphabet at least once).

Table: Hologralphabetic Sentence

1 5 11 17 20 27 32 35 40

[0500] The quick brown fox jumped over the lazy dog.

[S] quick

[C] brown

[TS] fox

[S] dog

and by default the unselected remainder is [U] The jumped over the lazy.

[0501] Sensitivity level: [U] < [C] < [S] < [TS] and defines the necessary order of reconstitution.

[0502] brown is removed first and coded at position 5 to reflect positional dependency on quick

[0503] quick is removed next and coded at position 5 without any positional dependency.

[0504] dog is removed next and coded at position 37 to reflect positional dependency on fox.

[0505] fox is coded and removed last and coded at position 17 because it is conditionally independent of all three other extractions.

[0506] On reconstitution, "brown" is inserted into the 5th position to yield The brown jumped over the lazy.

[0507] On reconstitution, "quick" is inserted into the 5th position to yield The quick brown jumped over the lazy.

[0508] On reconstitution, "dog" is inserted into the 37th position to yield The quick brown jumped over the lazy dog.

[0509] Reconstruction is dependent too. The position of all selections is a complex simultaneous choreography that depends on the selection and categorization of all prior selections.

[0510] This explains the dispersion variants and various levels of available conformance to BLM. Full support for all possible failures and complications is just not possible as this is N! problem. As a result, some variants provide survivability with full recovery or only partial recovery with a full security breach. There costs, speeds, and overheads fully reflect the level of service provided.

[0511] On reconstitution, "fox" is inserted into the 17th position to yield The quick brown fox jumped over the lazy dog. Its placement is contingent on the positional dependency of the three lower sensitivity level selections, but its displacement is unchanged because all precedence conditions have been previously met.

[0512] Linguistic syntax and inference issues plays into the automated selection and extraction of information. Names, locations, compound terms, data formats, common cultural phrases, aphorisms, white space, punctuation, numbers, quotations, parentheses, modifiers to nouns and verbs, and other constructs need to be excised in most cases so as not to leave a telltale gap in the Secured file. For this reason, substitution without any form of placeholders is advisable and careful removal of indications of special data formats is also indicated. For example, the removal of numbers within "05/06/23" into the

substitutions " / / " or "//" is very obvious and informative. Removal of the entire syntactical expression is indicated.

[0513] For example, the removal from "former president William Jefferson Clinton" leaving only "former president William" provides useful re-identification clues in that there are only a limited number of former presidents, few have the first name "William", and fewer still fall within a recent time period warranting security by partial selection and extraction of information. Inference requires assessment of selection by the full phrase, an entire line, an entire sentence, an entire paragraph. It is a complicated assessment of additional information within the data stream that refers to the sensitive item and exactly how it and all references need to be excised. As previously stated, information value is not specifically internalized because other internal content, external data sources, a priore knowledge, context, and concepts are useful for establishing aggregation, inference, and data-to-data interaction. In addition, usage, intent of secret protection, distribution, and secondary information sharing policies all dictate the actual scope of selection, selection categorization and sensitivity level encoding, and the granular control allowed in reconstitution.

AA.1.16.5 Reconstituting Information

[0514] At this point, information meets with policy, security law, governance issues, and compliance rules. Other existing security technology can provide similar results by tear lining and encrypting the streams; however, information is locked up and generally accessible only through a primitive filter that offers only the choice of all or nothing. This is also true of indexes and catalogs, which like tags, represent a condensation and synopsis of the original content. Instead, DigitalDoors provides functional Secured files and additionally granular content control for getting more of that information out to people and processes. Reconstituting files with granular content control means data

dispersal and information sharing is not an either/or scenario because specific reconstitution(s) of information can be delivered as needed to one or many different user, groups, parties, or processes.

[0515] Reconstitution is not a monolithic process, nor does it happen in exactly the same way for all dispersion locations and methods. Different dispersion methods utilize different storage locations and different reassembly protocols. Specifically, the DigitalDoors patents define several methods all utilized by the Secure/View DigitalDoors applications. Directory files (the map) details source information, recovery file locations, Secured file locations, and key file locations. The directory is optional and can be rolled up into recovery files. Note that various encryption options can be applied granularly to directory, recovery, secured, and key files, such that some or all are encrypted, encrypted with a one time pad for each file or each separate recovery selection line item, or encrypted through access to a public key infrastructure.

[0516] Security is herein rendered by reduction in metadata, format, encoding, and complexity. The information is dispersed, contexts are broken, and entropy is increased; this creates security. DigitalDoors does not affect security by methods presumed to be secure because they are proprietary, concealed, or obscured. The explanation for why this works can be found in the Third Law of Thermodynamics and the mathematics are completely consistent with statistical methods. Saltzer & Schroeder's Principles of Secure Design define an economy of design, failsafe defaults, complete mediation, open design, separation of privilege, and least privilege. The simplest mechanism to Secure and reconstitute the data streams, and simple enough so that everybody can understand the process. This explains the removal of metadata and formatting creating the simplest data stream that can be openly secures at all sensitivity levels and other dimensions of categorization.

[0517] Note that the reconstitution order proceeds from lowest sensitivity level to highest, from

lowest compartment to highest, and so forth because access to higher levels is presumed to be progressively limited and the reconstitution of higher sensitivity level elements depend on the prior placement of the lower level elements. This is a functional necessity of the previously defined multithreaded extraction process. Also note implementation minimizes the amount of the recovery metadata necessary to reconstitute a Secured data stream. It is possible to define recovery information through indexing, hashing, and positional designations more complicated than positional offset but only through the creation of more metadata compromising basic security that is the fundamental goal of DigitalDoors technology, complexity and slow speed of process, and the creation of new avenues of inference, aggregation, and interaction.

[0518] Security is effected herein by the destruction of context, the rendering of inferential clues, and dispersion of essential selections thereby leaving a redacted data stream that can be delivered as is or reversibly reconstituted under progressive granular control until it fulfills processing, sharing, or knowledge exploration needs.

[0519] A document can be redacted to a secure state and the recovery image of the extractions is created in situ as a reconstitution overlay, sort of like a positive/negative photograph. Extractions can be defined for each sensitivity level, compartment, mission, group, et cetera, and aggregated together to form one or more recovery files for each or every mix of granular specification.

[0520] A document can be redacted to a secure state and the recovery image of the extractions are created separately as line items within a flat ASCII recovery files. Extractions can be defined for each sensitivity level, compartment mission, group, et cetera, and aggregated together to form one or more recovery files for each or every mix of granular specification. The recovery file details for each extraction, its category, its starting location, its ending location, its length, any replacement length for

in situ substitutions (even if only one-for-one spaces), and the actual extraction. Note that variants on the format of the recovery file support overboard dispersion with separate encryption for each extraction applying unique keys for each. Details of the recovery variants are defined elsewhere. Recovery files can be stored locally or dispersed to MLS-compliant storage facilities.

[0521] A document can be redacted to different secured states (that is, versions) defined by each sensitivity level, compartment, mission, group, et cetera, with recovery data rolled up for recovery for each sensitivity level, compartment mission, group, et cetera. Rolled up recovery files contain line items up to and including sensitivity levels for each designated dispersed remote storage.

[0522] A document can be redacted to different secured states (that is, versions) defined by each sensitivity level, compartment, mission, group, et cetera, with recovery data dispersed to separate and unique storage for each line item. Note that the dispersion of line item extractions obstructs the basic intrinsic context to the Secured file while increasing the randomness (and quite literally the corresponding entropy within each package of directory, recovery, secured, and key file sets) and lack of context for each of the dispersed extractions.

[0523] The above outlined dispersion and recovery methods fulfill different functional and security requirements. The differing dispersion and recovery methods make tradeoffs in terms of reconstitution overhead, complexity, and processing with redundancy and accessibility to that offset the risks from aggregation, inference, and interaction. Although DigitalDoors enables security with functionality rather than as a linear tradeoff between the two, DigitalDoors dispersion methods engender a different tradeoff with reconstitution reliability. It is not a costless tradeoff, but it is a necessary one if you must enable a data stream security ecosystem within an open and functional network infrastructure lacking hardened perimeters.

AA.1.16.6 Directory Files

[0524] Directory files contain map information to recovery files and represent a physical and logical displacement to secured sensitive information in recovery files. This enables additional opportunities for granular security and privacy in conjunction with information sharing and knowledge discovery. The information is dispersed, contexts are broken, and entropy is increased; this creates security. You can envision this process like that of a railroad switching yard containing many uniform boxcars with different content. By dispersing content, the context and utility of the sensitive information is broken with respect to the source and the aggregation of randomized extractions adds to the implementation of security through the 4th law of thermodynamics. When directory files do not exist for a given redacted data source, the View applet seeks recovery files directly in anticipation of rolled up recovery files in any of the supported optional formats, accessible dispersion locations, and encryption or key management controls.

AA.1.16.7 A compound data stream that can be stripped into granular items

[0525] It is all about data. Data is the content qualified by a context and concept. Everything else is just proxy. Existing solutions are complex workarounds in the face of a unifying data theory within security ecosystem. A number of conceptual problems disappear with the simplification of workflow and communications into the storage, manipulation, and delivery of data. With this emphasis shifted to a data-centric view, we can successfully search, secure, and share with categorization, and better exploit information processing and communication workflow. The consequence of a simplified view is a simplified security environment not based on point solutions but that actually works as an ecosystem.

[0526] The complexity of information processing and communications, infrastructure, security, and data structure evolution have driven wedges into information technology workflow. As a result,

divergent insular solutions and the enforced separation of process and storage exacerbate the differences rather than lead to simplicity and integration. Insular solutions require additional effort not only to resolve the problems discretely for each area but also to create temporary bridges between areas, an effort disproportionate to the requirements. It is also counterproductive when false positives and negatives require frequent user override responses or when the point solutions preclude or corrupt workflow. This impasse is most evident with air-gapped legacy production facilities and the complex integration of distributed workflow. Additionally, insular solutions and bridges sow conflict, adding complexity to disrupt the workflow within each area or among them. Although vendors are integrating routers, guards, IDS, IPS, viral scanners into suites or unifying products, this approach addresses the proxy activity rather than the data itself, the fundamental basis for and result from activity.

[0527] Furthermore, virtualization and the distribution of storage, processing, and communications to the edge are showing that workflow and application processing are potentially independent but interrelated. Code is just a metadata package. In effect, job control code, macros, applets, and entire applications are structures that can be distributed and run almost anywhere. These structures are data, or metadata, and are delivered like any traditional data set. SOA and SaaS insinuate the viability of this theory. Any differences between a traditional notion of data and the application as data is summarized by the different methods used to collect, collate, distribute, and present them, but not at all in the data structure.

AA.1.16.8. Data Unification

[0528] Throughout this discussion, data is defined as any data stream regardless of internal format, purpose, intent, or structure. A data file is data. A document from Word is data (with imbedded metadata). An email message traversing a communications channel is data. An application stored on

local disk or invoked from a web-interface is a file and is also data. Anything that can be represented within a metadata structure is data even if it is all application code or contains macros with event-driven code. This paper asserts that applications are nothing more than data asserted as a code invoked through a process (like a CPU). It is a delivery and processing workflow. Applications become nothing more that destinations; rather than delivering data to storage, the data is delivered to the workflow.

[0529] This last assertion disconcerts some people who want to think that an application is a special set of instructions but overlook that it has structure and also because it often includes imbedded data is thus different entirely from a data store. The difference is purely contextual or conceptual, but the content is nevertheless the content of some data stream. Data acquisition, production, parsing, formatting, presentation, interpretation, categorization, and purpose are different between a pure data set and an application code, but this is also true between different data sets. Consider the difference between an MS Word document and an MS Access data set, but both are considered unstructured data streams.

[0530] Distinctions are completely contextual and predicated on what acts on the data sets, parts of the data sets, or how they are transformed for new and different purposes.

[0531] Is there a difference between an application and the data set? Of course there is, but differences are arbitrary based on usage and purpose. The difference exists mostly due to the historical development paths apparent within information technology. Most people are not likely to distinguish a difference, although support developers and architects will. But even developers and architects are users too most of the time, and these arbitrary and imposed differences become less consequential. Consider: A difference between a zip file and a self-executing zip archive? A difference between a document file, an Acrobat file, and an Acrobat file with a built-in e-book reader? A difference between

an application and a data set, a portable database application and the data set, and an application bound to the data set? Each can represent the same purpose, be repackaged in different ways, and be invoked in different ways -- implicitly or explicitly, or by the file type assignments in the desktop -- but desktop shortcuts mask such differences to the user.

[0532] All represent complex structures containing a traditional data core. Maps define execution entry points, module jump entry points, loaders and triggers for interspatial macros, and formats for parameter passing between functions. The map is (usually) a simple schema.

[0533] Consider that applications are specialized streams of data stored as a file. The structure includes a loader, the application, data space, buffer space, and increasingly the support for a digital signature validation of integrity. Malware insinuates itself inside the application file directly or as a jump vector to another file stored in memory or long-term storage. Although most data files preclude operative insertion of malware because it will force failure or be a visible attack, hierarchical and DOM data files can hide and execute malware through the built-in triggers. This malware lies dormant awaiting the specific activating trigger. Even when malware corruption is actually visible and obvious or violates the properties and attributes that are used to validate a data range, discovery is unlikely unless the actual data range is viewed or tested against acceptable properties and attributes.

[0534] Many workflows invoke the properties and attributes only at specific steps, for example, as a format immediately prior to presentation or after data is presented and altered by the user or process. Alternative steps can and do violate that intent, and this potential grows as data and applications are increasingly abstracted as hierarchical metadata and multi-purpose document object models (DOM).

[0535] Hierarchical and DOM data files frequently include links, references, macros, applets, helper tools, or even extensive applications. While it is equally correct to assert that data files are

becoming applications as it is to assert that applications are specialized data files, this paper demonstrates that is more useful and unifying to view everything as a data structure. In one sense, an application turned inside out is a data set with the application inclusion. Although the usage emphasis is on the application functionality rather than data content, the application is really nothing more than, and usefully visualized as, a compound inclusion within a data structure. Data complexity ranges from the simplest form to the hierarchical tree with imbedded compound inclusions, each of which can itself be a hierarchical tree with imbedded compound inclusions.

[0536] The most complex data structure we know about is DNA that includes some 4 billion combinations of UCAG amino acid types. Although DNA is clearly a data stream, it is a convoluted tree with hierarchical structures, external references, and imbedded compound inclusions. These imbedded compound inclusions are expressed as RNA chemical factories which create more complex amino acids and other precursors to life or even life itself. Aggregations and combinations of different parts of the same DNA chain or even external DNA, prions, proteins, free fragments of DNA or RNA, and other molecules with chemically-receptive sites for bonding to the UCAG pairs or the RNA receptors create new data, new triggers and events, and even new life forms. Other DigitalDoors white papers expand on the issue of DNA and RNA, MS Office metadata files, and the various document object models. The point herein is that DNA is primarily a data structure that is parsed and activated to initiate macros, applets, and full-scale applications.

[0537] This section defines data structures from the simplest to the most complex. Data structures include the simplest representation of data, such a single item. An item is represented by a range of granularity and purpose; it can be a value, an add and, a subtract and, a multiplier, a base, a numerator, a denominator, a single fact, and a DigitalDoors granular content item. Containment of even

that simplest of items is dependent on format, structure, purpose, infrastructure, and representation. The item could be the DNA UCAG or a compound UCAG-encoded molecule. More complex data formats include field-oriented data, such as a document with multiple position-oriented items. Padding, such as found in compound molecules, create the infrastructure to support these key items or bonding receptors. This document might be an e-mail message or a stripped announcement of a transaction, including a buyer, a seller, a product, a price, an extent; in other words, a line item transaction. The padding provides simple format or demarcation field control or more complex mixed-purpose usage for process functionality and also human readability. Field-oriented data formats support the transition to tagged data as seen with human-readable data files, application and system configuration files, or the MS Windows directory hive.

[0538] Tagged data transitions to row-oriented data transactions where the patterns among the records are repetitive with columns represent distinct fields.

[0539] These include sucrose, glucose, fructose and other energy sources for RNA process functionality. Although each sugar is different, the format is significantly similar. In the same way, row-oriented records provide the fodder for simple computerized processes.

[0540] Row-oriented data can include serialized or multiple complex indices (indexes).

[0541] Indices have multiple purposes both in terms of processing for assessment and interpretation for display. Indices are special items that provide context through association. When indexes become more complicated, they form the basis for locating information or establishing references between data within a structured database, either relational or object-oriented tables. The structure for the rows, the relationships between indexes and items, and other relationships are described by an implicit or explicit schema. The schema is an internal or a separate externalized skeleton that

supports the data structure. The issue of internalized or externalized schema, that is the application of structure and categorization, is purely an efficiency, reuse, and packaging issue. This is an another arbitrary complexity that is not needed for a data-centric model and the ensuing unifying security ecosystem.

[0542] The schema also defines the format and representation of items in terms of encoding, presentation, interpretation, and structure. For example, data can be encoded as binary, bits, decimal, ASCII, encryption, compression, or have meaningful utility only when used correctly. Binary large object (BLOB) fields can retain application code (source or machine code), sound, images, large documents, or other arbitrary data. The meaning for all of these BLOB fields is dependent on purpose, usage, format, structure, and intent. Consider executing a sound file or trying to hear an image.

[0543] When the schema is externalized or internalized as embedded tags and metatags within the data stream, the complexity of the data stream can increase. Hierarchical structures defined by a root node and branches form the basis for the DOM used by most modern desktop applications, such as the MS Office family. Not only does it support rich text formation, graphic layout, it also allows for unlimited complexity. The branches support items of any type, inclusions of any type, tags that define content, context and concept, establish links and references, or support compound macros, applets, filters, and code.

[0544] When the schema is referenced to an external template file, as with HTML or XML, the context is defined externally. When such schemas are self-documenting with tags, metatags, metadata, and the fractal-like self-similarity of nodes, the context is internally defined. However, associations to new schemas, (imposition of a new style sheet, for example), combinations with data from other sources, can alter the interpretation of the content, how it is contextual applied, and support new concepts not

obvious or overtly seen. This is the basis not only for data mining and search, but also for DigitalDoors-based security methods. Hierarchical files with both internal and external schemas pose new problems in terms of aggregation, inference, and data-to-data interaction, and this redraws the map of data content, context, and concept. Tags, metatags, metadata, comments, links, references, and imbedded code can be combined with intent improvised to disrupt, damage, or combine into new usage in conflict with anticipated point-security solutions.

[0545] Output from applications or complex SOA workflow are usually in the form of messages, code modules, reports, or human-usable triggers. These are also data streams in one format or another, but nevertheless still a data stream. Delivery of large hierarchical data is not different from the delivery of DNA. The execution of code within the data is not different from the expression of RNA. The paradigm shift is from viewing applications as a collection of methods and events to that of a compound data file with methods and events imbedded within.

[0546] When you accept this theory as to the predominance of data, you are well into your journey of the unifying data theory. Compound data is not useful until is parsed and applied per context correctly. For example, we previously asked what would happen if a sound file were executed or an audio stream viewed as a graphic. This is a process failure or disease situation. Therefore, we need to filter and apply each item by type. The process is best described as separating content from context from concept and dissecting compound formats into items, content into items, all of which are addressable by granular content control.

AA.1.16.9 Conclusion

[0547] The complexity of information processing and communications, infrastructure, security, and data structure evolution have driven wedges into computerized workflow. As a result, divergent

insular solutions and the enforce separation of process and storage exacerbate the differences rather than lead to simplicity and integration. Insular solutions require additional effort not only to resolve the problems discretely for each area but also to create temporary bridges between areas, an effort disproportionate to the requirements.

[0548] The primary benefit of unifying data is to see data as a compound stream that can be stripped into items. First, we can find information and apply it as intended or explore for new information. Second, we can assign tags to each item for security purposes. Third, we can determine what items to share and how to distribute the items in whole or in part, and within a generic data stream, application, document, or a structural context. The consequence of a simplified data view is a simplified security environment that is not based on point solutions but that actually works as an ecosystem. This supports a unified infrastructure based on categorization that enhances search, security, and sharing, but even more than that, it provides the same consistent, simplified, and functional basis for search, security, and sharing.

[0549] Why DigitalDoors security? What you are trying to protect. Is it the infrastructure, the communication channel, the application, the engine running the application, or the data? Today, we are trying to protect all of them at the same time with macro processes generally at the perimeter, with conflicting tools, and succeeding at very little. It is becoming increasingly expensive, time consuming, and disruptive within the current paradigm to impose security -- frequently counterproductively in terms of cost and degraded results. Instead, we need to recognize the convergence of application code and data, recognize that context and concept are the points of separation, and that security must be applied at the smallest discernable units. This explains the suitability for DigitalDoors granular content control.

AA.1.17.0 A New Information Security Paradigm: Neutralize the Loss of Perimeter

[0550] The existing shortfall in information security is indication of the amount of playing catch up now required. Information technology has always been driven by results in terms of accuracy, features, functionality, return on investment, and enabling wholly new business opportunities; security has always been the laggard afterthought. The current security metaphor pits functionality (or efficiency) in conflict to security, thus security is set at some practical level so that losses in functionality are tolerated to support security goals.

[0551] The desired balance between functionality and security changes with the perception of security threat levels. It is a lousy Hobbesian choice; a bad metaphor based on a flawed view of perimeter-based or point solution security, which is an evolutionary dead end.

[0552] Security goals usually lack formal definition, whereas products and services are tangible. Output is definable, the process is blueprinted, but where is security as an input, as an output, or as some multiplier effect on a production model. Security needs definition, perhaps as the ability to protect the resources and process that generates economic wealth within a production model.

[0553] It is all a matter of the paradigm you see. Security should not be a balancing act but rather a fundamental design parameter in information technology data and process. If you see security as elemental control, you see functionality as the enabler of security rather than a conflict. If you see security as a perimeter control, you see security as a stumbling block for functionality. This later view is the prevailing paradigm. Security is a stepchild to information technology rather than an amalgamated component for advanced functionality. In fact, effective element-wise security enables information sharing, multi-level access control, redistribution, controlled released and recovery options that can create totally new products and entirely business markets and models. DigitalDoors is doing all that now.

It is all a matter of the paradigm.

[0554] Recent high-profile failures in security, such as misuse of critical information, identity theft, and leakage of government secrets, the September 11, 2001 terrorist attacks, and the addition of new privacy directives and patriotic legislation have put the need for information security forefront. It is not that information has been unimportant and suddenly now is important; rather, there are limited methods and technologies for implementing information technologies that actually work. They work to a point and do not seem to be extensible to new situations or work beyond the borders for which they were invented.

[0555] Ad hoc efforts, band-aid policies, and application of best practices that aren't very good at all frankly reflect the really bad state of the art. Security as a step-child to functionality is a complex problem for governments, banks, brokerages, insurance companies, and many other fundamental human organizations that need to be revisited with new metaphors and a better paradigm.

[0556] Existing security technologies fall into three categories. They comprise access controls, encryption, and data or process filters. Access controls are based on usernames and passwords, or biometrics to a lesser degree. Encryption rebundles data into an impenetrable box opened only with an ungainly key management system that creates more problems than it addresses. Filters include virus protection, firewalls, and intrusion detection systems that search for known attack patterns at the periphery or once beyond that perimeter. All these systems represent a perimeter fortress mentality encircling the systems and data. It is a primitive burglar alarm that does not preclude inside thefts, accidental or incidental damage, or prevent destruction once the perimeter has been breached.

[0557] Any little hole or minor access is sufficient to provide full exposure with sufficient time, effort, and ingenuity. For this reason, top secret government data and processes are segregated

completely with air gaps from networks and external workflows. Of course, insiders like Hansen and Walker show the fallacy in the presumed effectiveness of such air-gap security methods. In fact, there is a common maxim that once the perimeter is breached, the soft underbelly is fully exposed, and the perimeter furthermore focuses the attack within to promote additional damage.

[0558] Sophisticated burglar alarms are augmented with internal sensors and content control. In fact, you see this with motion detectors, infrared heat sensors, glass-break detectors, and internal video and sound monitoring. Information security is developing ways to monitor internal processes and data access in a similar way. It actually makes sense, and is technological feasible to a point. However, cameras within the perimeter actually provide a venue for exploitation. Imagine have your built-in laptop camera showing the user, the user's location, other associates, and capturing the screen.

[0559] Advancing this burglar alarm concept further, museums understand the value of identifiable objects and alarm these objects to detect if they are moved or taken. Perimeter security has its place as a necessary first line of defense, but even for a museum, the perimeter is not invulnerable and not effective against an inside attack. Element-wise security a more sophisticated security system in that specific objects of value are designated as valuable and controlled individually. This is analogous for how data and process in information technology should advance. The information security paradigm requires the protection of individual data elements rather than a site-oriented perimeter access control.

[0560] There is unfortunately a great deal of difference between museum objects of value and data and process, and the corresponding methods of security. This paper details five major areas of logical discrimination. These are first, the logical similarity between data and process; second, the scale of data security dwarfs physical object security; third, data formats are becoming very complex; fourth, data elements combine to complicate security and privacy; and fifth, and information technology is

perniciously imbedded and intrinsically networked and cross-linked within workflows.

[0561] First, recognize that there is no difference in information processing between data and process. The process is series of steps that have been codified and represented in a data format. Therefore, a process is data. Worldwide, there are legal protections for process in terms of patent, trade secret, and copyright laws, although the emphasis is given to protecting new ideas rather than implemented concepts. This represents a serious shortfall that eventually will have to be addressed by new laws. Data on the other hand has limited protections. It can be protected by trade secret or copyright, but courts have limited the protections for structured and organized data gathered and owned by a custodian. However, the European Union has led the way in protecting individuals from misuse of such custodial data. In spite of this forward legislation, we need many new laws to protect data held by custodians and preserve the rights of ownership for personal data. This is especially important when process is perceived as data.

[0562] Second, control over an object of value is very different from the control over data. Damage, theft, or the reproduction of an object of value is obvious and limited to a single instance. Damage, theft, reproduction, or redistribution of data is not so obvious, particularly when data is replicated in multiple instances or dispersed as part of workflows. In addition, data is damaged just by exposure of that data, not so art in a museum, which is displayed expressly to be exposed. Exposure of data represents a new security threat, one of which you are undoubtedly well aware. The problem is that barrier-based security does not protect data from exposure for the previously outlined reasons and internal monitoring of process does not preclude digital reproduction, replication, or redistribution.

[0563] Furthermore, although a museum might have thousands of valuable objects, it is a discrete and controllable collection. Data on the other hand represents millions, billions, trillions, or more

discrete collection of elements. This data security problem is neither discrete nor controllable in any human scale. This is a complex problem for governments, banks, brokerages, insurance companies, and many other fundamental human organizations, and the bigger the organization and its processes, the greater the risk for exposure.

[0564] Third, data is changing from simple and primitive structures into compound hierarchical objects. You might have heard about metadata, and that is but only part of risk. If you haven't heard about object models, metadata, and hierarchical object structures, get with the program and learn about it quickly. The risk from objectified data structures is enormous and growing, albeit totally hidden from the security audit; it is solely at this point an architect and programmer issue. You already use object models if you use any application such a database, a graphical editor, or any word processing program.

[0565] For example, a Microsoft Word document is 75% metadata. That means that each document contains 25% of what you have actually typed but mostly things you would not believe when you review security risks. Since the data storage structure is partially unreadable binary, metadata is hidden. Some metadata is displayed from the file property pages; this probably includes your name, your organization's name, but the stuff that is not visible is a security nightmare. The name of everyone internal to the organization that proofed that document, and anything and everything you think you deleted from the document is still there. Metadata is only a part of the problem.

[0566] Consider also the external links, the metadata processes that are part of every Word document, the document blueprints, and the potential for embedded sounds, graphics, spreadsheets, and other data from other sources. The security risk is no longer discrete and part of any human scale. This is only for a Word document. There are far more complicated objects models and complications from data linked through the Internet and complex network workflows. Every imbedded applet has the

potential to be useful but also contain a Trojan Horse or destructive worm. Object models were developed for feature-rich applications and to enable accuracy, return on investment, and enabling wholly new business opportunities—remember that line?—but nonetheless are paradoxical to information security.

[0567] Fourth, we think of information security in terms of discrete units. For example, this might include a name, a social security number, or other elements that define an individual or overtly undermine security and personal privacy. We talked about the protection of discrete elements above and the failure of human scale in data protection. However, the problem is more egregious than the simple math of the number of discrete data elements. Realize that individual data elements are not only valuable individually but are even more valuable when combined with other such discrete elements. The sum of the whole is far greater than the sum of the discrete elements. For example, consider how a loan officer could correctly red-line your mortgage or make a business risk decision based on your blood type, your three-digit zip code zone, and your marital status without any other discrete personal information such as social security number, income level, or credit history.

[0568] There are at least three intersecting mathematical technologies you will hear more about in the future as part of any sophisticated consideration of data security and personal privacy. These are Set Theory, Polling, and Bayesian Inference. These mathematical technologies are not based on elementary sciences and thus seem like magic to most rational people. However, we take their results for granted everyday in election forecasting, weather predictions, that computer chips work, material management, and transportation logistics. Society would slow and stop without these statistical sciences. The bottom line is that these technologies undermine information security and exponentially drive any discrete security control from a human scale. For example, if you have just hundreds of discrete data

elements, you likely have billions of points of security exposures when information is extrapolated from the initial data elements.

[0569] Fifth, information technology has advanced from a simple one-for-one matching of data sets with computer processes into a scenario of multiple data sets interacting with multiple processes. This makes perimeter security impossible. With just 3 data sets and 3 processes, you have 27 permutations. With 42 data sets and 63 processes with a just 30% overlap, the security perimeter is represented by $1x10^{1943}$ permutations. That is a very big universe. Which perimeter—when there are so many? Data sets are frequently reused for many purposes or combined with other data sets. This is a principled result from Codd's database normalization concepts but also a fact of efficiency, necessity, integration, and advanced workflow development for accuracy, features, functions, return on investment, and enabling wholly new business opportunities.

[0570] Workflows intersect and overlap destroying any viable concept of a securable perimeter. Firewalls, intrusion detection, process monitoring for what? The perimeter is virtualized to the entire internal and external organization. It is a concept without reality. Of course, you could disassemble your networks, the Internet, distributed processing, and lock up each computer, although antithetical to what information process enables. You could add dongles for every piece of software to protect the data aspect of code and process. You could vet each and every user at each and every computer device. You could view functionality and security as mutually-conflicted goals and choice some balance between functionality and security. That is the metaphor in use today, it does not work, and will work less well as the security perimeter is virtualized beyond any securable perimeter.

[0571] The message is that security must be built from the inside. Data elements must be categorized individually for risk valuation. The exponential risk from many elements, data types, and

processes combining together must be assessed. Because the scale is beyond human scale, risk valuation, analysis, and security implementation must be applied by the information technology itself. It is not a human activity and correspondingly must be automated. The security must be built into the framework to bypass tensions with functionality, rather security must augment functionality instead of detracting from it. The paradigm suggests that data elements must be parsed and evaluated individually and within the context of the whole. The implementation thus demands a structure for measuring absolute security, for assessing risks and the requirements of each and every process, and applying a granular multi-level access control to individual data elements and sets or subsets of elements, or granular content control. This is certainly feasible. In fact, it has been implemented by DigitalDoors, Inc. Measurement is statistically-based. Likewise, the resulting security is statistical. Access control is multi-level and security is provided by an NP complete data dispersion implementation.

AA.1.18.0 Why security is ignored

[0572] The level of security provided in business processes and the production environment is rationalized to the lowest level possible to maintain its legitimacy. Although security is a desired attribute, it rarely if ever represents a constraint; security is perceived at best as a necessary evil. It has no return on investment (ROI). Attempts to create a security ROI rapidly unravel. Attempts to show how security itself or protection of assets as contribution to EPS (earnings per share) is really poor.

[0573] Attempts to ascribe an internal rate of return (IRR) for security in assessing alternative projects degenerate into a conflict as the security undermines the implementation of the selected project. Security is overlooked as functionality becomes essential. When security becomes an active constraint, it is removed from the equation. Security is usually optional.

[0574] Security impedes functionality. Even when it is understood that security is essential to

reputation, legal compliance, while avoiding fines and penalties, the risk that a breach is likely to occur is perceived to be so distant and so minimally invasive that business as usual trumps security. It is easy to recognize the lapses of security after the fact, but it is difficult to maintain the importance of security on an ongoing current basis.

[0575] Security is a fundamental building block for a sustainable competitive advantage. When security is couched in terms of its importance to creating a strategy, implementing that strategy, and maintaining an advantage within a competitive environment, or enabling cooperative sharing without revealing private and technical details, security still fails as an afterthought. People blame and then rationalize that it is security undermining success, lowering production efficiency, and contributing nothing immediately to ROI.

[0576] The subtle point is that there can be no ROI, no sustainable advantage, no multiplier effects in the production process, no development of patents, and protection for trade secrets without security. So while security has no current ROI, security is the integral foundation to those goals. It might be possible to perform an ROI-based functionality now with the additional costs of security versus the effects to future anticipated functionality with and without security. However, such long-term calculation of an annualized loss expectancy based on so many unknowns is tantamount to clairvoyance resulting in professional suicide. It is extraordinarily difficult in to explain the future at risk the boardroom when so much is instead measured in terms of current values and this year's bonuses.

[0577] In other words, security is necessary to the production model but not sufficient. Therein lays the rationalization to ignore security. Security does not create wealth. It is not an input to the production process. It is not even a multiplier to the production process, although it preserves the longevity of that multiplier. Security can be, and often is, rationalized as complication without portfolio

or return. However, security is the necessary and only ingredient that preserves the trade secret portfolio sustaining the competitive advantage provided by the wealth production model of the information revolution.

[0578] Without security, it is not possible to maintain the restricted knowledge that makes wealth generation from the information revolution possible. Without security, wealth generation reverts to the commodity of an industrial production model. In the end, security might be the only mechanism for protecting the innovation that is the driving engine of the information revolution. It is nothing but a mistake to rationalize security as a conflicting constraint with the immediate needs for functionality when security might well be the only thing sustaining a competitive advantage.

AA.1.19.0 Security integration and the lack of attention to it

[0579] In real practice, security is an adjective to data or an adverb to process. It is not a noun because security is not a viable standalone product or service. It is not a deliverable. Security can be delivered, per se, but only as a component of a product or service, the production methods to create those, or as method to protect and retain production trade secrets. Security is not an object. It is not a goal. Its importance is undeserved as a production input or output. It is important, but only with reference to the maintenance of a sustainable competitive advantage. Security lacks return on investment (ROI) while even an internal rate of return (IRR) is a difficult and sketchy equation. If ROI is hard, contribution to EPS is a long shot. Anticipated loss expectation (ALE) is calculable, but only in terms of losses to other quantifiable deliverables or resource inputs into profitable production activities.

[0580] Security does not stand by itself. Given the fear and attention given to all things security, it would seem that security was some concrete product and service, but it is not. Security is an attribute or property of other things. Security is not a data flow, a process that exists by itself, or an expression

of information. Security is not a discrete achievement and never apart from another discrete process. It cannot be divorced from other definitive deliverables. Security is not a building block or tangible foundation per se for other deliverables. Security is a state of control by management over access to superfluous, superficial, or fundamental information necessary to complete a task where a deliverable product or service is at stake.

[0581] Security is not a task by itself, of itself, for itself, or a reason for self-importance. The cover of security to protect a process is at best placebo. It seems that the integration of security for effectiveness is not something added as the event unfolds but rather somehow threaded within the structure itself. But it is not. Unfortunately, security is not a structure or a deliverable of any type, even as an outsourced service. Outsourced security services exist only as an attribute or property to a viable deliverable product or service. It is reckless and financially imprudent to outsource security without tie-ins to production. Security is only symbiotic with other resources and processes, never a resource or process unto itself. Security integrates into other activities, but it is not an activity creating product or service with standalone value.

[0582] Security is an economic entity. It is a marginal scalar without fixed asset value. Security does not alter production output value. Security is not a multiplier of product or service value. Security does not make a product or service worth more. Security has no basic asset or production value. Security does not make a better product or service. It only makes them sustainable. Security is not a production differentiator, but it is the sustainer of differentiation. Security protects the intellectual property value imbedded in the product or service or in the production methods used to produce them. Security protects an intellectual or economic monopoly as a multiplier to preserving this sustainable competitive advantage. Security is not a multiplier or marginal scalar of product or service value but a multiplier to

retain long-term differential value.

[0583] Security does not have a worth unless tied into the risks, potentials, threats, or anticipatory costs associated with production results or losses. If security is applied as a scalar to a worthless production activity, the output is still a more-expensively worthless activity. When security is imposed as a necessary process divorced from a clear attribution, as is often the case with the traditional security infrastructure, it applies to nothing, scales a null value, and generates no value added. Security can only be economically integrated into a process that warrants its value as a positive and non-zero value-added enhancement. Security is a long-term performance multiplier, never a performance enhancement that is integrated into a productive workflow and expressed in the information revolution as the necessary and perhaps only element to sustain a long-term competitive advantage.

AA.1.20.0 Security as the future foundation for information

[0584] Security is not a standalone entity. It is not a standalone activity. It is not a job or employment title separable from other tasks and workflows. Security is the basis for just about everything else. It is a foundation for all other activities. Security has no ROI, limited contribution to EPS, no real IRR, no real ALE, and no real payback. The value of security is continuity and having a future. Without security, there is no security at all. Security is a systemic requirement for success, for survivability and maintaining a sustainable competitive advantage. The problem with the future of security is that when we take it for granted we have none. When we think that security is an adjunct to our activities it fails and we have none.

[0585] Security is no longer something that is added to a process with the hope of any effectiveness. Security is not a chain and padlock that is separate from and attachable to raw materials, information, workflows, and production methods. Security is process that is unified with the production

process. With anything less raw materials, information, and intellectual property can be separated from the production process with ensuing leakage. When security is separable, not only is the production process attacked the add-on security itself is also stolen for its own value.

[0586] Security is part of the blueprint, part of the just-in-time (JIT) processes for handling raw materials, a protection over critical and proprietary workflows, integrated into the outputs, and part of the resulting products and services. Security is the process that identifies resources of value and, at least in our explicitly and unique view, disperses those resources beyond intrinsic value, immediate utility, and contextual recognition with a reversible process to recover that value when and as needed under ownership or custodial assess controls. That is a very important future of security, one that transcends merely physical assets and includes the increasingly more valuable ones of knowledge and intellectual property.

AA.2.0.0Design and implementation

AA.2.1.0 Implementation of Granular data control in data streams

[0587] Security with operational functionality and information sharing is possible by designating all network flows as data streams and controlling those streams. Applications and data of any kind are uniformly identified as data streams, since both are informational equivalents. Assessing an application, even a virtualized web-based one, as somehow representing a higher risk than data, overlooks the similarity in content and function of both, the higher-order representation of both as metadata, and their risk equivalence.

[0588] Security is possible through a process based on 1. identifying risk (with knowledge exploration) within the contained granular content of any data stream, 2. extracting the granular content at many defined sensitivity levels from the very matrix of any data stream, 3. providing access to a

redacted form of the data stream in its most primitive state and format, and 4. reconstituting that data stream on-the-fly as needed up to the sensitivity levels required to match the information, functional, and operational requirements.

[0589] This is the essence of granular content control. It enables the continuity missing from traditional paradigms based on the trade-off in security versus functionality. Because DigitalDoors allows granular control over arbitrary data streams, applications, and web-enabled applications, operation continue unabated. However, they may lack specific access to the crown jewels.

[0590] The crown jewels are defined specifically in relevance to each user, environment, and/or ecosystem, whether as sensitive information, risky applications, or metadata combinations of both. Redaction with substitution and recognition of required formatting enables continued operation even with gapped data streams, applications, or web applications. Functionality continues, not hobbled by the trade-off with security, but rather subject to realistic identifiable security constraints.

[0591] The Internet and the private but parallel GIG (Global Information Grid) represent the expansion of the localized network to everywhere at anytime. Any conduit, even guarded, represents a two-way transmission channel for the proliferation of sensitive information that traverses any edge, any perimeter, any geopolitical authority, or any functional demarcation. The idea that edge protection and perimeter protection can be effective within such an ecosystem without definitive boundaries is unsustainable. The idea that an arbitrary boundary for a community of interest (COI) can be defined within this ecosystem (the Internet, the GIG, or any attached subnetwork) to affect a perimeter security is also unsustainable.

[0592] While is certainly possible to define network edges and define a sharp boundary, this is possible only with complete isolation of distributed operations and the indiscriminate cessation of

information sharing, as indeed experienced under many existing security implementations today. Thus, any reliance on some sort of interspatial DMZ is just good intention without good security. DMZs include virtualized web applications, social networking sites, Wikis, remote management agents, and all Internet or GIG access, but DMZs are not a sustainable security technology.

[0593] Data streams and applications represent the top two internetwork traffic. These are necessary to sustain operations and information sharing regardless of the security barriers in place. SAAS, SOA, Web 2.0, Email, remote terminal services, virtualization, and remote services are examples of data streams with or without additional applications and with or without additional embedded data content. Unfortunately, data streams and applications are both forms of metadata with unlimited scope for covert channels. Traditional security implementations can only allow or deny access; it is a binary decision with no leeway for intermediate allowances. Therefore, traditional security either allows a security gap while enabling operational functionality or totally denies any throughput with the attendant non-operational effect.

[0594] Instead, DigitalDoors allows a redacted data stream with reconstruction. The allowable throughput can be set at any intermediate level (between none, some and all) with optional reconstruction up to any level (including some or all). Risk is adjustable, and risk can be measured and risk can be asserted. This specific aspect of DigitalDoors granular content control provides the missing functionality to enable compliant cross-domain security. In addition, automation of the selection of risk with the data stream allows for the unlimited assessment sensitivity levels and combinations of inference to make it possible to minimize risk with cross-domain information sharing and subsequent "uncontrollable" resharing with downstream partners. Cybercrime has shifted to attacking the plans, programs, designs of industry and national security.

[0595] DigitalDoors inventories the business practices, assess the need for business information, the risks from external threats and internal threats, the need for information in terms of the costs for acquisition, its retention, its reuse, and its potential for leakage and misuse both externally and internally. The resulting report details best production, contextual data value, information technology, and likely risk reduction practices. Security becomes a byproduct of better operations instead of a specific strategic goal or a forced mandate. Consider options deemed economically viable with competitive opportunities with a foundation of risk mitigation and control.

AA.2.2.0 Brief Functional Overview

[0596] DigitalDoors locates and selects critical information within a data flow, categorizes these selected elements, extracts them at any level of granularity, and disperses that critical information to different remote storage locations. The selected elements can be located manually (mouse selection), by dirty word lookup, or with sophisticated automatic tools to select content, meaning, and compounded risk factors. The downgraded data flow is stored in the lowest common risk factor and can be upgraded at wire speed to any authenticated MLS level. A trusted infrastructure with identification, authorization, and granular access control to granular storage represents a leap in security past traditional perimeter and containment methods.

AA.2.2.1 Brief Structural Overview

[0597] DigitalDoors was designed for networked and hosted infrastructures. The applications were designed and prototyped on MS Windows with IP networking, and widely demonstrated on that platform. We have shown automated content selection and categorization with AES encryption functioning on low-end laptops over an 11 Mbps wireless network with wireless SAN for remote storage and wireless Internet storage. RAID is supported in hardware. Because the user interface, the

downgrading engine, the upgrading engine, and other functions were modularized, DigitalDoors is able to function on standalone systems, networked systems, client/server architectures, and under remote terminal services. This has been demonstrated under Solaris v8 and Solaris v9 and TSOL v8. CMW labeling at the granular element level extends file, device, and process-oriented CMW labeling provided by Trusted Solaris. As the infrastructure migrates to Sun Microsystems, trusted applications running on trusted SunRays and trusted servers with dispersed trusted storage will enable multilevel side-by-side sessions.

AA.2.2.2 Remote Storage Overview

[0598] DigitalDoors downgrades data streams to any granular level as configured. This corresponds to the CMW definition of sensitivity level, compartment level, and information level. DigitalDoors also identifies data stream elements by content categorization, source, destination, and compound risk factors. The data streams are split in various configurable ways, with each stream or even each granular element stored to a different location. Backups can be automatically archived.

[0599] Elements can be replicated and stored remotely. The different remote storage locations are mapped as logical shares. This is fully compliant with MS Windows, and Solaris and TSOL with Samba, CIFS, and/or NFS. As the migration of the DigitalDoors infrastructure migrates to Sun Microsystems, NAS, SAN, iSCSI, FibreChannel, SNIA, and/or other storage methodologies, other data access methods will be supported in the same way. Currently, DigitalDoors supports at least 30 distinct storage locations for a single organization, with an unlimited parallel number for backup and replication.

[0600] The storage overhead for DigitalDoors adds only about 6% to the storage needs over baseline. However, organizational preference for archives, multilevel downgrades, separate sensitivity level storage, web-based or DOD distribution systems with catalogs each doubles the storage

requirements. Also, the need for failsafe operations, caching, proxy services, and redundancy is a separate and traditional storage provisioning issue. Five (99.999%) and six nines (99.9999%) statistical operational performance demand an infrastructure with many storage server farms and multi-homed networking access. The primary requirement for DigitalDoors is consistent and compatible storage facilities that are uniquely addressable. The secondary requirement is that this storage be granularized such that data stream access can be controlled by identification and authorization by user, process, sensitivity level, compartment level, and information level.

[0601] DigitalDoors is a series of differential process step that can occur on any platform, ranging from a mainframe to a low-end laptop. It enables content-level security in any type of data stream on any type of platform. It is inherently scalable from single to unlimited users. Content-level security can be affected at any layer of the ISO OSI model, from the bottom physical layer (level 1) to the top application layer (level 7). In addition, DigitalDoors uniquely functions within an application layer data stream to analysis content, concept, and context for security sensitivity to protect again brute force inference with statistical methods and data mining, assemblage of information, and human intelligence methods. In other words, DigitalDoors is processing any type of data stream for potential security lapses.

[0602] Most deployments are likely to require enterprise operation support with mixed client/server architecture. Automated OPSEC and COMSEC processing can occur as background processes on virtually any computing, Web/Internet, and networked platform. The DigitalDoors architecture was designed with run anywhere mentality, including hardware chipsets. However, in most cases, manual tagging presupposes a graphic user interface for selecting and identifying data stream content. As such, the core technologies require only visual element outlining, automated content

identification, list lookup and matching, and basic data processing. Processing can be offloaded to the client or performed on the server. Dispersion is a function of a secure networking infrastructure. Most operations occur at network wire speed, this being the primary performance limitation.

[0603] We have installed and demonstrated its effectiveness on currently deployed COTS environments. This includes a wireless laptop network with a wireless SAN, a 10BaseT Intel client-server architecture, a 100BaseT and Gigabit Intel client-server architecture (Win95 through Win 2003 AS and W2K EAL-4), several MS Terminal Server environments, and integrated with Solaris v8 and v9 and Trusted Solaris v8 servers and clients. We have tried the technology on Palm IV; we assume that DigitalDoors is portable to PDAs, handheld devices, cellular phones, and other hardware devices subject to their memory, connectivity, and CPU prerequisites.

[0604] For point of reference, we have integrated and demonstrated DigitalDoors with MS Windows, as add-ins for MS Outlook and MS Exchange, and presume simple functional integration for any MS Office, database, data workflow, signal transmission, or COTS products. We have demonstrated DigitalDoors with ASCII files, MS Word documents, HTML, and XML metadata.

The client requires only a basic GUI interface (at a minimum) with NetBEUI, NetBIOS, or IP networking over any network transmission medium. The server components require no GUI and can run as a background process with NetBEUI, NetBios, or IP networking over any inter-compatible network transmission media. Interprocess communications is enabled through client/server channels, OS messaging, and RPC. Support over LAN, WAN, Internet, VPN, NAS, SAN, with remote commercial storage services has been tested. DigitalDoors has also been successfully tested with encrypting network hardware (Intel) and native P2P encrypting protocols.

Table: The Data Declassification process

The Data Declassification process is defined by these steps:

Authorization of security levels:

Sensitivity

Compartment

Information

Categorization

Authorization of security roles:

Sensitivity

Compartment

Information

Categorization:

Identification of a data stream

Identification of the format for the data stream

Configuration of intended security levels

Sensitivity

Compartment

Information

Categorization

Categorization of data within the stream (e.g., tagging)

Selection of data within the stream for extraction

Extraction of data within the stream

Downgrading of data stream

Creation of upgrading recovery data files

Data backup

Data replication

Data dispersion

The Data Reclassification process is defined by these steps:

Authorization of security levels:

Sensitivity

Compartment

Information

Categorization

Authorization of security roles:

Sensitivity

Compartment

Information

Categorization

Access to downgraded data stream

Upgrading of data stream

[0605] Access security is established through a network ID and authorization process, such as Kerberos and enhanced with the compartmentalized trusted operating systems, such TSOL. Access security can be enabled for most networked devices.

[0606] Downgrading the data stream is manual or automatic, or assisted. Automatic OPSEC and COMSEC rules can be enforced prior or post manual selection (tagging of the sensitivity levels, etc.)

The processes are asymmetric by design. Downgrading can require substantial time, but upgrading is limited only by the network and the time lags to deliver a signal cross-country through multi-hop switches and routers that are part of the commercial public infrastructure. Realize that firewall, guard, filter, and router hops frequently entail 300 μsec unidirectional delays; this being the primary performance delays.

[0607] DigitalDoors has designed to the architecture and the process flows such that steps are optimized to each side of the firewall, guard, filter, router, and storage server mesh, including even optimization at the client for the extensive high-level content, concept, and context identification and dispersion. Manual tagging can be performed on any computer that supports a GUI interface. Processing is bound to disk or network performance limitations. Tagging information can be processed locally or delivered to a server for service-side processing. The user's ability to locate and highlight sensitive material is the functional performance limitation.

[0608] The data stream can be any of type of data. This includes signals, files, data streams, interprocess messaging in a workflow or high-volume transaction processing environment, email, instant messaging, logical information, and COTS data files. Data streams can be free form or structured, such as a COBOL data set, a RDBMS database, an MS Word document, or metadata, such as SGML, HTML, or XML. The DigitalDoors philosophy is to enhance security while maintaining traditional process workflows. Note that encrypted files and data streams require discrete and complete decryption before they can enter the transactional processing. Failure to decrypt such files abends systems and workflow because the format and content is unexpected. In addition, while encrypted XML and metadata is secure within that form, once decrypted is usually carries far more security-sensitive content than realized. Manual processing requires trivial overhead. CPU requirements are minimal for the client and server

components.

[0609] However, OPSEC, COMSEC, and enhanced automatic and assisted data selection requires additional CPU resources and approximately 400MB of storage. Even laptops are suitable. Automatically or assisted tagging information can be processed locally or delivered to a server for secure processing. Processing is bound to roundtrip network performance limitations, and database lookup seeks times. As the complexity is increased to automatically search data, metadata, content, context, and concept for sensitivity within the data stream, performs slows. Performance is affected by disk seek times and the depth of the searching and DigitalDoors categorizing process. Performance ranges from 34,000 to 360,000 instructions per recognizable element in the data stream. This ranges from 56 to 783 μsec on IBM Thinkpad laptop yielding a Winstone value of 23. DigitalDoors processes a document of 140 words in 70 seconds on that laptop, but within 19 seconds on Dual Intel CPU Server with SCSI drives (Winstone value of 89). A production SQL database server provides several orders of magnitude better lookup performance than a client database on a laptop; it also scales better. Experience with Sequoia or comparable scalable and fault-tolerant architectures indicate single digit μsec response times.

[0610] Dispersion performance is a function of network channel responsiveness and network storage performance characteristics. Wireless Ethernet 11Mbits/s down to 2.2 Mbits/s is sufficiently faster than database seek times and storage, but nonetheless effective in demonstrations.

[0611] ReClassification performance occurs at wire speed subject to the performance limitations of network infrastructure and remote network storage performance. We have recovered 2 MB documents within 3201 μsec using a multiple-hop router network and 10Mbits/s Ethernet hubs to Intel desktops. This is equivalent to raw file display on the same networks and systems.

[0612] The footprint is purposefully small. Storage requirements include application storage,

databases, temporary cache, and secure and dispersed network storage. Distribution is typically provided on a standard CD (640 MB) or business-card CD (120MB). User storage per user is as little as 15KB for a configuration profile. Application storage is approximately 30MB for clients, although this can be shared as a single store on a server or terminal server. Database storage requires approximately 231MB. This includes all language dictionaries, categorization elements, synonyms, and semiotic and grammatical inferences. Transactional processing, rollbacks, and security logs are optional, but on the same order of magnitude. All storage can use SSD, USB, PCMCIA, or flash RAM or networked devices. Performance is enhanced about 45 times with SSD. Flash RAM is marginally faster than mechanical hard drives due to lack of device driver optimization.

[0613] Data storage for user documents, data sets, data streams, metadata, and the like will be comparable to the size of the existing data store. Note that a design criteria is that all extra-security information, recovery files, auditing, and logs are stored in separate channels both for security reasons and format preservation. Expect additional overheads of only 6% for dispersion of recovery files. However, realize that minimum block sizes could increase that. Optimization for recovery files, which tend to be small, is possible by setting block sizes to 1KB or smaller on a special recovery file storage site.

[0614] When sources are dispersed and replicated for multiple location storage, or creation for individual TS, S, and C archives with support for compartments, information labels, and categorization, then each replication obviously increases the storage requirements by about 100%. Proxy caching is effective when documents are often requested; however, for security purposes, caching of recovery files is insecure. It it's also ineffective unless the recovery files are bound to the same sensitivity/compartment/category user because the storage server generates and only delivers at the level

dominated by the user clearance.

[0615] DigitalDoors assumes that the native OS encryption will be applied or that a certified encryption system is available within the network infrastructure. DigitalDoors will integrate with that facility. Encryption overheads are comparable to what exists now. When unclassified or downgraded documents and the recovery files are stored black and delivered black, the encryption overhead will be approximately 106% of what is now. In spite of that low overhead, this provides infinitely granular data storage and infinitely granular upgrading.

[0616] Interprocess communication and messaging for client/server is optimized to minimize traffic loads and point-to-point encryption overheads. This limited performance overhead is most evident when a terminal server environment is taxed by many users, who nonetheless see performance equivalent to that performed on the client desktop. User or process identification and authorization is interspersed throughout the DigitalDoors process—it is not a one time when the user logs in—so there is some ongoing validation chatter. This rate is granular and configurable by the security administrator on a range from seconds to hours.

[0617] There are true performance and storage benefits when downgraded documents are shortened and mechanically compressed. However, for functional reasons, a core design of DigitalDoors is that the data stream format is typically preserved intact (less extractions or substitutions) to allow for workflow processing without abends, interrupts, or crashes. Even HTML, XML, and other metadata are typically preserved to retain the original format so that DigitalDoors is providing security without compromising functionality.

AA.2.2.3 Operating Parameter

[0618] Any CPU platform with a development platform and as little as 64KB of RAM. Any

platform with support for either C, Java, or VB is suitable. DigitalDoors can run as a hardware process, an applet, an add-in, process filter, a dynamic link library, as an NT or RPC service, as a standalone client, as an in-process or out-of-process server, and as an enhancement to a firewall, IDS, IPS, guard, or operating system. DigitalDoors interacts with Kerberos, Active Directory, smart cards, biometrics, encrypting file systems, trusted operating systems, and many NAS, SAN, and other remote filing systems.

AA.2.2.4 Network Design Specifications

[0619] Because the core concept of DigitalDoors is to support collaboration, information sharing, dissemination, Internet and Web services, data mining, and data dispersion to remote storage, the client/server infrastructure requires only a core network. Integration preference is for a COTS network environment, such as Internet, Intranet, LAN, MAN, WAN, NetWare, NFS, or Microsoft networking, Network protocols can include NetBEUI, NetBIOS, and IP on a physical channel of 1Mbits/s wireless (Zenith), infrared, Bluetooth, 802.11a or 802.11b, or 802.11g, Ethernet, FDDI, iSCSI, frame relay, ISDN, DSL, T1, T3, OC3, and SONET.

AA.2.2.5 Remote Storage Infrastructure

[0620] DigitalDoors locates and selects critical information within a data flow, categorizes these selected elements, extracts them at any level of granularity, and disperses that critical information to different remote storage locations. The selected elements can be located manually (mouse selection), by dirty word lookup, or with sophisticated automatic tools to select content, meaning, and compounded risk factors. The downgraded data flow is stored in the lowest common risk factor and can be upgraded at wire speed to any authenticated MLS level. A trusted infrastructure with identification, authorization, and granular access control to granular storage represents a leap in security past traditional perimeter and

containment methods.

AA.2.2.6 Brief Structural Overview

[0621] DigitalDoors was designed for networked and hosted infrastructures. The applications were designed and prototyped on MS Windows with IP networking, and widely demonstrated on that platform. We have shown automated content selection and categorization with AES encryption functioning on low-end laptops over an 11 Mbps wireless network with wireless SAN for remote storage and wireless Internet storage. RAID is supported in hardware. Because the user interface, the downgrading engine, the upgrading engine, and other functions were modularized, DigitalDoors is able to function on standalone systems, networked systems, client/server architectures, and under remote terminal services.

[0622] This has been demonstrated under Solaris v8 and Solaris v9 and TSOL v8. CMW labeling at the granular element level extends file, device, and process-oriented CMW labeling provided by Trusted Solaris. As the infrastructure migrates to Sun Microsystems, trusted applications running on trusted SunRays and trusted servers with dispersed trusted storage will enable multilevel side-by-side sessions.

AA.2.2.7 Remote Storage Overview

[0623] DigitalDoors downgrades data streams to any granular level as configured. This corresponds to the CMW definition of sensitivity level, compartment level, and information level. DigitalDoors also identifies data stream elements by content categorization, source, destination, and compound risk factors. The data streams are split in various configurable ways, with each stream or even each granular element stored to a different location. Backups can be automatically archived. Elements can be replicated and stored remotely. The different remote storage locations are mapped as logical

shares. This is fully compliant with MS Windows, and Solaris and TSOL with Samba, CIFS, and/or NFS. As the migration of the DigitalDoors infrastructure migrates to Sun Microsystems, NAS, SAN, iSCSI, FibreChannel, SNIA, and/or other storage methodologies, other data access methods will be supported in the same way. Currently, DigitalDoors supports at least 30 distinct storage locations for a single organization, with an unlimited parallel number for backup and replication.

[0624] The storage overhead for DigitalDoors adds only about 6% to the storage needs over baseline. However, organizational preference for archives, multilevel downgrades, separate sensitivity level storage, web-based or DoD distribution systems with catalogs each doubles the storage requirements. Also, the need for failsafe operations, caching, proxy services, and redundancy is a separate and traditional storage provisioning issue. Five (99.999%) and six nines (99.9999%) statistical operational performance demand an infrastructure with many storage server farms and multi-homed networking access. The primary requirement for DigitalDoors is consistent and compatible storage facilities that are uniquely addressable. The secondary requirement is that this storage be granularized such that data stream access can be controlled by identification and authorization by user, process, sensitivity level, compartment level, and information level.

AA.2.3.0 Upfront Processing

[0625] DigitalDoors front-loads document and data stream processing. This upfront processing is dense, comprehensive, and deliberate. Such overhead is necessary to explore multitier categorization and security requirements beyond a limited perspective, potential emergent properties, and classification bias. Reconstitution occurs at wire and disk access speeds with enhanced asymmetric performance.

[0626] There are five reasons for this design architecture listed in this document. One, the comprehensive parsing of the document or data stream and all associated metadata, links, attachments,

inclusions, and structure provides a greater accuracy in the security assessment. Two, the comprehensive parsing balances content, context, and concept issues and provides a process milestone for assessing and balancing competing needs of security against sharing. Three, the front-loaded processing provides a focal point for word stemming, syntactical assessment, clustering, metasearching, index construction, data mining, data spanning, construction of topic maps, and a means to deal with the typical overabundance of information flow from a source (e.g. handling the fire hose concept of signal and data collection). Four, this architecture enables data, document, stream, process-, platform-, system-, network-, and grid-independence; the process is independent and allows anywhere, anytime, multi-platform access subject to access authentication and authorization procedures. Five, source documents tend to be processed once and thereafter distributed and read many times and this design optimizes for this.

Table: Upfront processing delivers:

Swarming intelligence to access aggregate value, risk, and granular value at the edge
MLS/MILS assessment
Determine what to protect and how much
Determine who should get indication of the information (directory or metadata)
Determine who should get information
Information aggregation and inference
Information prioritization as to what is needed now
Ability to redistribute multi-tier tagged materials without additional intervention
Support for inter-agency SBU tagging conflicts

AA.2.4.0 Proxy architecture

[0627] The DigitalDoors Identification and Authorization Server limits access to data and

process by vetting users through an identification and authorization process, a traditional security by perimeter and process control.

[0628] The DigitalDoors Clearinghouse Server resolves site IDs, vets for identification and authorization, and establishes access level for global reconstitution. The server matches the site ID and provides the recovery map and reconstitution files from wherever they are stored. It returns only the reconstitution files as allowed by the lowest condition of {local, networked, or global security} status and the designated security clearance for the requesting user. The clearinghouse server can also compare security hashes or digital file signatures to validate the integrity of the request for reconstitution files to preclude attacks using random site IDs and message keys.

[0629] The proxy server caches and handles reconstitution requests by providing a narrow service window to preclude non-vetted browsing, data storage mining, and bulk access to maps and reconstitution files. Multiple servers can sustain any level of performance necessary with network wire speed the most relevant performance-limiting factor.

[0630] The DigitalDoors Dispersion Server is a vital security function because it creates a division between each DeClassified document, often a document or other information sources downgraded to the lowest common denominator for public dissemination, and the corresponding recovery map and reconstitution files. It delivers only what is authorized specifically and minimally for each user. Dispersion security is significantly enhanced by creating a separation between the recovery map and the referenced reconstitution files and by providing specific multi-level access to a delivery stream of these vetted reconstitution files. System-level control with unique one-time pad encryption keys provides effective protection from intuitive forced re-assembly.

[0631] Although primary security is affected through the information hiding, dispersion, and

camouflage, the proxy architecture with the identification and authorization server, a generic proxy server, and the dispersion server precludes human intelligence operations, hacking for information value, and sophisticated inferential analysis applying set theory or Bayesian statistics against protected organizational data. In addition, this proxy architecture enables a distributed and replicated design such that network or Internet support is always available and available anywhere with complete multi-level security.

AA.2.5.0 Reclassification- Reconstitution

[0632] ReClassification of declassified materials, in effect reversible redaction, requires new paradigms because the process is not at all considered in the security or workflow literature. Because the DigitalDoors process declassifies documents using various pattern recognition, work and phrase lookup, and artificial intelligence, and then disperses the extractions to various locations, the reclassification (reconstitution) process is granular. Granularity is a function of location, user, time, threat modes, supervisory overrides, and other configurable limitations. This reclassification process protects the recovery map, recovery extracts, and other storage locations by a process of dispersion and granular storage access.

AA.2.5.1 ReClassification

[0633] On the receiving side, DeClassified e-mail, documents, or data are received as a standard e-mail message, documents, or data that required no special treatment, process, or interface. Size, shape, format, encoding, markup, and processing integrity are preserved. Data include formatted data sets, transactional data, or formatted XML and database tables, records, and queries, DeClassified e-mail, documents, or transactional data is the same as standard e-mail since there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for

ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration, as a secondary security access issue. However, the ReClassification process requires a downloadable user interface in the form of an e-mail client reader add-in, the use of our separate e-mail reader, or an integrated document or data reader.Selection of a DeClassified e-mail message, document, or data transaction initiates a workflow process consisting of:

Table: Workflow process

Identification and Authorization (either local or remote)

Lookup of recovery map information

Access control by time, location, threat mode, username, category, time of day, level hierarchy

Access to the recovery map (granular access controlled at all time)

Access control to dispersed storage locations

Recovery of dispersed classified critical and important information

Recovery of the encryption key per organization, per user, per document, and/or per extraction

Decryption of encrypted or MultiCrypted critical and important information (on the user side at the user machine through proprietary, standard, SSL, or other encryption add-ins)

Reconsitition (ReClassification of DeClassified e-mail message)

Display within standard user interface of ReClassified e-mail message

(such as within e-mail client, PDA, etc.)

Display within separate e-mail reader/ReClassifier application

Display within separate reader/ReClassifier application

Display within separate user application (MS Word, database, PowerPoint, Quicken)

[0634] Controlled release for ReClassification purposes can be limited such that different users will have leveled access to the extractions, for example, senior users can have access to the names and credit numbers on a granular basis, whereas junior users cannot.

[0635] Workflow and transactional messaging activities that require e-mail, POP3, SMTP, IP, or other similar message flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways. Pagers, wireless messaging, and other token delivery methods can be incorporated into access control, authentication, and differentiation for ReClassification levels. The DeClassification process and the display of the ReClassified e-mail, documents, or transaction data does not have to occur in any set or preset location. ReClassification can occur at network devices (in whole or in part) and be delivered to that same location or to other locations. Accumulation of recovery information need not occur at any one single location.

AA.2.5.2 Recovery Information

[0636] The recovery information is not a necessarily a single entity or file, or accumulated in one place or by a sequential single process at any single time. The recovery information is a starting point that describes the location of the recovery map. The recovery information for each declassified document is independent from all other declassified documents. Access to that information is either known or, not known but delivered after secured authentication. That location can be communicated in various methods, including system message, e-mail, by phone, encrypted as a message or semaphore, attachments, or other means. Access keys, encryption keys, and rights to that location are also communicated in various like methods.

[0637] The location might contain the entirety of the recovery map as a data record, table, XML document, or file in plain text, encrypted, or second order pointers to other locations. Support for other

locations is arbitrarily infinite. The recovery map might optionally contain information about the classified source document and the declassification and extraction process. Key sharing for storage access, authentication, and recovery is enabled with {L:N of M} key-sharing schemes. Typically, the recovery map defines offsets to extracted information (offset from initial starting point of the document, content structure, or data stream, offset from last extract, or from arbitrary tags and markers), lengths of the extracted information, lengths of any replacement for the extraction, extraction type, and the extraction itself. The offset is functional in all known types of media formats including text, formatted text, XML, audio, image, video, multimedia, database, and object-oriented descriptions. The starting point from content structure could include the offset from a natural break in the data stream format, a format block, or change in content type or structure.

[0638] For example, an MS Word document contains a metadata structure within which the user content is embedded; the reference to the user content forms the natural offset starting point. This is analogous for other types of documents, such a spreadsheet, data query, or OLE document embedded with the MS Word document. Two-pass placeholders support reconstitution in the event a DeClassified document is altered (such as edited, attacked, or damaged) subsequent to the DeClassification. This enables ongoing data processing with secured data and protects DeClassified data even in the event of a concerted semantic attack.

[0639] All of this information can be in plain text, encrypted, or represented by pointers to other storage locations. Numbered tag or two-pass tags (such as barcodes or URLs) allow ongoing editing, as with MS Word, wherein placement and even the existence of extractions could be altered in processes subsequent to DigitalDoors DeClassification, thus allowing alteration and functional ReClassification of the DeClassified document at any future time.

[0640] Support for other locations is arbitrarily infinite. Each unitary recovery unit for each extraction can be dispersed to an arbitrary storage location. Each storage location can be managed as a closed storage community, with access authenticated at a granular level. Storage locations in whole or in part and individual extractions can be distributed to enable survivability. Failure or loss of any single recovery information or extracted element does not preclude partial ReClassification or failure of any other ReClassification for others documents. A sufficient dispersed survival scheme provides 100% recovery in the event of catastrophic failure.

[0641] The location of each secondary or subsequent storage location is independent from all other storage locations. Access to one a particular storage location, file path, file name, access key, encryption key, or other access control (such as biometric token) in no way provides information about or access to any other location. Access to each secondary or subsequent storage location is also granular and separated authenticated. As a result, access to any part of the initial recovery information, even in plain text, without access to storage locations and access authentication prevents full or even partial recovery of the classified document. Plain-text access to any and all storage locations alone precludes manual recovery of a declassified document using human intelligence, as the magnitude of the accurate placement and reference is astronomical (the problem is N!, non-polynomial, or NP complete). Access to a complete plain-text recovery map with plain-text access to any and all storage locations could compromise a single declassified in full or in part, but no other declassified documents.

[0642] Authentication is granularly revocable at any time now or in the future. Access to classified documents and the reclassified levels is controllable at any time in the future. Access to recovery information, recovery maps, individual extractions, locations, and access to storage is both granular and revocable.

AA.2.5.3 Processing with DeClassified Data

[0643] The starting point from content structure could include the offset from a natural break in the data stream format, a format block, or change in content type or structure. For example, an MS Word document contains a metadata structure within which the user content is embedded; the reference to the user content forms the natural offset starting point. This is analogous for other types of documents, such a spreadsheet, data query, or OLE document embedded with the MS Word document. Two-pass placeholders support reconstitution in the event a DeClassified document is altered (such as edited, attacked, or damaged) subsequent to the DeClassification. This enables ongoing data processing with secured data and protects DeClassified data even in the event of a concerted semantic attack.

AA.2.5.4 Content Filtering

[0644] Content filtering includes declassification downgrading but also more typical filtering, document retention, and e-mail attachment retention policies, as implemented by other e-mail products. The major difference between typical filtering technology and DigitalDoors, is the ability of the e-mail, document, or data recipient to potential review the e-mail message and all attachments reconstituted in full, not just as filtered and delivered. The core engine supports classification by word type. Forty categories are currently supported. We are adding politically correct, hate mongering, job resumes, pornography, drug culture, games and jokes, spam, and others as specified. DigitalDoors can filter outgoing and forwarded e-mail for .com, .exe, sys, .ovl, .dll, .ocx, .prg, scr, and script attachments and even search within these data streams for known file infection patterns. Message subject, content, and attachments are processed. In addition to standard content filtering, DigitalDoors also assesses target e-mail addresses including to, cc, and bcc. The workflow is basically one of reviewing inbound and outbound messages for content. When excessive amounts of e-mail messages, documents, or

transactional data cannot be processed in real-time, they are parked for later processing in a temporary folder. Administrator (not user) configuration defines whether copies are made, saved, or deleted. Reports explain actions taken and the event logging databases provide analysis into the e-mail declassification and filtering process.

AA.2.5.5 Channel Transmission Protection

[0645] The transmission channel(s) for source data to the user's desktop, server, or transaction process, external generation of the recovery file, and external generation of a recovery map are all authenticated with multi-factor authentication, such as PKI certificates, tokens, or other methods, and encrypted with point-to-point encryption of the transmission channel with SSL or in hardware. This process can be controlled by DigitalDoors or be part of the core networking environment and independent from DigitalDoors functionality.

AA.2.6.0 The Recovery map

[0646] DigitalDoors enables information security and privacy protection through the selection of critical elements, the granular extraction of those critical elements to dispersed storage, the distribution of the redacted (and encrypted) information while preserving functional work flow format. The system enables reconstruction of the original document and the multi-level reconstitution of the document to users with different clearance levels and different rolls based on their identification and access authorization. Selection of critical elements can be done: automatically, fully-manual, or assisted with automatic features and manual overrides. Reconstitution and multi-level reconstitution is driven by a "recovery map" that details the selections, replacements, and pointers to actual extractions that are restored with appropriate authorization.

AA.2.6.1 Map Structure Variants and Granularity

[0647] The DigitalDoors "recovery map" can be in different forms depending on application, configuration, encryption, signatures, and administrative options. The map in effect is a starting path to a file (database record, script-macro, or data stream) represented by an unqualified or fully-qualified path name and file name, resource reference, a URL, or an XML pointer. The path might not even be recorded, or it might be encrypted. This initial file might contain no references whatsoever to the source and/or redacted file, or even overt references to the collections of associated files and processes. Each file in the collections of associated files and processes might be encrypted in whole or in part, encrypted extraction-wise, element-wise, or distributed as additional collections of associated files and processes, dispersed to an arbitrary number of networked storage locations.

The "recovery map" is partitioned to improved reliability, replicability, survivability, integrity, validity, security, and preclude forced inferential reconstitution activities. The "recovery map" exists as a series of steps and separated pieces of information, with minimal top-down dependence and also stepwise and separated independence. Each step and dispersion further precludes reverse engineering the source and decreases the amount of contextual value to an extraction recovery file. There is no down-top dependence. Each step, each price of information, and each file is independent from all others with no or limited reference to the original source, redacted document, or extraction files. File names, file locations, file date and time stamps, file sizes, and file ownership ring rights might be altered to preclude source document group association, user relationships, process or procedural connections, event correlation, or indication or sources and methods.

[0648] Options for the redaction and how extractions are represented within the redacted document are important in achieving a high-level security while maintaining interoperability. As such, replacements for extractions can be null, randomized in both content, context, and size, or substituted

with misleading information. Because the various recovery maps can also include efficient minimal or fully chaffed listings of all document elements, combinations of redacted documents with even plain-text recovery maps can severely hamper forced inferential reconstitution activities.

[0649] The DigitalDoors recovery map information in its simplest form contains the offsets to the extractions, length of extractions, length of redacted document replacements (which can range from 0 to thousands of characters). This file, database record, script-macro, or data stream might be encrypted in whole or in parts, encrypted extraction-wise, element-wise, or distributed whole and in parts. Encryption is through various algorithms including 3DES, Rijndael, or any third-party encryption product. Granularity with randomness, chaos in association, and dispersion might make the need for encryption necessary merely to meet legal requirements. The recovery map is not necessarily a single entity, but rather usually a collection of associated files and processes, dispersed to an arbitrary number of networked storage locations each accessible through hierarchical controls. The redacted file must have an association to at least one recovery map under any expectation of full or partial reconstitution. Existence of the recovery map alone does not provide any ability to access, decrypt, or apply the map for increased comprehension of the redacted document through forced inferential reconstitution activities. Access to the recovery map is controlled by identification and authorization, by 1-factor, 2-factor, 3-factor authentication or third-party products integrated or callable within the DigitalDoors process.

[0650] Metaphorically, you might think of the recovery map as a pirate treasure map or such a map broken in small pieces so that only a sufficient number of pieces have practical value. Furthermore, without ownership, map location reference (the starting point), a time relevancy element, a time frame in history, an understanding of the map relevance to a treasure, the age of the map, or the authenticity

and integrity of the map, you have only an interesting artifact. Sometimes it is possible to infer one of those elements and thereby establish values for other elements and thus possible context to the map. Consider that you have thousands of such maps,you have only the context of enforced secrecy and no singular context for any one map, let alone many of these maps. A mapped mismatched with an anticipated treasure is not likely to provide the right results, even for pirate treasure. Given a correct context and a correct starting point for the map, you still need direction and scale for perspective. Now consider that the map provides location for a treasure, you will still need conveyance to the treasure (network access), ability to get the treasure itself (access control), and a means to break open the box containing the treasure (decryption key).

[0651] The availability of new information analysis techniques requires that security experts counter threats not only from the obvious but also the probable, possible, and unknown. Information analysis includes not only capture of native communication streams, but also human intelligence methods, assemblage, data mining, intuition, interception and alteration, hypothesis construction and validation, logical inference, Bayesian inference, polling, and set theory. These methods represent a dramatic advancement through the statistical quantification of threats and unknowns with decision-making based on game theory and risk management. As a result, information security methods must include protection against the obvious and also the probable and possible. New information security methods must be flexible enough to protect again the accidental, intentional with systems knowledge, unforeseen, and unknown as well. Concrete security methods do not anticipate probable, possible, and wholly unknown and unforeseen attacks.

[0652] The range of known to unknown, expected, probable, possible, promising, and the only-likely predicate the need for new security methods. As a result, DigitalDoors feels the best defense is

based on statistical methods, chaos theory, Bayesian inference, and set theory. DigitalDoors is aware of these attacking technologies and inverts these same techniques to counter these actual and perceived threats in both defensive and offensive processes.

[0653] The DigitalDoors technology includes the recognition of information as content, context, and concept. Content is the obvious textual or streaming information. Context is the broader information gleaned from work flows, situational utility, how the content is gathered, created, processed, used, and distributed, and what other content establishes an informational basis for semiotic connotation and denotation. Concept is created through process structure, data structure, data tagging and encoding, metadata, and reference and inference to the textual information environment. The ultimate situation is a large number of independent extraction recovery files, generally inaccessible by users, further individually encrypted with one-time pads, which even in aggregate indicate only activity but not content, context, or concept. This explains the value for selection, extraction, granularity, and dispersion.

[0654] A relationship exists between the source document, the recovery map, reconstitition extraction files, process flows, encryption/decryption steps, integrity checksums, digital signatures, key generation, key distribution, and anonymity. Note that separate and independent channels can mask any relationship among the source, its recovery map(s), and its extraction file(s). Independence, randomness, dispersion, secure erasure (per MilSpec), optional date and time stamping erasure, initial creation of random files to preclude sequential dated backups analysis for creating grouped file associations.

[0655] DigitalDoors protects the redacted documents from the recovery maps, and these from the extraction files by implementation of chaos (such as, the law of large numbers), the independence of the files, file names, location of files, access to files, and the infinite granularity of individual recovery maps and extraction files. In other words, a recovery map can be processed through the DigitalDoors

technology n-times to create n-levels of indirection. In the same way, extracted components can be dispersed n-times to create n-levels of access control and indirection.

[0656] DigitalDoors technology is implemented under MS Windows. Client, server, and remote components exist, support networking, dispersion, encryption, and add-in integration with other Windows-compliant applications. The technology is generally partitioned into user interface, processing, and storage modules, thereby foreseeing deployment as remote, host-based, server-based, and appliance-based code. For example, the aspects within MS word that create insecurity, links to external sites and other documents, imbedded digital content in formats not known to DigitalDoors, redlining, metadata, and hidden elements would be automatically excluded, the remainder would be redacted and published, and the source Word file itself could be purged from the system or stored within a secure footprint. Since some of the necessary processes compromise the nominal security of the desktop and the communication channels, different parts of this process would occur within different vetted environments; the only insecurity would include what is actually visible on the screen at any time and/or saved as PrntScrn, and the parts of the document buffered in RAM or disk swap space.

[0657] DigitalDoors technology can be deployed to secure environments within an open and insecure network infrastructure (IP over Ethernet and Cisco routers) on notoriously insecure components (laptops, PCs, handheld devices, and general-purpose appliances) using COTS applications (MS Word, MS Powerpoint, and new file sharing features embedded in Windows 2003). Application of digital signatures, key management with tokens, proxies, and special key servers provides the necessary infrastructure to vet the integrity of a process work flow.

AA.2.7.0 Concurrency in granular data control

[0658] Because the code base for the DigitalDoors applications is based on object-oriented

programming, many processes, functions, features, events, and user options are inherently concurrent. However, many processes, functions, and features include an active process status test to disable certain aspects of this concurrency.

[0659] This prevents process clashes (for example when two or classifications processes run at the same time on different multiprocessor threads) but primary enhances consistency and security to contain risk. Concurrency and parallelism represent new doorways through the perimeter, an obvious failure in any perimeter security model. Reduction and control over concurrency enables the new deperimeterized DigitalDoors security model by controlling what and how much goes through the doors. This is granular content control.

[0660] For example, it is possible to start the Secure process on a document and leave the computer; the process will complete, but no manual activity or options are allowed once the timed login has expired. Likewise queued documents will be processed safety in the background with an active user login, but the process is locked into whatever administrator and user options are in effect at the timed login expires.

[0661] Likewise, many categorization functions will preclude manual overrides until they are completed. This prevents clashes and security lapses should a user try to distort the tagging. Generally, the automatic recoding feature will code to the highest sensitivity level even when prior processes assert some tag lower than the highest level in the various databases. This helps achieve stability and consistency as well. Manual overrides against the document itself or the relined selections can always occur at any stage, but not during an automated process. However, note that COMSEC will preclude any manual overrides, as it is meant to do, so as to assert that demanded highest organizational sensitivity level.

AA.2.8.0 DeClassifying and ReClassifying E-Mail

[0662] Security for workflows, transactions, and E-mail requires new paradigms because it is only partially addressed by current encryption methods. The current technology fails to address workflow, convenience, and distribution issues. Specifically, before and after encryption, the security gap reveals a window of opportunity to attack the plain-text. DigitalDoors specifically addresses the security of workflow, transmission, and e-mail with methods to distribute and provide granular leveled access to information. The two control points for securing e-mail with DeClassification are represented by the generation and distribution of the e-mail and the receiving and ReClassification (reconstitution of the critical and important information) of the e-mail.

AA.2.8.1 Generation and Distribution

[0663] There are several issues when securing e-mail. First, is the issue of sourcing e-mail. Although this generally occurs at the e-mail server or the e-mail client, frequently other applications directly initiate the e-mail transmission from within the applications themselves. Internet enablement and interprocess messaging represent serious growth areas in application development, for which DigitalDoors is expressly useful. We implement DeClassification e-mail generation from within applications directly by initiating the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Second, we intercept the generation of e-mail and DeClassify it on the the fly by initiating the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Attachments can be unpacked and examined, declassified, and reprocessed as attachments. Third, we intercept outboxes and outgoing e-mail caches in order to initiate the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Classified information is encrypted, MultiCrypted, and dispersed to local and/or remote storage

locations. Storage of critical and important information is configured and dispersed to different storage locations including Internet and hosted storage options. Dispersal is controlled by policy, configuration options, and threat mode based on users, recipients, situational needs, e-mail content, and critical and important information categorization. For example, an e-mail containing names and credit card numbers can be DeClassified with each of the two (and other categories) dispersed to separate storage locations.\

[0664] Distribution represents no changes to standard e-mail process in any form as standard address book processes are used. We can match the address book, names, or e-mail addresses against known threats or alter processing subject to the target addresses. Remailing could initiate a further downgrading (or even upgrading) of content when compared against target addresses. Distribution is nominally assumed to exist within a decentralized, distributed, and non-secured environment. There is no special need, effort, or consideration for distributing DeClassified e-mail as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes.

ReClassification is a function that is controlled and limited by the standard identification and authorization process. This includes the use of agents, and leveled access through secret sharing, alerts, threats, location, timing, policy, access to remote storage, and dispersive storage.

AA.2.8.2 Receiving and ReClassification

[0665] On the receiving side, DeClassified e-mail is received as a standard e-mail message that required no special treatment, process, or interface. DeClassified e-mail is the same as standard e-mail as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration.

[0666] The ReClassification process requires a downloadable user interface in the form of a e-mail client reader add-in or the use of our separate e-mail reader. Selection of a DeClassified e-mail message initiates a workflow process.

Table: DeClassified e-mail message workflow process

Identification and Authorization (either local or remote)

Lookup of recovery map information

Access control by time, location, threat mode, username, category, time of day, level hierarchy

Access to the recovery map (granular access controlled at all time)

Access control to dispersed storage locations

Recovery of dispersed classified critical and important information

Decryption of encrypted or MultiCrypted critical and important information

Reconstitution (ReClassification of DeClassified e-mail message)

Display within standard user interface of ReClassified e-mail message

(Such as within e-mail client, PDA, etc.)

Display within separate e-mail reader/ReClassifier application

[0667] Controlled release for ReClassification purposes can be limited such that different e-mail users will have leveled access to the e-mail message, for example, senior users can have access to the names and credit numbers on a granular basis whereas junior users cannot.

[0668] Workflow and transactional messaging activities that require e-mail, POP3, SMTP, IP, or other similar message flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways.

AA.2.8.3 Forwarding and Redelivery of E-Mail

[0669] The standard ReClassification process protects access to DeClassified e-mail while it is ReClassified. This retains control over information flow. Redelivery, forwarding, or remailing of the e-mail message prevents general exposure to critical and important elements, as only the DeClassified message can be forwarded or remailed. All subsequent readers of the e-mail will see only this DeClassified version as a minimum in security. However, access to the partial leveled or fully ReClassified document will require the standard access identification and authorization process.

[0670] It is also possible to embed within the browser tools--assuming control over the e-mail reading environment within an organization--indicators to further DeClassify e-mail prior to forwarding or redelivery of previously downgraded e-mail messages.

AA.2.8.4 Content Filtering

[0671] Content filtering includes declassification downgrading but also more typical filtering, document retention, and e-mail attachment retention policies, as implemented by other e-mail products. The major difference between typical filtering technology and DigitalDoors, is the ability of the mail recipient to potential review the e-mail message and all attachments reconstituted in full, not just as filtered and delivered. The core engine supports classification by word type. Forty categories are currently supported. We are adding politically correct, hate mongering, job resumes, pornography, drug culture, games and jokes, spam, and others as specified. DigitalDoors can filter outgoing and forwarded e-mail for .com, .exe, sys, .ovl, .dll, .ocx, .prg, and script attachments and even search within these data streams for known file infection patterns. Message subject, content, and attachments are processed. In addition to standard content filtering, DigitalDoors also assesses target e-mail addresses including to, cc, and bcc. The workflow is basically one of reviewing inbound and outbound messages for content.

When excessive amounts of messages cannot be processed in real-time, they are parked for later processing in a temporary folder. Administrator (not user) configuration defines whether copies are made, saved, or deleted. Reports explain actions taken and the event logging databases provide analysis into the e-mail declassification and filtering process.

AA.2.8.5 Processing E-Mail

[0672] DigitalDoors was envisioned as a solution for security in complex workflows as opposed to point solutions for files, records, or databases. As such, the modularized structure of the suite, the support data structures, and databases, and the flexibility for semaphores, interprocess communication, and command line arguments have allowed us to integrate the DigitalDoors suite into Microsoft Outlook, one of the most proprietary and awkward software development platforms imaginable. As a result of this integration effort, DigitalDoors for Outlook runs virtually transparently as part of the normal Outlook workflow. In fact, it is a structured as a standard .DLL add-in library.

[0673] New messages are DeClassified prior to outgoing delivery and received messages are automatically ReClassified for users according to security levels. Recovery information widely dispersed and distributed on VPNs or Internet-attached networks are easily mapped into the received e-mail for real-time display. Reconstitution times compare to wire speed and are several magnitudes faster than traditional encryption technologies. Revocation of access, revocation of messages sent in error, and location-specific security is enforced as with all DigitalDoors technologies.

[0674] As with most workflow integration efforts, maintenance of security and smoothness of operations dwarfed programming issues. System timing, order of process, access control, workability, and other reliability overshadowed code concerns. Interception of events inside the Outlook framework proved the most fearsome constraints; these problems have since been overcome. However, it is

important to document the workflow for security reasons.

[0675] All DigitalDoors process within Outlook are transparent where possible, all standard outlook functions, features, and workflows are maintained intact, and we succeeded at not altering the user interface with the exception of several toolbar buttons for online help and specific DigitalDoors functions. The creation of a new or forwarded e-mail message occurs without change in the user interface. An additional DigitalDoors user logon is required for all DeClassification and ReClassification consistent with the DigitalDoors desktop Suite. Administrative and user controls are configured through the DigitalDoors Suite. Outlook username names, MS Exchange interfaces, e-mail addresses, passwords, and other Outlook-specific settings are maintained and function identically with or without the DigitalDoors for Outlook.DLL add-in.

[0676] While the external workflow is not altered from the user's perspective, several internal changes are important to document. First, e-mail messages are constructed through four primary mechanisms. First, they are created as new. Second, they are forwarded or responded to. Third, messages are constructed as drafts. Fourth messages are copied and pasted. Other workflows are possible with MS Exchange, other Microsoft workgroup messaging tools, and third-party .DLL add-ins. User-created or third-party .OPM or VB scripts should function normally with the DigitalDoors for Outlook.DLL installed, however, we frown on such tools as significant security holes and for the potential conflicts with other add-ins, overlapping scripts, and potential to create holes for viruses and opportunities for script kiddies. Our initial scripts were converted and compiled into the more-secure MS-preferred .DLL add-in formats. Note that VBA, VB scripts, and other tools can undermine Outlook security, the Outlook security patches that prevent scripts from running, and create many security concerns. We suggest you avoid that path. Nevertheless, the DigitalDoors process remains the same throughout whether other

scripts are installed.

[0677] New message events are trapped. The outgoing Outbox message is moved into a cache folder for processing. DeClassification processing occurs per administrative configuration. The message is parsed completely by means of the MS Outlook message object structure. A temporary input file is constructed. This file is passed through a system shell function to the DeClassify application. The temporary file is processed by this application and a temporary output DeClassified file is created. The recovery map and files are also created and dispersed per administrative configuration. The DeClassify application deletes all traces of the temporary input file as per administrative configuration. This includes overwrites per SAS, Orange Book, and other security standards for file erasure. The DeClassify application completes. The temporary output DeClassified file replaces the contents of the original outlook message. The message is returned to the Outlook Outbox folder and delivered per standard Outlook configuration.

[0678] ReClassification occurs when a user selects an Inbox message. If the message has been DeClassify, the selection event is intercepted and the temporary input DeClassified file is created. This file is passed through a system shell function to the ReClassify application. The temporary file is processed by this application and a temporary ReClassified output is displayed. All user-specific DigitalDoors security measures are applied to the level of granular recovery, access controls, location, and other limitations. The DeClassify application deletes all traces of the temporary input file as per administrative configuration. This includes overwrites per SAS, Orange Book, and other security standards for file erasure. The ReClassify application completes.

AA.2.9.0 Declassifying and ReClassifying while Browsing the web

[0679] Security for workflows, transactions, and information stores while browsing the Internet

and Intranet webs require new paradigms because it is only partially addressed by current methods. The current technology fails to address workflow, convenience, and distribution issues for public and private information, the granularity of delivery for that information, or the centralized and decentralized access to that information. DigitalDoors specifically addresses the security of workflow, transmission, and browsing with methods to distribute and provide granular leveled access to information. The two control points for securing browsing with DeClassification are represented by the de facto delivery of declassified information with access to immediately reclassified information subject to secondary identification and authorization.

AA.2.9.1Generation and Distribution:

[0680] There are several issues when securing browsing and access to data (web pages, tables, data sets, and linked files). First, is the issue of sourcing information and downgrading it for general presentation. DigitalDoors implements DeClassification for all data stored directly by initiating the DeClassification process with multiple secured outputs where the declassified data is an output. Second, access to the ReClassified material is generated on-the-fly subject to identification and authorization. This is accomplished by initiating the DeClassification process with multiple secured outputs where the declassified data is an input along with the critical and important extracts.

[0681] Classified information is encrypted, MultiCrypted, and dispersed to local and/or remote storage locations. Storage of critical and important information is configured and dispersed to different storage locations including Internet and hosted storage options. Dispersal is controlled by policy, configuration options, and threat mode based on users, recipients, situational needs, browsing content, and critical and important information categorization. For example, a static web page or even one generated dynamically from a database containing names and credit card numbers can be DeClassified

with each of the two (and other categories) dispersed to separate storage locations.

[0682] Distribution represents no changes to standard browsing process in any form except for dynamically-generated pages require a dynamic recovery file. Distribution is nominally assumed to exist within a decentralized, distributed, and non-secured environment. There is no special need, effort, or consideration for distributing DeClassified browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. Because all static pages are stored in the DeClassified format, access to all pages in a ReClassified format is on a per page basis.

[0683] ReClassification is a function that is controlled and limited by the standard identification and authorization process. This includes the use of agents, and leveled access through secret sharing, alerts, threats, location, timing, policy, access to remote storage, and dispersive storage.

AA.2.9.2 Receiving and ReClassification

[0684] On the receiving side, DeClassified browsing is received as a standard data that required no special treatment, process, or interface. DeClassified browsing is the same as standard browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration.

[0685] A healthy security state for an organization is the state in which its files are redacted/ scrubbed and are in essence frames populated with skeleton content that allow continuation of operations.

[0686] The ideal security scenario for an organization is the state of formlessness where almost all of the information in a file is extracted and the result is a state of formless the form has evaporated.

A withdrawal of sensitive extracts calls for a highly organized process. To achieve that every content element needs to be automatically or manually categorized, and inventoried. The categorization of all content elements in the file is the basic and beginning step. Following that extraction of sensitive elements takes place. The extraction can be done manually, automatically of a blend of both manual and automatic.

[0687] Automatic extraction is based on combination of different semantic filters.

AA.1.14.0 Taking advantage of a major economic imbalance

[0688] Society gives lip service to information being a valuable asset but in reality it is mostly considered valueless. As an example the Internet which contains many Billions of documents which are there for any taker for free. Another example is the current accounting methodologies do not value information as an asset on the balance sheet.

[0689] The lack of recognition of the value of information leads to:

[0690] 1) Lack of inventorying of information assets in organizations (60%-85% of the data is unstructured)

[0691] 2) The lack of information inventory precludes from using current information assets as economic assets as well as from building upon those information assets to create more assets. The current environment of valueless information assets precludes us from monetizing those assets. The bottom line is we are losing economically by not being able to inventory our existing assets, and not being able to monetize them.

[0692] The crucial imbalance in the basic recognition of information as valuable assets and thereafter an underutilization of such information assets is an economic flow and an unreasonable aberration. Understanding such an imbalance and taking advantage of it will lead to substantial economic

gains.

[0693] Society for 20 years has focused on the perimeter on the tools storing and processing information and not on the substance it self: the information and the content. The focus on the computing and storage tools was spearheaded by vendors who have made hundreds of billions of dollars selling their products. Dealing with the content became the niche of only few small companies. The resulting situation is that 60% -to 85% of the information of organizations is unstructured, which means the information is un- inventoried and therefore mostly unusable.

[0694] In essence even with trillions of dollars of purchases of IT tools and capabilities over the past 2 decades most organizations can't access process and leverage between 60 to 85% of their information.

[0695] Organizations can't show Internal Rates of Return (IRR) on their information investments because they really never treated the information as an asset class on its own merit with an economic value. There was no inventory, therefore no value could have been attached to those assets and now there is nothing to show for.

AA.1.15.0 Monetizing the organization's information assets:

[0696] To take full advantage of the major economic imbalance in information the following steps are proposed:

[0697] 1) Inventorying the information assets

[0698] There is a need to establish an information inventorying process in the organization. The same way that retail stores or supermarkets inventory all their merchandise using barcodes, information assets throughout the organization must be inventoried. All information assets must be classified and tagged. This inventorying effort must cover, structured, unstructured and semi structured (e-mail)

information.

[0699] 2) Data mining - enlarging the information asset base

[0700] After assets are accounted for in an inventory, organized efforts must be undertaken to enlarge the base of those assets that are important to the organization.

[0701] The subjects of importance will be inputted into a data-mining engine, which will search the Internet and other sources for relevant data assets.

[0702] 3) Leveraging automation and legal power – the Innovation Production Line

[0703] After relevant assets are located, they will be analyzed automatically and categorized into assets that are to be treated as trade secrets- (potential contacts, leads etc to be stored in data bases.) and ideas that are of important values and need to be secured by patent protection.

[0704] Innovative ideas will be processed automatically and inserted automatically into templates of Provisional patents. The provisional patents will be sent automatically by e-mail to the patent office with automated payment.

[0705] By inserting legal framework such as patents trademarks etc into the newly formed information products the value of the information product grows substantially. To maximize the value of the created information assets people should be involved in the process to twick the input the system receives for example inputting additional key words.

[0706] 4) Leveraging web scalability – the Innovation Factory

[0707] To take full advantage of being early in the market to benefit from the big information imbalance, there is a need to leverage the scalability of the web. Specifically to leverage the new social graphs capabilities.

[0708] Combining an automatic information assets production line with interaction of many

people across the web to fine tune the production process will create a powerful combination – bringing an automated information creation process with the innovative input of many people.

[0709] 5) Monetization

[0710] As the organization creates information products, the organization can sell such products to partners as well as to the public over the web.

[0711] Proposed model is using DigitalDoors controlled release process to sell parts of the created assets over the Internet. For example releasing first part for free the other parts for payment.

AA.1.16.0 Information disassembly and Reassembly

AA.1.16.1 Disassemblying Information

[0712] Anyone can easily encrypt, shred, delete, or destroy information so that it no longer retains utility. However, DigitalDoors takes information apart, specifically information represented by any arbitrary data stream, removing any utility, so that it can be reassembled for later use. Information is selected in various ways, selections containing a paragraph, sentence, line, word, or character, and are extracted from the data stream.

[0713] When information is taken apart, the message is changed, obscured, broken, and the entropy of the parts is increased. Entropy is the solution to government-sponsored terrorism, criminal-initiated attacks for financial gain, or preserving the economic stability of entire sectors or even nations. This creates a flexible, enduring, and inherently independent security method. This information disassembly is consistent with the need to implement privacy, confidentiality, resilience, continuity, and survivability in the ecosystem of previously inconsistent goals of mutual security, information sharing, and knowledge discovery.

[0714] Information reassembly ranges from zero to full reconstitution in minute rheostat degrees

under granular control. In other words, you can get none, some little, some more, or all of the data stream back. Information is encoded with one or more systems of labels and tags or classifications

[0715] DigitalDoors supports a system of arbitrary, unlimited, overlapping tags so that preplanned, far-ranging, or multipurpose utility is supported, and a utility that is not preconditioned on information source, intended usage, or the application of increasingly sophisticated knowledge management tools that can rework or expose information secured under existing systems.

[0716] Tagging is supported both as in-channel codes, a metadata, and out-of-channel markup. The out-of-channel is the preferred method because of its enhance security and lack of additional information for inference. Granular control means that arbitrary systems of user access control can be applied to the rheostat. Granular control includes sensitivity level, compartment, word or phrase categorization, role, group membership, syntactical element, mission, location, time of day, payment or trade, and other tags or labels.

[0717] Realize that tags condense and synopsize data streams, usually within a cultural context, and tags represent a denser risk than the initial content. Separation of content, randomization of context, and limited connection to tags are important security features.

[0718] Why disassemble information? DigitalDoors does this to enable security, privacy, confidentiality, functionality, survivability, and compliance, among other reasons.

AA.1.16.2 The reasons for disassembly

[0719] Security establishes a state of freedom from danger, harm, or risk of loss.

[0720] Privacy is instituted by a method, desire, or process to maintain personal, business, or corporate identity without releasing it to others.

[0721] Confidentiality is an expectation and workflow process that ensures that the content of

a transmitted message cannot be determined except by the possessor of a key associated with the message, even if the transmitted bits are observed.

[0722] Survivability means that dispersed information can be regrouped to reconstitute the whole.

[0723] Compliance means that the technology fulfills the legal or organizational requirements to protect and retain confidentiality of data and process protection.

[0724] Information generally exists within a structure dictated by purpose or the software tool used to create, collate, organize, store, distribute, or share it. The structure can be a document format, a file format, a database table or many associated tables, a Microsoft Word document, or an explicit structure (such as a TIF image format).

[0725] More often than not the structure is a metadata hierarchy. All of these structures, whether defined internally or externally or by the application, are ultimately a purpose-specific implementation of metadata. Even where these structures are buried within a code library or specific instructions that shave off pieces of information from an application data file, every organization of any data stream can be converted and represented by a metadata definition, and, is ultimately a metadata hierarchy.

[0726] A metadata hierarchy is a security, privacy, and confidentiality risk combined with but also apart from the data stream. The problem is that metadata is an all-purpose file structure until itself that also defines data streams within, redefines purpose and usage of imbedded data streams, and supports the inclusion of functionality. Metadata is so flexible and generic that infinite layers of purported but obscured utility can mask, distort, or hide data within. Often, the metadata hides more information and risk than the expected data stream within. Metadata in any form, whether data, application, or mixed from both provides powerful clues that enable aggregation, inference, and

interaction undermining traditional security methods.

[0727] DigitalDoors removes this metadata and creates raw data streams in primitive forms that can be secured. DigitalDoors can secure the data stream and reintroduce the metadata. However, reintroducing metadata restores the risk from the metadata, maybe not the same metadata, but enough that recreates measurable and significant security risk. However, disassembling the data stream specifically and reconstituting the information as needed enables a wide range of security, and depending on the substitution techniques applied, also maintains the consistency and functionality of the data for most information processing operations.

[0728] Note that some structured financial data can be secured in this manner -- although the need to maintain posting, trial balances, and consistency between accounts and roll-ups complicates -- but does not preclude -- the process of creating secured but working data streams through selective data extraction and granular content control.

AA.1.16.3 Finding Value within Information

[0729] Value of information is found in the data stream and through its meaning, utility, and purpose. Value is tied to paragraphs, sentences, phrases, words, or characters. However, value is also greater than the sum of the parts, so that aggregation, inference, and interaction both internally and with additional external data sources represents an important security consideration. Value is modified by encoding method and can be covertly enhanced by the encoding method or metadata structure. Every computer file exists in storage in a linear representation of "0s" and "1's" that ultimately is organized or clumped into units of meaning.

[0730] The data stream is an information flow with a starting point and a length. It does not matter how it is encoded or how the data stream represents information. Text, text in any language, text

defined by words or larger units of paragraphs, numbers, names, words, images, sound, video, signals, data sets, and so forth are all represented by a data stream. One of the key benefits of the DigitalDoors technology is the recognition of risk in units far less monolithic than an entire document or document package. Granularity is achievable at word, character, symbol, or image bit levels. As such, security can be effected so that the choice is not whether to distribute a document or not, but what parts of the document can be distributed to what sensitivity levels, to which users, and under what controls. Documents do not have to be rewritten, ghosted, or endlessly tear lined but can be automatically processed once and distributed in Secured form or granular reconstituted streams at any time now or in the future. Recallable granular content control overcomes the time element to existing security paradigms; so in effect you can "put the cat back into the bag."

[0731] DigitalDoors support uncontrolled secondary or tertiary distribution without fear of compromise in sources, methods, security, or repurposing.

[0732] DigitalDoors applies automatic methods, white lists, gray lists, black lists, complex Boolean and placement associations, symbol recognition, categorization to value selection and subsequent information extraction. These lists (actually database entries) need only be consistent with the type of data stream, the language, and the encoding methods. An organization can create multiple databases for different languages, for example, and apply one, some, or many. Database entries to match sound, images, or video overlay apply just the same way as text, although sub-message element selection is specific to channel, encoding, or fuzzy matches in terms of non-linear placement. This is pertinent to mixed media, as in sounds or images.

[0733] It is important to recognize that selection is multitasking, supporting manual, assisted, and automatic features, supporting any order of use or reuse, with manual editing and removal of any

and all selections at any time with reapplication of any of the manual, assisted, or automatic methods again. Note, however, that COMSEC recognition finalizes selection that cannot be overridden, for obvious reasons of enforcement. However, new selection methods (not currently imagined) can added into DigitalDoors to augment or enhance the recognition of aggregation, inference, and interaction.

AA1.16.4 The Doctrine behind DigitalDoors

[0734] As explained, finding information in a data stream is necessary and, independent from the DigitalDoors process. DigitalDoors applies automatic methods, white lists, gray lists, black lists, complex Boolean and placement associations, categorization, manual user selection, manual overrides, and can included other methods as they are developed and proven effective. However, after selection takes places, all of these selections must be reconciled for overlap and conflict. We do this by analyzing selections in a table.

[0735] All selections are defined by a starting offset and an ending offset. This technology is consistent for all types of data streams and encoding methods. Extracting the selections is not always simple or straightforward because excising the information is contingent on foolproof reconstitution and the application or MilSpec security techniques embodied by the Bell-Lapadula Model (BLM), Saltzer & Schroeder's Principles of Secure Design, and others.

[0736] It is a significant complication applicable to the architecture of security ecosystem design for the military or for business. DigitalDoors stores selections that are extracted in one, many, or distributed locations. Reconstitution can occur in steps (like defenses-in-depth), up to a predefined sensitivity level, or to full reconstitution. Because of this, starting and ending offsets as defined by the selection process, change as a function of selection priorities and positional precedence. Selections are not independent until DigitalDoors extracts them and creates a recovery file. Selection dispersion creates

data independence by virtue of contextual dissociation; this can be further enhanced uniform data stores by creating random chaff. In other words, the position of all selections is a complex simultaneous choreography that depends on the selection and categorization of all prior selections.

[0737] For example, selection priorities and positional precedence becomes important when extracting selections to create directory, recovery, and secured files. This can be a complicate multi-path choreographed hierarchy for conformance to cross-domain MLS information sharing ecosystems. Specifically, certain security and versioning options requires a multithreaded pass first ordered by position (starting to ending) with each sensitivity level (starting to ending) for the recovery files to conform with BLM. Then certain security and versioning options requires a second multithreaded pass ordered only by position (high to low) to create the secured files since each subsequent selection is dependent on the (missing) positions of any and all prior selections.

[0738] For example, consider this well-known 35 letter pangram (e.g. holoalphabetic sentence uses every letter of the alphabet at least once).

Table: Hologralphabetic Sentence

1 5 11 17 20 27 32 35 40

The quick brown fox jumped over the lazy dog.

[S] quick

[C] brown

[TS] fox

[S] dog

and by default the unselected remainder is [U] The jumped over the lazy.

Sensitivity level: [U] < [C] < [S] < [TS] and defines the necessary order of reconstitution.

brown is removed first and coded at position 5 to reflect positional dependency on quick.

`quick` is removed next and coded at position 5 without any positional dependency.

`dog` is removed next and coded at position 37 to reflect positional dependency on fox.

`fox` is coded and removed last and coded at position 17 because it is conditionally independent of all three other extractions.

On reconstitution, "brown" is inserted into the 5th position to yield

The brown jumped over the lazy.

On reconstitution, "quick" is inserted into the 5th position to yield The quick brown jumped over the lazy.

On reconstitution, "dog" is inserted into the 37th position to yield The quick brown jumped over the lazy dog.

[0739] Reconstruction is dependent too. The position of all selections is a complex simultaneous choreography that depends on the selection and categorization of all prior selections. This explains the dispersion variants and various levels of available conformance to BLM. Full support for all possible failures and complications is just not possible as this is N! problem. As a result, some variants provide survivability with full recovery or only partial recovery with a full security breach. There costs, speeds, and overheads fully reflect the level of service provided.

[0740] On reconstitution, "fox" is inserted into the 17th position to yield The quick brown fox jumped over the lazy dog. Its placement is contingent on the positional dependency of the three lower sensitivity level selections, but its displacement is unchanged because all precedence conditions have been previously met.

[0741] Linguistic syntax and inference issues plays into the automated selection and extraction

of information. Names, locations, compound terms, data formats, common cultural phrases, aphorisms, white space, punctuation, numbers, quotations, parentheses, modifiers to nouns and verbs, and other constructs need to be excised in most cases so as not to leave a telltale gap in the Secured file. For this reason, substitution without any form of placeholders is advisable and careful removal of indications of special data formats is also indicated. For example, the removal of numbers within "05/06/23" into the substitutions " / / " or "//" is very obvious and informative. Removal of the entire syntactical expression is indicated.

[0742] For example, the removal from "former president William Jefferson Clinton" leaving only "former president William" provides useful re-identification clues in that there are only a limited number of former presidents, few have the first name "William", and fewer still fall within a recent time period warranting security by partial selection and extraction of information. Inference requires assessment of selection by the full phrase, an entire line, an entire sentence, an entire paragraph. It is a complicated assessment of additional information within the data stream that refers to the sensitive item and exactly how it and all references need to be excised. As previously stated, information value is not specifically internalized because other internal content, external data sources, a priore knowledge, context, and concepts are useful for establishing aggregation, inference, and data-to-data interaction. In addition, usage, intent of secret protection, distribution, and secondary information sharing policies all dictate the actual scope of selection, selection categorization and sensitivity level encoding, and the granular control allowed in reconstitution.

AA.1.16.5 Reconstituting Information

[0743] At this point, information meets with policy, security law, governance issues, and compliance rules. Other existing security technology can provide similar results by tear lining and

encrypting the streams; however, information is locked up and generally accessible only through a primitive filter that offers only the choice of all or nothing. This is also true of indexes and catalogs, which like tags, represent a condensation and synopsis of the original content. Instead, DigitalDoors provides functional Secured files and additionally granular content control for getting more of that information out to people and processes. Reconstituting files with granular content control means data dispersal and information sharing is not an either/or scenario because specific reconstitution(s) of information can be delivered as needed to one or many different user, groups, parties, or processes.

[0744] Reconstitution is not a monolithic process, nor does it happen in exactly the same way for all dispersion locations and methods. Different dispersion methods utilize different storage locations and different reassembly protocols. Specifically, the DigitalDoors patents define several methods all utilized by the Secure/View DigitalDoors applications. Directory files (the map) details source information, recovery file locations, Secured file locations, and key file locations. The directory is optional and can be rolled up into recovery files. Note that various encryption options can be applied granularly to directory, recovery, secured, and key files, such that some or all are encrypted, encrypted with a one time pad for each file or each separate recovery selection line item, or encrypted through access to a public key infrastructure.

[0745] Security is herein rendered by reduction in metadata, format, encoding, and complexity. The information is dispersed, contexts are broken, and entropy is increased; this creates security. DigitalDoors does not affect security by methods presumed to be secure because they are proprietary, concealed, or obscured. The explanation for why this works can be found in the Third Law of Thermodynamics and the mathematics are completely consistent with statistical methods. Saltzer & Schroeder's Principles of Secure Design define an economy of design, failsafe defaults, complete

mediation, open design, separation of privilege, and least privilege. The simplest mechanism to Secure and reconstitute the data streams, and simple enough so that everybody can understand the process. This explains the removal of metadata and formatting creating the simplest data stream that can be openly secures at all sensitivity levels and other dimensions of categorization.

[0746] Note that the reconstitution order proceeds from lowest sensitivity level to highest, from lowest compartment to highest, and so forth because access to higher levels is presumed to be progressively limited and the reconstitution of higher sensitivity level elements depend on the prior placement of the lower level elements. This is a functional necessity of the previously defined multithreaded extraction process. Also note implementation minimizes the amount of the recovery metadata necessary to reconstitute a Secured data stream. It is possible to define recovery information through indexing, hashing, and positional designations more complicated than positional offset but only through the creation of more metadata compromising basic security that is the fundamental goal of DigitalDoors technology, complexity and slow speed of process, and the creation of new avenues of inference, aggregation, and interaction.

[0747] Security is effected herein by the destruction of context, the rendering of inferential clues, and dispersion of essential selections thereby leaving a redacted data stream that can be delivered as is or reversibly reconstituted under progressive granular control until it fulfills processing, sharing, or knowledge exploration needs.

[0748] A document can be redacted to a secure state and the recovery image of the extractions is created in situ as a reconstitution overlay, sort of like a positive/negative photograph. Extractions can be defined for each sensitivity level, compartment, mission, group, et cetera, and aggregated together to form one or more recovery files for each or every mix of granular specification.

[0749] A document can be redacted to a secure state and the recovery image of the extractions are created separately as line items within a flat ASCII recovery files. Extractions can be defined for each sensitivity level, compartment mission, group, et cetera, and aggregated together to form one or more recovery files for each or every mix of granular specification. The recovery file details for each extraction, its category, its starting location, its ending location, its length, any replacement length for in situ substitutions (even if only one-for-one spaces), and the actual extraction. Note that variants on the format of the recovery file support overboard dispersion with separate encryption for each extraction applying unique keys for each. Details of the recovery variants are defined elsewhere. Recovery files can be stored locally or dispersed to MLS-compliant storage facilities.

[0750] A document can be redacted to different secured states (that is, versions) defined by each sensitivity level, compartment, mission, group, et cetera, with recovery data rolled up for recovery for each sensitivity level, compartment mission, group, et cetera. Rolled up recovery files contain line items up to and including sensitivity levels for each designated dispersed remote storage.

[0751] A document can be redacted to different secured states (that is, versions) defined by each sensitivity level, compartment, mission, group, et cetera, with recovery data dispersed to separate and unique storage for each line item. Note that the dispersion of line item extractions obstructs the basic intrinsic context to the Secured file while increasing the randomness (and quite literally the corresponding entropy within each package of directory, recovery, secured, and key file sets) and lack of context for each of the dispersed extractions.

[0752] The above outlined dispersion and recovery methods fulfill different functional and security requirements. The differing dispersion and recovery methods make tradeoffs in terms of reconstitution overhead, complexity, and processing with redundancy and accessibility to that offset the

risks from aggregation, inference, and interaction. Although DigitalDoors enables security with functionality rather than as a linear tradeoff between the two, DigitalDoors dispersion methods engender a different tradeoff with reconstitution reliability. It is not a costless tradeoff, but it is a necessary one if you must enable a data stream security ecosystem within an open and functional network infrastructure lacking hardened perimeters.

AA.1.16.6 Directory Files

[0753] Directory files contain map information to recovery files and represent a physical and logical displacement to secured sensitive information in recovery files. This enables additional opportunities for granular security and privacy in conjunction with information sharing and knowledge discovery. The information is dispersed, contexts are broken, and entropy is increased; this creates security. You can envision this process like that of a railroad switching yard containing many uniform boxcars with different content. By dispersing content, the context and utility of the sensitive information is broken with respect to the source and the aggregation of randomized extractions adds to the implementation of security through the 4th law of thermodynamics. When directory files do not exist for a given redacted data source, the View applet seeks recovery files directly in anticipation of rolled up recovery files in any of the supported optional formats, accessible dispersion locations, and encryption or key management controls.

AA.1.16.7 A compound data stream that can be stripped into granular items

[0754] It is all about data. Data is the content qualified by a context and concept. Everything else is just proxy. Existing solutions are complex workarounds in the face of a unifying data theory within security ecosystem. A number of conceptual problems disappear with the simplification of workflow and communications into the storage, manipulation, and delivery of data. With this emphasis shifted to a

data-centric view, we can successfully search, secure, and share with categorization, and better exploit information processing and communication workflow. The consequence of a simplified view is a simplified security environment not based on point solutions but that actually works as an ecosystem.

[0755] The complexity of information processing and communications, infrastructure, security, and data structure evolution have driven wedges into information technology workflow. As a result, divergent insular solutions and the enforced separation of process and storage exacerbate the differences rather than lead to simplicity and integration. Insular solutions require additional effort not only to resolve the problems discretely for each area but also to create temporary bridges between areas, an effort disproportionate to the requirements. It is also counterproductive when false positives and negatives require frequent user override responses or when the point solutions preclude or corrupt workflow. This impasse is most evident with air-gapped legacy production facilities and the complex integration of distributed workflow. Additionally, insular solutions and bridges sow conflict, adding complexity to disrupt the workflow within each area or among them. Although vendors are integrating routers, guards, IDS, IPS, viral scanners into suites or unifying products, this approach addresses the proxy activity rather than the data itself, the fundamental basis for and result from activity.

[0756] Furthermore, virtualization and the distribution of storage, processing, and communications to the edge are showing that workflow and application processing are potentially independent but interrelated. Code is just a metadata package. In effect, job control code, macros, applets, and entire applications are structures that can be distributed and run almost anywhere. These structures are data, or metadata, and are delivered like any traditional data set. SOA and SaaS insinuate the viability of this theory. Any differences between a traditional notion of data and the application as data is summarized by the different methods used to collect, collate, distribute, and present them, but

not at all in the data structure.

AA.1.16.8. Data Unification

[0757] Throughout this discussion, data is defined as any data stream regardless of internal format, purpose, intent, or structure. A data file is data. A document from Word is data (with imbedded metadata). An email message traversing a communications channel is data. An application stored on local disk or invoked from a web-interface is a file and is also data. Anything that can be represented within a metadata structure is data even if it is all application code or contains macros with event-driven code. This paper asserts that applications are nothing more than data asserted as a code invoked through a process (like a CPU). It is a delivery and processing workflow. Applications become nothing more that destinations; rather than delivering data to storage, the data is delivered to the workflow

[0758] This last assertion disconcerts some people who want to think that an application is a special set of instructions but overlook that it has structure and also because it often includes imbedded data is thus different entirely from a data store. The difference is purely contextual or conceptual, but the content is nevertheless the content of some data stream. Data acquisition, production, parsing, formatting, presentation, interpretation, categorization, and purpose are different between a pure data set and an application code, but this is also true between different data sets. Consider the difference between an MS Word document and an MS Access data set, but both are considered unstructured data streams.

[0759] Distinctions are completely contextual and predicated on what acts on the data sets, parts of the data sets, or how they are transformed for new and different purposes.

[0760] Is there a difference between an application and the data set? Of course there is, but differences are arbitrary based on usage and purpose. The difference exists mostly due to the historical

development paths apparent within information technology. Most people are not likely to distinguish a difference, although support developers and architects will. But even developers and architects are users too most of the time, and these arbitrary and imposed differences become less consequential. Consider: A difference between a zip file and a self-executing zip archive? A difference between a document file, an Acrobat file, and an Acrobat file with a built-in e-book reader? A difference between an application and a data set, a portable database application and the data set, and an application bound to the data set? Each can represent the same purpose, be repackaged in different ways, and be invoked in different ways -- implicitly or explicitly, or by the file type assignments in the desktop -- but desktop shortcuts mask such differences to the user.

[0761] All represent complex structures containing a traditional data core. Maps define execution entry points, module jump entry points, loaders and triggers for interspatial macros, and formats for parameter passing between functions. The map is (usually) a simple schema.

[0762] Consider that applications are specialized streams of data stored as a file. The structure includes a loader, the application, data space, buffer space, and increasingly the support for a digital signature validation of integrity. Malware insinuates itself inside the application file directly or as a jump vector to another file stored in memory or long-term storage. Although most data files preclude operative insertion of malware because it will force failure or be a visible attack, hierarchical and DOM data files can hide and execute malware through the built-in triggers. This malware lies dormant awaiting the specific activating trigger. Even when malware corruption is actually visible and obvious or violates the properties and attributes that are used to validate a data range, discovery is unlikely unless the actual data range is viewed or tested against acceptable properties and attributes.

[0763] Many workflows invoke the properties and attributes only at specific steps, for example,

as a format immediately prior to presentation or after data is presented and altered by the user or process. Alternative steps can and do violate that intent, and this potential grows as data and applications are increasingly abstracted as hierarchical metadata and multi-purpose document object models (DOM).

[0764] Hierarchical and DOM data files frequently include links, references, macros, applets, helper tools, or even extensive applications. While it is equally correct to assert that data files are becoming applications as it is to assert that applications are specialized data files, this paper demonstrates that is more useful and unifying to view everything as a data structure. In one sense, an application turned inside out is a data set with the application inclusion. Although the usage emphasis is on the application functionality rather than data content, the application is really nothing more than, and usefully visualized as, a compound inclusion within a data structure. Data complexity ranges from the simplest form to the hierarchical tree with imbedded compound inclusions, each of which can itself be a hierarchical tree with imbedded compound inclusions.

[0765] The most complex data structure we know about is DNA that includes some 4 billion combinations of UCAG amino acid types. Although DNA is clearly a data stream, it is a convoluted tree with hierarchical structures, external references, and imbedded compound inclusions. These imbedded compound inclusions are expressed as RNA chemical factories which create more complex amino acids and other precursors to life or even life itself. Aggregations and combinations of different parts of the same DNA chain or even external DNA, prions, proteins, free fragments of DNA or RNA, and other molecules with chemically-receptive sites for bonding to the UCAG pairs or the RNA receptors create new data, new triggers and events, and even new life forms. Other DigitalDoors white papers expand on the issue of DNA and RNA, MS Office metadata files, and the various document object models. The point herein is that DNA is primarily a data structure that is parsed and activated to initiate macros,

applets, and full-scale applications.

[0766] This section defines data structures from the simplest to the most complex. Data structures include the simplest representation of data, such a single item. An item is represented by a range of granularity and purpose; it can be a value, an add and, a subtract and, a multiplier, a base, a numerator, a denominator, a single fact, and a DigitalDoors granular content item. Containment of even that simplest of items is dependent on format, structure, purpose, infrastructure, and representation. The item could be the DNA UCAG or a compound UCAG-encoded molecule. More complex data formats include field-oriented data, such as a document with multiple position-oriented items. Padding, such as found in compound molecules, create the infrastructure to support these key items or bonding receptors. This document might be an e-mail message or a stripped announcement of a transaction, including a buyer, a seller, a product, a price, an extent; in other words, a line item transaction. The padding provides simple format or demarcation field control or more complex mixed-purpose usage for process functionality and also human readability. Field-oriented data formats support the transition to tagged data as seen with human-readable data files, application and system configuration files, or the MS Windows directory hive.

[0767] Tagged data transitions to row-oriented data transactions where the patterns among the records are repetitive with columns represent distinct fields.

[0768] These include sucrose, glucose, fructose and other energy sources for RNA process functionality. Although each sugar is different, the format is significantly similar. In the same way, row-oriented records provide the fodder for simple computerized processes.

[0769] Row-oriented data can include serialized or multiple complex indices (indexes).

[0770] Indices have multiple purposes both in terms of processing for assessment and

interpretation for display. Indices are special items that provide context through association. When indexes become more complicated, they form the basis for locating information or establishing references between data within a structured database, either relational or object-oriented tables. The structure for the rows, the relationships between indexes and items, and other relationships are described by an implicit or explicit schema. The schema is an internal or a separate externalized skeleton that supports the data structure. The issue of internalized or externalized schema, that is the application of structure and categorization, is purely an efficiency, reuse, and packaging issue. This is an another arbitrary complexity that is not needed for a data-centric model and the ensuing unifying security ecosystem.

[0771] The schema also defines the format and representation of items in terms of encoding, presentation, interpretation, and structure. For example, data can be encoded as binary, bits, decimal, ASCII, encryption, compression, or have meaningful utility only when used correctly. Binary large object (BLOB) fields can retain application code (source or machine code), sound, images, large documents, or other arbitrary data. The meaning for all of these BLOB fields is dependent on purpose, usage, format, structure, and intent. Consider executing a sound file or trying to hear an image.

[0772] When the schema is externalized or internalized as embedded tags and metatags within the data stream, the complexity of the data stream can increase. Hierarchical structures defined by a root node and branches form the basis for the DOM used by most modern desktop applications, such as the MS Office family. Not only does it support rich text formation, graphic layout, it also allows for unlimited complexity. The branches support items of any type, inclusions of any type, tags that define content, context and concept, establish links and references, or support compound macros, applets, filters, and code.

[0773] When the schema is referenced to an external template file, as with HTML or XML, the context is defined externally. When such schemas are self-documenting with tags, metatags, metadata, and the fractal-like self-similarity of nodes, the context is internally defined. However, associations to new schemas, (imposition of a new style sheet, for example), combinations with data from other sources, can alter the interpretation of the content, how it is contextual applied, and support new concepts not obvious or overtly seen. This is the basis not only for data mining and search, but also for DigitalDoors-based security methods. Hierarchical files with both internal and external schemas pose new problems in terms of aggregation, inference, and data-to-data interaction, and this redraws the map of data content, context, and concept. Tags, metatags, metadata, comments, links, references, and imbedded code can be combined with intent improvised to disrupt, damage, or combine into new usage in conflict with anticipated point-security solutions.

[0774] Output from applications or complex SOA workflow are usually in the form of messages, code modules, reports, or human-usable triggers. These are also data streams in one format or another, but nevertheless still a data stream. Delivery of large hierarchical data is not different from the delivery of DNA. The execution of code within the data is not different from the expression of RNA. The paradigm shift is from viewing applications as a collection of methods and events to that of a compound data file with methods and events imbedded within.

[0775] When you accept this theory as to the predominance of data, you are well into your journey of the unifying data theory. Compound data is not useful until is parsed and applied per context correctly. For example, we previously asked what would happen if a sound file were executed or an audio stream viewed as a graphic. This is a process failure or disease situation. Therefore, we need to filter and apply each item by type. The process is best described as separating content from context from

concept and dissecting compound formats into items, content into items, all of which are addressable by granular content control.

AA.1.16.9 Conclusion

[0776] The complexity of information processing and communications, infrastructure, security, and data structure evolution have driven wedges into computerized workflow. As a result, divergent insular solutions and the enforce separation of process and storage exacerbate the differences rather than lead to simplicity and integration. Insular solutions require additional effort not only to resolve the problems discretely for each area but also to create temporary bridges between areas, an effort disproportionate to the requirements.

[0777] The primary benefit of unifying data is to see data as a compound stream that can be stripped into items. First, we can find information and apply it as intended or explore for new information. Second, we can assign tags to each item for security purposes. Third, we can determine what items to share and how to distribute the items in whole or in part, and within a generic data stream, application, document, or a structural context. The consequence of a simplified data view is a simplified security environment that is not based on point solutions but that actually works as an ecosystem. This supports a unified infrastructure based on categorization that enhances search, security, and sharing, but even more than that, it provides the same consistent, simplified, and functional basis for search, security, and sharing.

[0778] Why DigitalDoors security? What you are trying to protect. Is it the infrastructure, the communication channel, the application, the engine running the application, or the data? Today, we are trying to protect all of them at the same time with macro processes generally at the perimeter, with conflicting tools, and succeeding at very little. It is becoming increasingly expensive, time consuming,

and disruptive within the current paradigm to impose security -- frequently counterproductively in terms of cost and degraded results. Instead, we need to recognize the convergence of application code and data, recognize that context and concept are the points of separation, and that security must be applied at the smallest discernable units. This explains the suitability for DigitalDoors granular content control.

AA.1.17.0 A New Information Security Paradigm: Neutralize the Loss of Perimeter

[0779] The existing shortfall in information security is indication of the amount of playing catch up now required. Information technology has always been driven by results in terms of accuracy, features, functionality, return on investment, and enabling wholly new business opportunities; security has always been the laggard afterthought. The current security metaphor pits functionality (or efficiency) in conflict to security, thus security is set at some practical level so that losses in functionality are tolerated to support security goals.

[0780] The desired balance between functionality and security changes with the perception of security threat levels. It is a lousy Hobbesian choice; a bad metaphor based on a flawed view of perimeter-based or point solution security, which is an evolutionary dead end.

[0781] Security goals usually lack formal definition, whereas products and services are tangible. Output is definable, the process is blueprinted, but where is security as an input, as an output, or as some multiplier effect on a production model. Security needs definition, perhaps as the ability to protect the resources and process that generates economic wealth within a production model.

[0782] It is all a matter of the paradigm you see. Security should not be a balancing act but rather a fundamental design parameter in information technology data and process. If you see security as elemental control, you see functionality as the enabler of security rather than a conflict. If you see

security as a perimeter control, you see security as a stumbling block for functionality. This later view is the prevailing paradigm. Security is a stepchild to information technology rather than an amalgamated component for advanced functionality. In fact, effective element-wise security enables information sharing, multi-level access control, redistribution, controlled released and recovery options that can create totally new products and entirely business markets and models. DigitalDoors is doing all that now. It is all a matter of the paradigm.

[0783] Recent high-profile failures in security, such as misuse of critical information, identity theft, and leakage of government secrets, the September 11, 2001 terrorist attacks, and the addition of new privacy directives and patriotic legislation have put the need for information security forefront. It is not that information has been unimportant and suddenly now is important; rather, there are limited methods and technologies for implementing information technologies that actually work. They work to a point and do not seem to be extensible to new situations or work beyond the borders for which they were invented.

[0784] Ad hoc efforts, band-aid policies, and application of best practices that aren't very good at all frankly reflect the really bad state of the art. Security as a step-child to functionality is a complex problem for governments, banks, brokerages, insurance companies, and many other fundamental human organizations that need to be revisited with new metaphors and a better paradigm.

[0785] Existing security technologies fall into three categories. They comprise access controls, encryption, and data or process filters. Access controls are based on usernames and passwords, or biometrics to a lesser degree. Encryption rebundles data into an impenetrable box opened only with an ungainly key management system that creates more problems than it addresses. Filters include virus protection, firewalls, and intrusion detection systems that search for known attack patterns at the

periphery or once beyond that perimeter. All these systems represent a perimeter fortress mentality encircling the systems and data. It is a primitive burglar alarm that does not preclude inside thefts, accidental or incidental damage, or prevent destruction once the perimeter has been breached.

[0786] Any little hole or minor access is sufficient to provide full exposure with sufficient time, effort, and ingenuity. For this reason, top secret government data and processes are segregated completely with air gaps from networks and external workflows. Of course, insiders like Hansen and Walker show the fallacy in the presumed effectiveness of such air-gap security methods. In fact, there is a common maxim that once the perimeter is breached, the soft underbelly is fully exposed, and the perimeter furthermore focuses the attack within to promote additional damage.

[0787] Sophisticated burglar alarms are augmented with internal sensors and content control. In fact, you see this with motion detectors, infrared heat sensors, glass-break detectors, and internal video and sound monitoring. Information security is developing ways to monitor internal processes and data access in a similar way. It actually makes sense, and is technological feasible to a point. However, cameras within the perimeter actually provide a venue for exploitation. Imagine have your built-in laptop camera showing the user, the user's location, other associates, and capturing the screen.

[0788] Advancing this burglar alarm concept further, museums understand the value of identifiable objects and alarm these objects to detect if they are moved or taken. Perimeter security has its place as a necessary first line of defense, but even for a museum, the perimeter is not invulnerable and not effective against an inside attack. Element-wise security a more sophisticated security system in that specific objects of value are designated as valuable and controlled individually. This is analogous for how data and process in information technology should advance. The information security paradigm requires the protection of individual data elements rather than a site-oriented perimeter access control.

[0789] There is unfortunately a great deal of difference between museum objects of value and data and process, and the corresponding methods of security. This paper details five major areas of logical discrimination. These are first, the logical similarity between data and process; second, the scale of data security dwarfs physical object security; third, data formats are becoming very complex; fourth, data elements combine to complicate security and privacy; and fifth, and information technology is perniciously imbedded and intrinsically networked and cross-linked within workflows.

[0790] First, recognize that there is no difference in information processing between data and process. The process is series of steps that have been codified and represented in a data format. Therefore, a process is data. Worldwide, there are legal protections for process in terms of patent, trade secret, and copyright laws, although the emphasis is given to protecting new ideas rather than implemented concepts. This represents a serious shortfall that eventually will have to be addressed by new laws. Data on the other hand has limited protections. It can be protected by trade secret or copyright, but courts have limited the protections for structured and organized data gathered and owned by a custodian. However, the European Union has led the way in protecting individuals from misuse of such custodial data. In spite of this forward legislation, we need many new laws to protect data held by custodians and preserve the rights of ownership for personal data. This is especially important when process is perceived as data.

[0791] Second, control over an object of value is very different from the control over data. Damage, theft, or the reproduction of an object of value is obvious and limited to a single instance. Damage, theft, reproduction, or redistribution of data is not so obvious, particularly when data is replicated in multiple instances or dispersed as part of workflows. In addition, data is damaged just by exposure of that data, not so art in a museum, which is displayed expressly to be exposed. Exposure of

data represents a new security threat, one of which you are undoubtedly well aware. The problem is that barrier-based security does not protect data from exposure for the previously outlined reasons and internal monitoring of process does not preclude digital reproduction, replication, or redistribution.

[0792] Furthermore, although a museum might have thousands of valuable objects, it is a discrete and controllable collection. Data on the other hand represents millions, billions, trillions, or more discrete collection of elements. This data security problem is neither discrete nor controllable in any human scale. This is a complex problem for governments, banks, brokerages, insurance companies, and many other fundamental human organizations, and the bigger the organization and its processes, the greater the risk for exposure.

[0793] Third, data is changing from simple and primitive structures into compound hierarchical objects. You might have heard about metadata, and that is but only part of risk. If you haven't heard about object models, metadata, and hierarchical object structures, get with the program and learn about it quickly. The risk from objectified data structures is enormous and growing, albeit totally hidden from the security audit; it is solely at this point an architect and programmer issue. You already use object models if you use any application such a database, a graphical editor, or any word processing program.

[0794] For example, a Microsoft Word document is 75% metadata. That means that each document contains 25% of what you have actually typed but mostly things you would not believe when you review security risks. Since the data storage structure is partially unreadable binary, metadata is hidden. Some metadata is displayed from the file property pages; this probably includes your name, your organization's name, but the stuff that is not visible is a security nightmare. The name of everyone internal to the organization that proofed that document, and anything and everything you think you deleted from the document is still there. Metadata is only a part of the problem.

[0795] Consider also the external links, the metadata processes that are part of every Word document, the document blueprints, and the potential for embedded sounds, graphics, spreadsheets, and other data from other sources. The security risk is no longer discrete and part of any human scale. This is only for a Word document. There are far more complicated objects models and complications from data linked through the Internet and complex network workflows. Every imbedded applet has the potential to be useful but also contain a Trojan Horse or destructive worm. Object models were developed for feature-rich applications and to enable accuracy, return on investment, and enabling wholly new business opportunities—remember that line?—but nonetheless are paradoxical to information security.

[0796] Fourth, we think of information security in terms of discrete units. For example, this might include a name, a social security number, or other elements that define an individual or overtly undermine security and personal privacy. We talked about the protection of discrete elements above and the failure of human scale in data protection. However, the problem is more egregious than the simple math of the number of discrete data elements. Realize that individual data elements are not only valuable individually but are even more valuable when combined with other such discrete elements. The sum of the whole is far greater than the sum of the discrete elements. For example, consider how a loan officer could correctly red-line your mortgage or make a business risk decision based on your blood type, your three-digit zip code zone, and your marital status without any other discrete personal information such as social security number, income level, or credit history.

[0797] There are at least three intersecting mathematical technologies you will hear more about in the future as part of any sophisticated consideration of data security and personal privacy. These are Set Theory, Polling, and Bayesian Inference. These mathematical technologies are not based on

elementary sciences and thus seem like magic to most rational people. However, we take their results for granted everyday in election forecasting, weather predictions, that computer chips work, material management, and transportation logistics. Society would slow and stop without these statistical sciences. The bottom line is that these technologies undermine information security and exponentially drive any discrete security control from a human scale. For example, if you have just hundreds of discrete data elements, you likely have billions of points of security exposures when information is extrapolated from the initial data elements.

[0798] Fifth, information technology has advanced from a simple one-for-one matching of data sets with computer processes into a scenario of multiple data sets interacting with multiple processes. This makes perimeter security impossible. With just 3 data sets and 3 processes, you have 27 permutations. With 42 data sets and 63 processes with a just 30% overlap, the security perimeter is represented by $1x10^{1943}$ permutations. That is a very big universe. Which perimeter—when there are so many? Data sets are frequently reused for many purposes or combined with other data sets. This is a principled result from Codd's database normalization concepts but also a fact of efficiency, necessity, integration, and advanced workflow development for accuracy, features, functions, return on investment, and enabling wholly new business opportunities.

[0799] Workflows intersect and overlap destroying any viable concept of a securable perimeter. Firewalls, intrusion detection, process monitoring for what? The perimeter is virtualized to the entire internal and external organization. It is a concept without reality. Of course, you could disassemble your networks, the Internet, distributed processing, and lock up each computer, although antithetical to what information process enables. You could add dongles for every piece of software to protect the data aspect of code and process. You could vet each and every user at each and every computer device. You could

view functionality and security as mutually-conflicted goals and choice some balance between functionality and security. That is the metaphor in use today, it does not work, and will work less well as the security perimeter is virtualized beyond any securable perimeter.

[0800] The message is that security must be built from the inside. Data elements must be categorized individually for risk valuation. The exponential risk from many elements, data types, and processes combining together must be assessed. Because the scale is beyond human scale, risk valuation, analysis, and security implementation must be applied by the information technology itself. It is not a human activity and correspondingly must be automated. The security must be built into the framework to bypass tensions with functionality, rather security must augment functionality instead of detracting from it. The paradigm suggests that data elements must be parsed and evaluated individually and within the context of the whole. The implementation thus demands a structure for measuring absolute security, for assessing risks and the requirements of each and every process, and applying a granular multi-level access control to individual data elements and sets or subsets of elements, or granular content control. This is certainly feasible. In fact, it has been implemented by DigitalDoors, Inc. Measurement is statistically-based. Likewise, the resulting security is statistical. Access control is multi-level and security is provided by an NP complete data dispersion implementation.

AA.1.18.0 Why security is ignored

[0801] The level of security provided in business processes and the production environment is rationalized to the lowest level possible to maintain its legitimacy. Although security is a desired attribute, it rarely if ever represents a constraint; security is perceived at best as a necessary evil. It has no return on investment (ROI). Attempts to create a security ROI rapidly unravel. Attempts to show how security itself or protection of assets as contribution to EPS (earnings per share) is really poor.

[0802] Attempts to ascribe an internal rate of return (IRR) for security in assessing alternative projects degenerate into a conflict as the security undermines the implementation of the selected project. Security is overlooked as functionality becomes essential. When security becomes an active constraint, it is removed from the equation. Security is usually optional.

[0803] Security impedes functionality. Even when it is understood that security is essential to reputation, legal compliance, while avoiding fines and penalties, the risk that a breach is likely to occur is perceived to be so distant and so minimally invasive that business as usual trumps security. It is easy to recognize the lapses of security after the fact, but it is difficult to maintain the importance of security on an ongoing current basis.

[0804] Security is a fundamental building block for a sustainable competitive advantage. When security is couched in terms of its importance to creating a strategy, implementing that strategy, and maintaining an advantage within a competitive environment, or enabling cooperative sharing without revealing private and technical details, security still fails as an afterthought. People blame and then rationalize that it is security undermining success, lowering production efficiency, and contributing nothing immediately to ROI.

[0805] The subtle point is that there can be no ROI, no sustainable advantage, no multiplier effects in the production process, no development of patents, and protection for trade secrets without security. So while security has no current ROI, security is the integral foundation to those goals. It might be possible to perform an ROI-based functionality now with the additional costs of security versus the effects to future anticipated functionality with and without security. However, such long-term calculation of an annualized loss expectancy based on so many unknowns is tantamount to clairvoyance resulting in professional suicide. It is extraordinarily difficult in to explain the future at risk the boardroom when

so much is instead measured in terms of current values and this year's bonuses.

[0806] In other words, security is necessary to the production model but not sufficient. Therein lays the rationalization to ignore security. Security does not create wealth. It is not an input to the production process. It is not even a multiplier to the production process, although it preserves the longevity of that multiplier. Security can be, and often is, rationalized as complication without portfolio or return. However, security is the necessary and only ingredient that preserves the trade secret portfolio sustaining the competitive advantage provided by the wealth production model of the information revolution.

[0807] Without security, it is not possible to maintain the restricted knowledge that makes wealth generation from the information revolution possible. Without security, wealth generation reverts to the commodity of an industrial production model. In the end, security might be the only mechanism for protecting the innovation that is the driving engine of the information revolution. It is nothing but a mistake to rationalize security as a conflicting constraint with the immediate needs for functionality when security might well be the only thing sustaining a competitive advantage.

AA.1.19.0 Security integration and the lack of attention to it

[0808] In real practice, security is an adjective to data or an adverb to process. It is not a noun because security is not a viable standalone product or service. It is not a deliverable. Security can be delivered, per se, but only as a component of a product or service, the production methods to create those, or as method to protect and retain production trade secrets. Security is not an object. It is not a goal. Its importance is undeserved as a production input or output. It is important, but only with reference to the maintenance of a sustainable competitive advantage. Security lacks return on investment (ROI) while even an internal rate of return (IRR) is a difficult and sketchy equation. If ROI is hard,

contribution to EPS is a long shot. Anticipated loss expectation (ALE) is calculable, but only in terms of losses to other quantifiable deliverables or resource inputs into profitable production activities.

[0809] Security does not stand by itself. Given the fear and attention given to all things security, it would seem that security was some concrete product and service, but it is not. Security is an attribute or property of other things. Security is not a data flow, a process that exists by itself, or an expression of information. Security is not a discrete achievement and never apart from another discrete process. It cannot be divorced from other definitive deliverables. Security is not a building block or tangible foundation per se for other deliverables. Security is a state of control by management over access to superfluous, superficial, or fundamental information necessary to complete a task where a deliverable product or service is at stake.

[0810] Security is not a task by itself, of itself, for itself, or a reason for self-importance. The cover of security to protect a process is at best placebo. It seems that the integration of security for effectiveness is not something added as the event unfolds but rather somehow threaded within the structure itself. But it is not. Unfortunately, security is not a structure or a deliverable of any type, even as an outsourced service. Outsourced security services exist only as an attribute or property to a viable deliverable product or service. It is reckless and financially imprudent to outsource security without tie-ins to production. Security is only symbiotic with other resources and processes, never a resource or process unto itself. Security integrates into other activities, but it is not an activity creating product or service with standalone value.

[0811] Security is an economic entity. It is a marginal scalar without fixed asset value. Security does not alter production output value. Security is not a multiplier of product or service value. Security does not make a product or service worth more. Security has no basic asset or production value. Security

does not make a better product or service. It only makes them sustainable. Security is not a production differentiator, but it is the sustainer of differentiation. Security protects the intellectual property value imbedded in the product or service or in the production methods used to produce them. Security protects an intellectual or economic monopoly as a multiplier to preserving this sustainable competitive advantage. Security is not a multiplier or marginal scalar of product or service value but a multiplier to retain long-term differential value.

[0812] Security does not have a worth unless tied into the risks, potentials, threats, or anticipatory costs associated with production results or losses. If security is applied as a scalar to a worthless production activity, the output is still a more-expensively worthless activity. When security is imposed as a necessary process divorced from a clear attribution, as is often the case with the traditional security infrastructure, it applies to nothing, scales a null value, and generates no value added. Security can only be economically integrated into a process that warrants its value as a positive and non-zero value-added enhancement. Security is a long-term performance multiplier, never a performance enhancement that is integrated into a productive workflow and expressed in the information revolution as the necessary and perhaps only element to sustain a long-term competitive advantage.

AA.1.20.0 Security as the future foundation for information

[0813] Security is not a standalone entity. It is not a standalone activity. It is not a job or employment title separable from other tasks and workflows. Security is the basis for just about everything else. It is a foundation for all other activities. Security has no ROI, limited contribution to EPS, no real IRR, no real ALE, and no real payback. The value of security is continuity and having a future. Without security, there is no security at all. Security is a systemic requirement for success, for survivability and maintaining a sustainable competitive advantage. The problem with the future of

security is that when we take it for granted we have none. When we think that security is an adjunct to our activities it fails and we have none.

[0814] Security is no longer something that is added to a process with the hope of any effectiveness. Security is not a chain and padlock that is separate from and attachable to raw materials, information, workflows, and production methods. Security is process that is unified with the production process. With anything less raw materials, information, and intellectual property can be separated from the production process with ensuing leakage. When security is separable, not only is the production process attacked the add-on security itself is also stolen for its own value.

[0815] Security is part of the blueprint, part of the just-in-time (JIT) processes for handling raw materials, a protection over critical and proprietary workflows, integrated into the outputs, and part of the resulting products and services. Security is the process that identifies resources of value and, at least in our explicitly and unique view, disperses those resources beyond intrinsic value, immediate utility, and contextual recognition with a reversible process to recover that value when and as needed under ownership or custodial assess controls. That is a very important future of security, one that transcends merely physical assets and includes the increasingly more valuable ones of knowledge and intellectual property.

AA.2.0.0Design and implementation

AA.2.1.0 Implementation of Granular data control in data streams

[0816] Security with operational functionality and information sharing is possible by designating all network flows as data streams and controlling those streams. Applications and data of any kind are uniformly identified as data streams, since both are informational equivalents. Assessing an application, even a virtualized web-based one, as somehow representing a higher risk than data, overlooks the

similarity in content and function of both, the higher-order representation of both as metadata, and their risk equivalence.

[0817] Security is possible through a process based on 1. identifying risk (with knowledge exploration) within the contained granular content of any data stream, 2. extracting the granular content at many defined sensitivity levels from the very matrix of any data stream, 3. providing access to a redacted form of the data stream in its most primitive state and format, and 4. reconstituting that data stream on-the-fly as needed up to the sensitivity levels required to match the information, functional, and operational requirements.

[0818] This is the essence of granular content control. It enables the continuity missing from traditional paradigms based on the trade-off in security versus functionality. Because DigitalDoors allows granular control over arbitrary data streams, applications, and web-enabled applications, operation continue unabated. However, they may lack specific access to the crown jewels.

[0819] The crown jewels are defined specifically in relevance to each user, environment, and/or ecosystem, whether as sensitive information, risky applications, or metadata combinations of both. Redaction with substitution and recognition of required formatting enables continued operation even with gapped data streams, applications, or web applications. Functionality continues, not hobbled by the trade-off with security, but rather subject to realistic identifiable security constraints.

[0820] The Internet and the private but parallel GIG (Global Information Grid) represent the expansion of the localized network to everywhere at anytime. Any conduit, even guarded, represents a two-way transmission channel for the proliferation of sensitive information that traverses any edge, any perimeter, any geopolitical authority, or any functional demarcation. The idea that edge protection and perimeter protection can be effective within such an ecosystem without definitive boundaries is

unsustainable. The idea that an arbitrary boundary for a community of interest (COI) can be defined within this ecosystem (the Internet, the GIG, or any attached subnetwork) to affect a perimeter security is also unsustainable.

[0821] While is certainly possible to define network edges and define a sharp boundary, this is possible only with complete isolation of distributed operations and the indiscriminate cessation of information sharing, as indeed experienced under many existing security implementations today. Thus, any reliance on some sort of interspatial DMZ is just good intention without good security. DMZs include virtualized web applications, social networking sites, Wikis, remote management agents, and all Internet or GIG access, but DMZs are not a sustainable security technology.

[0822] Data streams and applications represent the top two internetwork traffic. These are necessary to sustain operations and information sharing regardless of the security barriers in place. SAAS, SOA, Web 2.0, Email, remote terminal services, virtualization, and remote services are examples of data streams with or without additional applications and with or without additional embedded data content. Unfortunately, data streams and applications are both forms of metadata with unlimited scope for covert channels. Traditional security implementations can only allow or deny access; it is a binary decision with no leeway for intermediate allowances. Therefore, traditional security either allows a security gap while enabling operational functionality or totally denies any throughput with the attendant non-operational effect.

[0823] Instead, DigitalDoors allows a redacted data stream with reconstruction. The allowable throughput can be set at any intermediate level (between none, some and all) with optional reconstruction up to any level (including some or all). Risk is adjustable, and risk can be measured and risk can be asserted. This specific aspect of DigitalDoors granular content control provides the missing

functionality to enable compliant cross-domain security. In addition, automation of the selection of risk with the data stream allows for the unlimited assessment sensitivity levels and combinations of inference to make it possible to minimize risk with cross-domain information sharing and subsequent "uncontrollable" resharing with downstream partners. Cybercrime has shifted to attacking the plans, programs, designs of industry and national security.

[0824] DigitalDoors inventories the business practices, assess the need for business information, the risks from external threats and internal threats, the need for information in terms of the costs for acquisition, its retention, its reuse, and its potential for leakage and misuse both externally and internally. The resulting report details best production, contextual data value, information technology, and likely risk reduction practices. Security becomes a byproduct of better operations instead of a specific strategic goal or a forced mandate. Consider options deemed economically viable with competitive opportunities with a foundation of risk mitigation and control.

AA.2.2.0 Brief Functional Overview

[0825] DigitalDoors locates and selects critical information within a data flow, categorizes these selected elements, extracts them at any level of granularity, and disperses that critical information to different remote storage locations. The selected elements can be located manually (mouse selection), by dirty word lookup, or with sophisticated automatic tools to select content, meaning, and compounded risk factors. The downgraded data flow is stored in the lowest common risk factor and can be upgraded at wire speed to any authenticated MLS level. A trusted infrastructure with identification, authorization, and granular access control to granular storage represents a leap in security past traditional perimeter and containment methods.

AA.2.2.1 Brief Structural Overview

[0826] DigitalDoors was designed for networked and hosted infrastructures. The applications were designed and prototyped on MS Windows with IP networking, and widely demonstrated on that platform. We have shown automated content selection and categorization with AES encryption functioning on low-end laptops over an 11 Mbps wireless network with wireless SAN for remote storage and wireless Internet storage. RAID is supported in hardware. Because the user interface, the downgrading engine, the upgrading engine, and other functions were modularized, DigitalDoors is able to function on standalone systems, networked systems, client/server architectures, and under remote terminal services. This has been demonstrated under Solaris v8 and Solaris v9 and TSOL v8. CMW labeling at the granular element level extends file, device, and process-oriented CMW labeling provided by Trusted Solaris. As the infrastructure migrates to Sun Microsystems, trusted applications running on trusted SunRays and trusted servers with dispersed trusted storage will enable multilevel side-by-side sessions.

AA.2.2.2 Remote Storage Overview

[0827] DigitalDoors downgrades data streams to any granular level as configured. This corresponds to the CMW definition of sensitivity level, compartment level, and information level. DigitalDoors also identifies data stream elements by content categorization, source, destination, and compound risk factors. The data streams are split in various configurable ways, with each stream or even each granular element stored to a different location. Backups can be automatically archived.

[0828] Elements can be replicated and stored remotely. The different remote storage locations are mapped as logical shares. This is fully compliant with MS Windows, and Solaris and TSOL with Samba, CIFS, and/or NFS. As the migration of the DigitalDoors infrastructure migrates to Sun Microsystems, NAS, SAN, iSCSI, FibreChannel, SNIA, and/or other storage methodologies, other data

access methods will be supported in the same way. Currently, DigitalDoors supports at least 30 distinct storage locations for a single organization, with an unlimited parallel number for backup and replication.

[0829] The storage overhead for DigitalDoors adds only about 6% to the storage needs over baseline. However, organizational preference for archives, multilevel downgrades, separate sensitivity level storage, web-based or DOD distribution systems with catalogs each doubles the storage requirements. Also, the need for failsafe operations, caching, proxy services, and redundancy is a separate and traditional storage provisioning issue. Five (99.999%) and six nines (99.9999%) statistical operational performance demand an infrastructure with many storage server farms and multi-homed networking access. The primary requirement for DigitalDoors is consistent and compatible storage facilities that are uniquely addressable. The secondary requirement is that this storage be granularized such that data stream access can be controlled by identification and authorization by user, process, sensitivity level, compartment level, and information level.

[0830] DigitalDoors is a series of differential process step that can occur on any platform, ranging from a mainframe to a low-end laptop. It enables content-level security in any type of data stream on any type of platform. It is inherently scalable from single to unlimited users. Content-level security can be affected at any layer of the ISO OSI model, from the bottom physical layer (level 1) to the top application layer (level 7). In addition, DigitalDoors uniquely functions within an application layer data stream to analysis content, concept, and context for security sensitivity to protect again brute force inference with statistical methods and data mining, assemblage of information, and human intelligence methods. In other words, DigitalDoors is processing any type of data stream for potential security lapses.

[0831] Most deployments are likely to require enterprise operation support with mixed

client/server architecture. Automated OPSEC and COMSEC processing can occur as background processes on virtually any computing, Web/Internet, and networked platform. The DigitalDoors architecture was designed with run anywhere mentality, including hardware chipsets. However, in most cases, manual tagging presupposes a graphic user interface for selecting and identifying data stream content. As such, the core technologies require only visual element outlining, automated content identification, list lookup and matching, and basic data processing. Processing can be offloaded to the client or performed on the server. Dispersion is a function of a secure networking infrastructure. Most operations occur at network wire speed, this being the primary performance limitation.

[0832] We have installed and demonstrated its effectiveness on currently deployed COTS environments. This includes a wireless laptop network with a wireless SAN, a 10BaseT Intel client-server architecture, a 100BaseT and Gigabit Intel client-server architecture (Win95 through Win 2003 AS and W2K EAL-4), several MS Terminal Server environments, and integrated with Solaris v8 and v9 and Trusted Solaris v8 servers and clients. We have tried the technology on Palm IV; we assume that DigitalDoors is portable to PDAs, handheld devices, cellular phones, and other hardware devices subject to their memory, connectivity, and CPU prerequisites.

[0833] For point of reference, we have integrated and demonstrated DigitalDoors with MS Windows, as add-ins for MS Outlook and MS Exchange, and presume simple functional integration for any MS Office, database, data workflow, signal transmission, or COTS products. We have demonstrated DigitalDoors with ASCII files, MS Word documents, HTML, and XML metadata.

[0834] The client requires only a basic GUI interface (at a minimum) with NetBEUI, NetBIOS, or IP networking over any network transmission medium. The server components require no GUI and can run as a background process with NetBEUI, NetBios, or IP networking over any inter-compatible

network transmission media. Interprocess communications is enabled through client/server channels, OS messaging, and RPC. Support over LAN, WAN, Internet, VPN, NAS, SAN, with remote commercial storage services has been tested. DigitalDoors has also been successfully tested with encrypting network hardware (Intel) and native P2P encrypting protocols.

Table: The Data Declassification process

The Data Declassification process is defined by these steps:

Authorization of security levels:

Sensitivity

Compartment

Information

Categorization

Authorization of security roles:

Sensitivity

Compartment

Information

Categorization:

Identification of a data stream

Identification of the format for the data stream

Configuration of intended security levels

Sensitivity

Compartment

Information

Categorization

Categorization of data within the stream (e.g., tagging)

Selection of data within the stream for extraction

Extraction of data within the stream

Downgrading of data stream

Creation of upgrading recovery data files

Data backup

Data replication

Data dispersion

The Data Reclassification process is defined by these steps:

Authorization of security levels:

Sensitivity

Compartment

Information

Categorization

Authorization of security roles:

Sensitivity

Compartment

Information

Categorization

Access to downgraded data stream

Upgrading of data stream

[0835] Access security is established through a network ID and authorization process, such as Kerberos and enhanced with the compartmentalized trusted operating systems, such TSOL. Access security can be enabled for most networked devices.

[0836] Downgrading the data stream is manual or automatic, or assisted. Automatic OPSEC and COMSEC rules can be enforced prior or post manual selection (tagging of the sensitivity levels, etc.) The processes are asymmetric by design. Downgrading can require substantial time, but upgrading is limited only by the network and the time lags to deliver a signal cross-country through multi-hop switches and routers that are part of the commercial public infrastructure. Realize that firewall, guard, filter, and router hops frequently entail 300 μsec unidirectional delays; this being the primary performance delays.

[0837] DigitalDoors has designed to the architecture and the process flows such that steps are optimized to each side of the firewall, guard, filter, router, and storage server mesh, including even optimization at the client for the extensive high-level content, concept, and context identification and dispersion. Manual tagging can be performed on any computer that supports a GUI interface. Processing is bound to disk or network performance limitations. Tagging information can be processed locally or delivered to a server for service-side processing. The user's ability to locate and highlight sensitive material is the functional performance limitation.

[0838] The data stream can be any of type of data. This includes signals, files, data streams, interprocess messaging in a workflow or high-volume transaction processing environment, email, instant messaging, logical information, and COTS data files. Data streams can be free form or structured, such as a COBOL data set, a RDBMS database, an MS Word document, or metadata, such as SGML, HTML, or XML. The DigitalDoors philosophy is to enhance security while maintaining traditional process

workflows. Note that encrypted files and data streams require discrete and complete decryption before they can enter the transactional processing. Failure to decrypt such files abends systems and workflow because the format and content is unexpected. In addition, while encrypted XML and metadata is secure within that form, once decrypted is usually carries far more security-sensitive content than realized. Manual processing requires trivial overhead. CPU requirements are minimal for the client and server components.

[0839] However, OPSEC, COMSEC, and enhanced automatic and assisted data selection requires additional CPU resources and approximately 400MB of storage. Even laptops are suitable. Automatically or assisted tagging information can be processed locally or delivered to a server for secure processing. Processing is bound to roundtrip network performance limitations, and database lookup seeks times. As the complexity is increased to automatically search data, metadata, content, context, and concept for sensitivity within the data stream, performs slows. Performance is affected by disk seek times and the depth of the searching and DigitalDoors categorizing process. Performance ranges from 34,000 to 360,000 instructions per recognizable element in the data stream. This ranges from 56 to 783 μsec on IBM Thinkpad laptop yielding a Winstone value of 23. DigitalDoors processes a document of 140 words in 70 seconds on that laptop, but within 19 seconds on Dual Intel CPU Server with SCSI drives (Winstone value of 89). A production SQL database server provides several orders of magnitude better lookup performance than a client database on a laptop; it also scales better. Experience with Sequoia or comparable scalable and fault-tolerant architectures indicate single digit μsec response times.

[0840] Dispersion performance is a function of network channel responsiveness and network storage performance characteristics. Wireless Ethernet 11Mbits/s down to 2.2 Mbits/s is sufficiently faster than database seek times and storage, but nonetheless effective in demonstrations.

[0841] ReClassification performance occurs at wire speed subject to the performance limitations of network infrastructure and remote network storage performance. We have recovered 2 MB documents within 3201 μsec using a multiple-hop router network and 10Mbits/s Ethernet hubs to Intel desktops. This is equivalent to raw file display on the same networks and systems.

[0842] The footprint is purposefully small. Storage requirements include application storage, databases, temporary cache, and secure and dispersed network storage. Distribution is typically provided on a standard CD (640 MB) or business-card CD (120MB). User storage per user is as little as 15KB for a configuration profile. Application storage is approximately 30MB for clients, although this can be shared as a single store on a server or terminal server. Database storage requires approximately 231MB. This includes all language dictionaries, categorization elements, synonyms, and semiotic and grammatical inferences. Transactional processing, rollbacks, and security logs are optional, but on the same order of magnitude. All storage can use SSD, USB, PCMCIA, or flash RAM or networked devices. Performance is enhanced about 45 times with SSD. Flash RAM is marginally faster than mechanical hard drives due to lack of device driver optimization.

[0843] Data storage for user documents, data sets, data streams, metadata, and the like will be comparable to the size of the existing data store. Note that a design criteria is that all extra-security information, recovery files, auditing, and logs are stored in separate channels both for security reasons and format preservation. Expect additional overheads of only 6% for dispersion of recovery files. However, realize that minimum block sizes could increase that. Optimization for recovery files, which tend to be small, is possible by setting block sizes to 1KB or smaller on a special recovery file storage site.

[0844] When sources are dispersed and replicated for multiple location storage, or creation for

individual TS, S, and C archives with support for compartments, information labels, and categorization, then each replication obviously increases the storage requirements by about 100%. Proxy caching is effective when documents are often requested; however, for security purposes, caching of recovery files is insecure. It it's also ineffective unless the recovery files are bound to the same sensitivity/compartment/category user because the storage server generates and only delivers at the level dominated by the user clearance.

[0845] DigitalDoors assumes that the native OS encryption will be applied or that a certified encryption system is available within the network infrastructure. DigitalDoors will integrate with that facility. Encryption overheads are comparable to what exists now. When unclassified or downgraded documents and the recovery files are stored black and delivered black, the encryption overhead will be approximately 106% of what is now. In spite of that low overhead, this provides infinitely granular data storage and infinitely granular upgrading.

[0846] Interprocess communication and messaging for client/server is optimized to minimize traffic loads and point-to-point encryption overheads. This limited performance overhead is most evident when a terminal server environment is taxed by many users, who nonetheless see performance equivalent to that performed on the client desktop. User or process identification and authorization is interspersed throughout the DigitalDoors process—it is not a one time when the user logs in—so there is some ongoing validation chatter. This rate is granular and configurable by the security administrator on a range from seconds to hours.

[0847] There are true performance and storage benefits when downgraded documents are shortened and mechanically compressed. However, for functional reasons, a core design of DigitalDoors is that the data stream format is typically preserved intact (less extractions or substitutions) to allow for

workflow processing without abends, interrupts, or crashes. Even HTML, XML, and other metadata are typically preserved to retain the original format so that DigitalDoors is providing security without compromising functionality.

AA.2.2.3 Operating Parameters

[0848] Any CPU platform with a development platform and as little as 64KB of RAM. Any platform with support for either C, Java, or VB is suitable. DigitalDoors can run as a hardware process, an applet, an add-in, process filter, a dynamic link library, as an NT or RPC service, as a standalone client, as an in-process or out-of-process server, and as an enhancement to a firewall, IDS, IPS, guard, or operating system. DigitalDoors interacts with Kerberos, Active Directory, smart cards, biometrics, encrypting file systems, trusted operating systems, and many NAS, SAN, and other remote filing systems.

AA.2.2.4 Network Design Specifications

[0849] Because the core concept of DigitalDoors is to support collaboration, information sharing, dissemination, Internet and Web services, data mining, and data dispersion to remote storage, the client/server infrastructure requires only a core network. Integration preference is for a COTS network environment, such as Internet, Intranet, LAN, MAN, WAN, NetWare, NFS, or Microsoft networking, Network protocols can include NetBEUI, NetBIOS, and IP on a physical channel of 1Mbits/s wireless (Zenith), infrared, Bluetooth, 802.11a or 802.11b, or 802.11g, Ethernet, FDDI, iSCSI, frame relay, ISDN, DSL, T1, T3, OC3, and SONET.

AA.2.2.5 Remote Storage Infrastructure

[0850] DigitalDoors locates and selects critical information within a data flow, categorizes these selected elements, extracts them at any level of granularity, and disperses that critical information to

different remote storage locations. The selected elements can be located manually (mouse selection), by dirty word lookup, or with sophisticated automatic tools to select content, meaning, and compounded risk factors. The downgraded data flow is stored in the lowest common risk factor and can be upgraded at wire speed to any authenticated MLS level. A trusted infrastructure with identification, authorization, and granular access control to granular storage represents a leap in security past traditional perimeter and containment methods.

AA.2.2.6 Brief Structural Overview

[0851] DigitalDoors was designed for networked and hosted infrastructures. The applications were designed and prototyped on MS Windows with IP networking, and widely demonstrated on that platform. We have shown automated content selection and categorization with AES encryption functioning on low-end laptops over an 11 Mbps wireless network with wireless SAN for remote storage and wireless Internet storage. RAID is supported in hardware. Because the user interface, the downgrading engine, the upgrading engine, and other functions were modularized, DigitalDoors is able to function on standalone systems, networked systems, client/server architectures, and under remote terminal services.

[0852] This has been demonstrated under Solaris v8 and Solaris v9 and TSOL v8. CMW labeling at the granular element level extends file, device, and process-oriented CMW labeling provided by Trusted Solaris. As the infrastructure migrates to Sun Microsystems, trusted applications running on trusted SunRays and trusted servers with dispersed trusted storage will enable multilevel side-by-side sessions.

AA.2.2.7 Remote Storage Overview

[0853] DigitalDoors downgrades data streams to any granular level as configured. This

corresponds to the CMW definition of sensitivity level, compartment level, and information level. DigitalDoors also identifies data stream elements by content categorization, source, destination, and compound risk factors. The data streams are split in various configurable ways, with each stream or even each granular element stored to a different location. Backups can be automatically archived. Elements can be replicated and stored remotely. The different remote storage locations are mapped as logical shares. This is fully compliant with MS Windows, and Solaris and TSOL with Samba, CIFS, and/or NFS. As the migration of the DigitalDoors infrastructure migrates to Sun Microsystems, NAS, SAN, iSCSI, FibreChannel, SNIA, and/or other storage methodologies, other data access methods will be supported in the same way. Currently, DigitalDoors supports at least 30 distinct storage locations for a single organization, with an unlimited parallel number for backup and replication.

[0854] The storage overhead for DigitalDoors adds only about 6% to the storage needs over baseline. However, organizational preference for archives, multilevel downgrades, separate sensitivity level storage, web-based or DoD distribution systems with catalogs each doubles the storage requirements. Also, the need for failsafe operations, caching, proxy services, and redundancy is a separate and traditional storage provisioning issue. Five (99.999%) and six nines (99.9999%) statistical operational performance demand an infrastructure with many storage server farms and multi-homed networking access. The primary requirement for DigitalDoors is consistent and compatible storage facilities that are uniquely addressable. The secondary requirement is that this storage be granularized such that data stream access can be controlled by identification and authorization by user, process, sensitivity level, compartment level, and information level.

AA.2.3.0 Upfront Processing

[0855] DigitalDoors front-loads document and data stream processing. This upfront processing

is dense, comprehensive, and deliberate. Such overhead is necessary to explore multitier categorization and security requirements beyond a limited perspective, potential emergent properties, and classification bias. Reconstitution occurs at wire and disk access speeds with enhanced asymmetric performance.

[0856] There are five reasons for this design architecture listed in this document. One, the comprehensive parsing of the document or data stream and all associated metadata, links, attachments, inclusions, and structure provides a greater accuracy in the security assessment. Two, the comprehensive parsing balances content, context, and concept issues and provides a process milestone for assessing and balancing competing needs of security against sharing. Three, the front-loaded processing provides a focal point for word stemming, syntactical assessment, clustering, metasearching, index construction, data mining, data spanning, construction of topic maps, and a means to deal with the typical overabundance of information flow from a source (e.g. handling the fire hose concept of signal and data collection). Four, this architecture enables data, document, stream, process-, platform-, system-, network-, and grid-independence; the process is independent and allows anywhere, anytime, multi-platform access subject to access authentication and authorization procedures. Five, source documents tend to be processed once and thereafter distributed and read many times and this design optimizes for this.

Table: Upfront processing delivers:

Swarming intelligence to access aggregate value, risk, and granular value at the edge
MLS/MILS assessment
Determine what to protect and how much
Determine who should get indication of the information (directory or metadata)
Determine who should get information
Information aggregation and inference

Information prioritization as to what is needed now

Ability to redistribute multi-tier tagged materials without additional intervention

Support for inter-agency SBU tagging conflicts

AA.2.4.0 Proxy architecture

[0857] The DigitalDoors Identification and Authorization Server limits access to data and process by vetting users through an identification and authorization process, a traditional security by perimeter and process control.

[0858] The DigitalDoors Clearinghouse Server resolves site IDs, vets for identification and authorization, and establishes access level for global reconstitution. The server matches the site ID and provides the recovery map and reconstitution files from wherever they are stored. It returns only the reconstitution files as allowed by the lowest condition of {local, networked, or global security} status and the designated security clearance for the requesting user. The clearinghouse server can also compare security hashes or digital file signatures to validate the integrity of the request for reconstitution files to preclude attacks using random site IDs and message keys.

[0859] The proxy server caches and handles reconstitution requests by providing a narrow service window to preclude non-vetted browsing, data storage mining, and bulk access to maps and reconstitution files. Multiple servers can sustain any level of performance necessary with network wire speed the most relevant performance-limiting factor.

[0860] The DigitalDoors Dispersion Server is a vital security function because it creates a division between each DeClassified document, often a document or other information sources downgraded to the lowest common denominator for public dissemination, and the corresponding recovery map and reconstitution files. It delivers only what is authorized specifically and minimally for

each user. Dispersion security is significantly enhanced by creating a separation between the recovery map and the referenced reconstitution files and by providing specific multi-level access to a delivery stream of these vetted reconstitution files. System-level control with unique one-time pad encryption keys provides effective protection from intuitive forced re-assembly.

[0861] Although primary security is affected through the information hiding, dispersion, and camouflage, the proxy architecture with the identification and authorization server, a generic proxy server, and the dispersion server precludes human intelligence operations, hacking for information value, and sophisticated inferential analysis applying set theory or Bayesian statistics against protected organizational data. In addition, this proxy architecture enables a distributed and replicated design such that network or Internet support is always available and available anywhere with complete multi-level security.

AA.2.5.0 Reclassification- Reconstitution

[0862] ReClassification of declassified materials, in effect reversible redaction, requires new paradigms because the process is not at all considered in the security or workflow literature. Because the DigitalDoors process declassifies documents using various pattern recognition, work and phrase lookup, and artificial intelligence, and then disperses the extractions to various locations, the reclassification (reconstitution) process is granular. Granularity is a function of location, user, time, threat modes, supervisory overrides, and other configurable limitations. This reclassification process protects the recovery map, recovery extracts, and other storage locations by a process of dispersion and granular storage access.

AA.2.5.1 ReClassification

[0863] On the receiving side, DeClassified e-mail, documents, or data are received as a standard

e-mail message, documents, or data that required no special treatment, process, or interface. Size, shape, format, encoding, markup, and processing integrity are preserved. Data include formatted data sets, transactional data, or formatted XML and database tables, records, and queries, DeClassified e-mail, documents, or transactional data is the same as standard e-mail since there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration, as a secondary security access issue. However, the ReClassification process requires a downloadable user interface in the form of an e-mail client reader add-in, the use of our separate e-mail reader, or an integrated document or data reader.Selection of a DeClassified e-mail message, document, or data transaction initiates a workflow process consisting of:

Table: Workflow process

Identification and Authorization (either local or remote)

Lookup of recovery map information

Access control by time, location, threat mode, username, category, time of day, level hierarchy

Access to the recovery map (granular access controlled at all time)

Access control to dispersed storage locations

Recovery of dispersed classified critical and important information

Recovery of the encryption key per organization, per user, per document, and/or per extraction

Decryption of encrypted or MultiCrypted critical and important information (on the user side at the user machine through proprietary, standard, SSL, or other encryption add-ins)

Reconsitition (ReClassification of DeClassified e-mail message)

Display within standard user interface of ReClassified e-mail message

(such as within e-mail client, PDA, etc.)

Display within separate e-mail reader/ReClassifier application

Display within separate reader/ReClassifier application

Display within separate user application (MS Word, database, PowerPoint, Quicken)

[0864] Controlled release for ReClassification purposes can be limited such that different users will have leveled access to the extractions, for example, senior users can have access to the names and credit numbers on a granular basis, whereas junior users cannot.

[0865] Workflow and transactional messaging activities that require e-mail, POP3, SMTP, IP, or other similar message flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways. Pagers, wireless messaging, and other token delivery methods can be incorporated into access control, authentication, and differentiation for ReClassification levels. The DeClassification process and the display of the ReClassified e-mail, documents, or transaction data does not have to occur in any set or preset location. ReClassification can occur at network devices (in whole or in part) and be delivered to that same location or to other locations. Accumulation of recovery information need not occur at any one single location.

AA.2.5.2 Recovery Information

[0866] The recovery information is not a necessarily a single entity or file, or accumulated in one place or by a sequential single process at any single time. The recovery information is a starting point that describes the location of the recovery map. The recovery information for each declassified document is independent from all other declassified documents. Access to that information is either known or, not known but delivered after secured authentication. That location can be communicated in various

methods, including system message, e-mail, by phone, encrypted as a message or semaphore, attachments, or other means. Access keys, encryption keys, and rights to that location are also communicated in various like methods.

[0867] The location might contain the entirety of the recovery map as a data record, table, XML document, or file in plain text, encrypted, or second order pointers to other locations. Support for other locations is arbitrarily infinite. The recovery map might optionally contain information about the classified source document and the declassification and extraction process. Key sharing for storage access, authentication, and recovery is enabled with {L:N of M} key-sharing schemes. Typically, the recovery map defines offsets to extracted information (offset from initial starting point of the document, content structure, or data stream, offset from last extract, or from arbitrary tags and markers), lengths of the extracted information, lengths of any replacement for the extraction, extraction type, and the extraction itself. The offset is functional in all known types of media formats including text, formatted text, XML, audio, image, video, multimedia, database, and object-oriented descriptions. The starting point from content structure could include the offset from a natural break in the data stream format, a format block, or change in content type or structure.

[0868] For example, an MS Word document contains a metadata structure within which the user content is embedded; the reference to the user content forms the natural offset starting point. This is analogous for other types of documents, such a spreadsheet, data query, or OLE document embedded with the MS Word document. Two-pass placeholders support reconstitution in the event a DeClassified document is altered (such as edited, attacked, or damaged) subsequent to the DeClassification. This enables ongoing data processing with secured data and protects DeClassified data even in the event of a concerted semantic attack.

[0869] All of this information can be in plain text, encrypted, or represented by pointers to other storage locations. Numbered tag or two-pass tags (such as barcodes or URLs) allow ongoing editing, as with MS Word, wherein placement and even the existence of extractions could be altered in processes subsequent to DigitalDoors DeClassification, thus allowing alteration and functional ReClassification of the DeClassified document at any future time.

[0870] Support for other locations is arbitrarily infinite. Each unitary recovery unit for each extraction can be dispersed to an arbitrary storage location. Each storage location can be managed as a closed storage community, with access authenticated at a granular level. Storage locations in whole or in part and individual extractions can be distributed to enable survivability. Failure or loss of any single recovery information or extracted element does not preclude partial ReClassification or failure of any other ReClassification for others documents. A sufficient dispersed survival scheme provides 100% recovery in the event of catastrophic failure.

[0871] The location of each secondary or subsequent storage location is independent from all other storage locations. Access to one a particular storage location, file path, file name, access key, encryption key, or other access control (such as biometric token) in no way provides information about or access to any other location. Access to each secondary or subsequent storage location is also granular and separated authenticated. As a result, access to any part of the initial recovery information, even in plain text, without access to storage locations and access authentication prevents full or even partial recovery of the classified document. Plain-text access to any and all storage locations alone precludes manual recovery of a declassified document using human intelligence, as the magnitude of the accurate placement and reference is astronomical (the problem is N!, non-polynomial, or NP complete). Access to a complete plain-text recovery map with plain-text access to any and all storage locations could

compromise a single declassified in full or in part, but no other declassified documents.

[0872] Authentication is granularly revocable at any time now or in the future. Access to classified documents and the reclassified levels is controllable at any time in the future. Access to recovery information, recovery maps, individual extractions, locations, and access to storage is both granular and revocable.

AA.2.5.3 Processing with DeClassified Data

[0873] The starting point from content structure could include the offset from a natural break in the data stream format, a format block, or change in content type or structure. For example, an MS Word document contains a metadata structure within which the user content is embedded; the reference to the user content forms the natural offset starting point. This is analogous for other types of documents, such a spreadsheet, data query, or OLE document embedded with the MS Word document. Two-pass placeholders support reconstitution in the event a DeClassified document is altered (such as edited, attacked, or damaged) subsequent to the DeClassification. This enables ongoing data processing with secured data and protects DeClassified data even in the event of a concerted semantic attack.

AA.2.5.4 Content Filtering

[0874] Content filtering includes declassification downgrading but also more typical filtering, document retention, and e-mail attachment retention policies, as implemented by other e-mail products. The major difference between typical filtering technology and DigitalDoors, is the ability of the e-mail, document, or data recipient to potential review the e-mail message and all attachments reconstituted in full, not just as filtered and delivered. The core engine supports classification by word type. Forty categories are currently supported. We are adding politically correct, hate mongering, job resumes, pornography, drug culture, games and jokes, spam, and others as specified. DigitalDoors can filter

outgoing and forwarded e-mail for .com, .exe, sys, .ovl, .dll, .ocx, .prg, scr, and script attachments and even search within these data streams for known file infection patterns. Message subject, content, and attachments are processed. In addition to standard content filtering, DigitalDoors also assesses target e-mail addresses including to, cc, and bcc. The workflow is basically one of reviewing inbound and outbound messages for content. When excessive amounts of e-mail messages, documents, or transactional data cannot be processed in real-time, they are parked for later processing in a temporary folder. Administrator (not user) configuration defines whether copies are made, saved, or deleted. Reports explain actions taken and the event logging databases provide analysis into the e-mail declassification and filtering process.

AA.2.5.5 Channel Transmission Protection

[0875] The transmission channel(s) for source data to the user's desktop, server, or transaction process, external generation of the recovery file, and external generation of a recovery map are all authenticated with multi-factor authentication, such as PKI certificates, tokens, or other methods, and encrypted with point-to-point encryption of the transmission channel with SSL or in hardware. This process can be controlled by DigitalDoors or be part of the core networking environment and independent from DigitalDoors functionality.

AA.2.6.0 The Recovery map

[0876] DigitalDoors enables information security and privacy protection through the selection of critical elements, the granular extraction of those critical elements to dispersed storage, the distribution of the redacted (and encrypted) information while preserving functional work flow format. The system enables reconstruction of the original document and the multi-level reconstitution of the document to users with different clearance levels and different rolls based on their identification and

access authorization. Selection of critical elements can be done: automatically, fully-manual, or assisted with automatic features and manual overrides. Reconstitution and multi-level reconstitution is driven by a "recovery map" that details the selections, replacements, and pointers to actual extractions that are restored with appropriate authorization.

AA.2.6.1 Map Structure Variants and Granularity

[0877] The DigitalDoors "recovery map" can be in different forms depending on application, configuration, encryption, signatures, and administrative options. The map in effect is a starting path to a file (database record, script-macro, or data stream) represented by an unqualified or fully-qualified path name and file name, resource reference, a URL, or an XML pointer. The path might not even be recorded, or it might be encrypted. This initial file might contain no references whatsoever to the source and/or redacted file, or even overt references to the collections of associated files and processes. Each file in the collections of associated files and processes might be encrypted in whole or in part, encrypted extraction-wise, element-wise, or distributed as additional collections of associated files and processes, dispersed to an arbitrary number of networked storage locations.

[0878] The "recovery map" is partitioned to improved reliability, replicability, survivability, integrity, validity, security, and preclude forced inferential reconstitution activities. The "recovery map" exists as a series of steps and separated pieces of information, with minimal top-down dependence and also stepwise and separated independence. Each step and dispersion further precludes reverse engineering the source and decreases the amount of contextual value to an extraction recovery file. There is no down-top dependence. Each step, each price of information, and each file is independent from all others with no or limited reference to the original source, redacted document, or extraction files. File names, file locations, file date and time stamps, file sizes, and file ownership ring rights might be altered

to preclude source document group association, user relationships, process or procedural connections, event correlation, or indication or sources and methods.

[0879] Options for the redaction and how extractions are represented within the redacted document are important in achieving a high-level security while maintaining interoperability. As such, replacements for extractions can be null, randomized in both content, context, and size, or substituted with misleading information. Because the various recovery maps can also include efficient minimal or fully chaffed listings of all document elements, combinations of redacted documents with even plain-text recovery maps can severely hamper forced inferential reconstitution activities.

[0880] The DigitalDoors recovery map information in its simplest form contains the offsets to the extractions, length of extractions, length of redacted document replacements (which can range from 0 to thousands of characters). This file, database record, script-macro, or data stream might be encrypted in whole or in parts, encrypted extraction-wise, element-wise, or distributed whole and in parts. Encryption is through various algorithms including 3DES, Rijndael, or any third-party encryption product. Granularity with randomness, chaos in association, and dispersion might make the need for encryption necessary merely to meet legal requirements. The recovery map is not necessarily a single entity, but rather usually a collection of associated files and processes, dispersed to an arbitrary number of networked storage locations each accessible through hierarchical controls. The redacted file must have an association to at least one recovery map under any expectation of full or partial reconstitution. Existence of the recovery map alone does not provide any ability to access, decrypt, or apply the map for increased comprehension of the redacted document through forced inferential reconstitution activities. Access to the recovery map is controlled by identification and authorization, by 1-factor, 2-factor, 3-factor authentication or third-party products integrated or callable within the DigitalDoors

process.

[0881] Metaphorically, you might think of the recovery map as a pirate treasure map or such a map broken in small pieces so that only a sufficient number of pieces have practical value. Furthermore, without ownership, map location reference (the starting point), a time relevancy element, a time frame in history, an understanding of the map relevance to a treasure, the age of the map, or the authenticity and integrity of the map, you have only an interesting artifact. Sometimes it is possible to infer one of those elements and thereby establish values for other elements and thus possible context to the map. Consider that you have thousands of such maps,you have only the context of enforced secrecy and no singular context for any one map, let alone many of these maps. A mapped mismatched with an anticipated treasure is not likely to provide the right results, even for pirate treasure. Given a correct context and a correct starting point for the map, you still need direction and scale for perspective. Now consider that the map provides location for a treasure, you will still need conveyance to the treasure (network access), ability to get the treasure itself (access control), and a means to break open the box containing the treasure (decryption key).

[0882] The availability of new information analysis techniques requires that security experts counter threats not only from the obvious but also the probable, possible, and unknown. Information analysis includes not only capture of native communication streams, but also human intelligence methods, assemblage, data mining, intuition, interception and alteration, hypothesis construction and validation, logical inference, Bayesian inference, polling, and set theory. These methods represent a dramatic advancement through the statistical quantification of threats and unknowns with decision-making based on game theory and risk management. As a result, information security methods must include protection against the obvious and also the probable and possible. New information security

methods must be flexible enough to protect again the accidental, intentional with systems knowledge, unforeseen, and unknown as well. Concrete security methods do not anticipate probable, possible, and wholly unknown and unforeseen attacks.

[0883] The range of known to unknown, expected, probable, possible, promising, and the only-likely predicate the need for new security methods. As a result, DigitalDoors feels the best defense is based on statistical methods, chaos theory, Bayesian inference, and set theory. DigitalDoors is aware of these attacking technologies and inverts these same techniques to counter these actual and perceived threats in both defensive and offensive processes.

[0884] The DigitalDoors technology includes the recognition of information as content, context, and concept. Content is the obvious textual or streaming information. Context is the broader information gleaned from work flows, situational utility, how the content is gathered, created, processed, used, and distributed, and what other content establishes an informational basis for semiotic connotation and denotation. Concept is created through process structure, data structure, data tagging and encoding, metadata, and reference and inference to the textual information environment. The ultimate situation is a large number of independent extraction recovery files, generally inaccessible by users, further individually encrypted with one-time pads, which even in aggregate indicate only activity but not content, context, or concept. This explains the value for selection, extraction, granularity, and dispersion.

[0885] A relationship exists between the source document, the recovery map, reconstitition extraction files, process flows, encryption/decryption steps, integrity checksums, digital signatures, key generation, key distribution, and anonymity. Note that separate and independent channels can mask any relationship among the source, its recovery map(s), and its extraction file(s). Independence, randomness, dispersion, secure erasure (per MilSpec), optional date and time stamping erasure, initial creation of

random files to preclude sequential dated backups analysis for creating grouped file associations.

[0886] DigitalDoors protects the redacted documents from the recovery maps, and these from the extraction files by implementation of chaos (such as, the law of large numbers), the independence of the files, file names, location of files, access to files, and the infinite granularity of individual recovery maps and extraction files. In other words, a recovery map can be processed through the DigitalDoors technology n-times to create n-levels of indirection. In the same way, extracted components can be dispersed n-times to create n-levels of access control and indirection.

[0887] DigitalDoors technology is implemented under MS Windows. Client, server, and remote components exist, support networking, dispersion, encryption, and add-in integration with other Windows-compliant applications. The technology is generally partitioned into user interface, processing, and storage modules, thereby foreseeing deployment as remote, host-based, server-based, and appliance-based code. For example, the aspects within MS word that create insecurity, links to external sites and other documents, imbedded digital content in formats not known to DigitalDoors, redlining, metadata, and hidden elements would be automatically excluded, the remainder would be redacted and published, and the source Word file itself could be purged from the system or stored within a secure footprint. Since some of the necessary processes compromise the nominal security of the desktop and the communication channels, different parts of this process would occur within different vetted environments; the only insecurity would include what is actually visible on the screen at any time and/or saved as PrntScrn, and the parts of the document buffered in RAM or disk swap space.

[0888] DigitalDoors technology can be deployed to secure environments within an open and insecure network infrastructure (IP over Ethernet and Cisco routers) on notoriously insecure components (laptops, PCs, handheld devices, and general-purpose appliances) using COTS applications (MS Word,

MS Powerpoint, and new file sharing features embedded in Windows 2003). Application of digital signatures, key management with tokens, proxies, and special key servers provides the necessary infrastructure to vet the integrity of a process work flow.

AA.2.7.0 Concurrency in granular data control

[0889] Because the code base for the DigitalDoors applications is based on object-oriented programming, many processes, functions, features, events, and user options are inherently concurrent. However, many processes, functions, and features include an active process status test to disable certain aspects of this concurrency.

[0890] This prevents process clashes (for example when two or classifications processes run at the same time on different multiprocessor threads) but primary enhances consistency and security to contain risk. Concurrency and parallelism represent new doorways through the perimeter, an obvious failure in any perimeter security model. Reduction and control over concurrency enables the new deperimeterized DigitalDoors security model by controlling what and how much goes through the doors. This is granular content control.

[0891] For example, it is possible to start the Secure process on a document and leave the computer; the process will complete, but no manual activity or options are allowed once the timed login has expired. Likewise queued documents will be processed safety in the background with an active user login, but the process is locked into whatever administrator and user options are in effect at the timed login expires.

[0892] Likewise, many categorization functions will preclude manual overrides until they are completed. This prevents clashes and security lapses should a user try to distort the tagging. Generally, the automatic recoding feature will code to the highest sensitivity level even when prior processes assert

some tag lower than the highest level in the various databases. This helps achieve stability and consistency as well. Manual overrides against the document itself or the relined selections can always occur at any stage, but not during an automated process. However, note that COMSEC will preclude any manual overrides, as it is meant to do, so as to assert that demanded highest organizational sensitivity level.

AA.2.8.0 DeClassifying and ReClassifying E-Mail

[0893] Security for workflows, transactions, and E-mail requires new paradigms because it is only partially addressed by current encryption methods. The current technology fails to address workflow, convenience, and distribution issues. Specifically, before and after encryption, the security gap reveals a window of opportunity to attack the plain-text. DigitalDoors specifically addresses the security of workflow, transmission, and e-mail with methods to distribute and provide granular leveled access to information. The two control points for securing e-mail with DeClassification are represented by the generation and distribution of the e-mail and the receiving and ReClassification (reconstitution of the critical and important information) of the e-mail.

AA.2.8.1 Generation and Distribution

[0894] There are several issues when securing e-mail. First, is the issue of sourcing e-mail. Although this generally occurs at the e-mail server or the e-mail client, frequently other applications directly initiate the e-mail transmission from within the applications themselves. Internet enablement and interprocess messaging represent serious growth areas in application development, for which DigitalDoors is expressly useful. We implement DeClassification e-mail generation from within applications directly by initiating the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Second, we intercept the generation of e-mail and DeClassify it on the

the fly by initiating the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Attachments can be unpacked and examined, declassified, and reprocessed as attachments. Third, we intercept outboxes and outgoing e-mail caches in order to initiate the DeClassification process with multiple secured outputs where the declassified e-mail is an output. Classified information is encrypted, MultiCrypted, and dispersed to local and/or remote storage locations. Storage of critical and important information is configured and dispersed to different storage locations including Internet and hosted storage options. Dispersal is controlled by policy, configuration options, and threat mode based on users, recipients, situational needs, e-mail content, and critical and important information categorization. For example, an e-mail containing names and credit card numbers can be DeClassified with each of the two (and other categories) dispersed to separate storage locations.

[0895] Distribution represents no changes to standard e-mail process in any form as standard address book processes are used. We can match the address book, names, or e-mail addresses against known threats or alter processing subject to the target addresses. Remailing could initiate a further downgrading (or even upgrading) of content when compared against target addresses. Distribution is nominally assumed to exist within a decentralized, distributed, and non-secured environment. There is no special need, effort, or consideration for distributing DeClassified e-mail as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes.

[0896] ReClassification is a function that is controlled and limited by the standard identification and authorization process. This includes the use of agents, and leveled access through secret sharing, alerts, threats, location, timing, policy, access to remote storage, and dispersive storage.

AA.2.8.2 Receiving and ReClassification

[0897] On the receiving side, DeClassified e-mail is received as a standard e-mail message that

required no special treatment, process, or interface. DeClassified e-mail is the same as standard e-mail as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration.

[0898] The ReClassification process requires a downloadable user interface in the form of a e-mail client reader add-in or the use of our separate e-mail reader. Selection of a DeClassified e-mail message initiates a workflow process.

Table: DeClassified e-mail message workflow process

Identification and Authorization (either local or remote)
Lookup of recovery map information
Access control by time, location, threat mode, username, category, time of day, level hierarchy
Access to the recovery map (granular access controlled at all time)
Access control to dispersed storage locations
Recovery of dispersed classified critical and important information
Decryption of encrypted or MultiCrypted critical and important information
Reconstitution (ReClassification of DeClassified e-mail message)
Display within standard user interface of ReClassified e-mail message
(Such as within e-mail client, PDA, etc.)
Display within separate e-mail reader/ReClassifier application

[0899] Controlled release for ReClassification purposes can be limited such that different e-mail

users will have leveled access to the e-mail message, for example, senior users can have access to the names and credit numbers on a granular basis whereas junior users cannot.

[0900] Workflow and transactional messaging activities that require e-mail, POP3, SMTP, IP, or other similar message flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways.

AA.2.8.3 Forwarding and Redelivery of E-Mail

[0901] The standard ReClassification process protects access to DeClassified e-mail while it is ReClassified. This retains control over information flow. Redelivery, forwarding, or remailing of the e-mail message prevents general exposure to critical and important elements, as only the DeClassified message can be forwarded or remailed. All subsequent readers of the e-mail will see only this DeClassified version as a minimum in security. However, access to the partial leveled or fully ReClassified document will require the standard access identification and authorization process.

It is also possible to embed within the browser tools--assuming control over the e-mail reading environment within an organization--indicators to further DeClassify e-mail prior to forwarding or redelivery of previously downgraded e-mail messages.

AA.2.8.4 Content Filtering

[0902] Content filtering includes declassification downgrading but also more typical filtering, document retention, and e-mail attachment retention policies, as implemented by other e-mail products. The major difference between typical filtering technology and DigitalDoors, is the ability of the mail recipient to potential review the e-mail message and all attachments reconstituted in full, not just as filtered and delivered. The core engine supports classification by word type. Forty categories are currently supported. We are adding politically correct, hate mongering, job resumes, pornography, drug

culture, games and jokes, spam, and others as specified. DigitalDoors can filter outgoing and forwarded e-mail for .com, .exe, sys, .ovl, .dll, .ocx, .prg, and script attachments and even search within these data streams for known file infection patterns. Message subject, content, and attachments are processed. In addition to standard content filtering, DigitalDoors also assesses target e-mail addresses including to, cc, and bcc. The workflow is basically one of reviewing inbound and outbound messages for content. When excessive amounts of messages cannot be processed in real-time, they are parked for later processing in a temporary folder. Administrator (not user) configuration defines whether copies are made, saved, or deleted. Reports explain actions taken and the event logging databases provide analysis into the e-mail declassification and filtering process.

AA.2.8.5 Processing E-Mail

[0903] DigitalDoors was envisioned as a solution for security in complex workflows as opposed to point solutions for files, records, or databases. As such, the modularized structure of the suite, the support data structures, and databases, and the flexibility for semaphores, interprocess communication, and command line arguments have allowed us to integrate the DigitalDoors suite into Microsoft Outlook, one of the most proprietary and awkward software development platforms imaginable. As a result of this integration effort, DigitalDoors for Outlook runs virtually transparently as part of the normal Outlook workflow. In fact, it is a structured as a standard .DLL add-in library.

[0904] New messages are DeClassified prior to outgoing delivery and received messages are automatically ReClassified for users according to security levels. Recovery information widely dispersed and distributed on VPNs or Internet-attached networks are easily mapped into the received e-mail for real-time display. Reconstitution times compare to wire speed and are several magnitudes faster than traditional encryption technologies. Revocation of access, revocation of messages sent in error, and

location-specific security is enforced as with all DigitalDoors technologies.

[0905] As with most workflow integration efforts, maintenance of security and smoothness of operations dwarfed programming issues. System timing, order of process, access control, workability, and other reliability overshadowed code concerns. Interception of events inside the Outlook framework proved the most fearsome constraints; these problems have since been overcome. However, it is important to document the workflow for security reasons.

[0906] All DigitalDoors process within Outlook are transparent where possible, all standard outlook functions, features, and workflows are maintained intact, and we succeeded at not altering the user interface with the exception of several toolbar buttons for online help and specific DigitalDoors functions. The creation of a new or forwarded e-mail message occurs without change in the user interface. An additional DigitalDoors user logon is required for all DeClassification and ReClassification consistent with the DigitalDoors desktop Suite. Administrative and user controls are configured through the DigitalDoors Suite. Outlook username names, MS Exchange interfaces, e-mail addresses, passwords, and other Outlook-specific settings are maintained and function identically with or without the DigitalDoors for Outlook.DLL add-in.

[0907] While the external workflow is not altered from the user's perspective, several internal changes are important to document. First, e-mail messages are constructed through four primary mechanisms. First, they are created as new. Second, they are forwarded or responded to. Third, messages are constructed as drafts. Fourth messages are copied and pasted. Other workflows are possible with MS Exchange, other Microsoft workgroup messaging tools, and third-party .DLL add-ins. User-created or third-party .OPM or VB scripts should function normally with the DigitalDoors for Outlook.DLL installed, however, we frown on such tools as significant security holes and for the potential conflicts

with other add-ins, overlapping scripts, and potential to create holes for viruses and opportunities for script kiddies. Our initial scripts were converted and compiled into the more-secure MS-preferred .DLL add-in formats. Note that VBA, VB scripts, and other tools can undermine Outlook security, the Outlook security patches that prevent scripts from running, and create many security concerns. We suggest you avoid that path. Nevertheless, the DigitalDoors process remains the same throughout whether other scripts are installed.

[0908] New message events are trapped. The outgoing Outbox message is moved into a cache folder for processing. DeClassification processing occurs per administrative configuration. The message is parsed completely by means of the MS Outlook message object structure. A temporary input file is constructed. This file is passed through a system shell function to the DeClassify application. The temporary file is processed by this application and a temporary output DeClassified file is created. The recovery map and files are also created and dispersed per administrative configuration. The DeClassify application deletes all traces of the temporary input file as per administrative configuration. This includes overwrites per SAS, Orange Book, and other security standards for file erasure. The DeClassify application completes. The temporary output DeClassified file replaces the contents of the original outlook message. The message is returned to the Outlook Outbox folder and delivered per standard Outlook configuration.

[0909] ReClassification occurs when a user selects an Inbox message. If the message has been DeClassify, the selection event is intercepted and the temporary input DeClassified file is created. This file is passed through a system shell function to the ReClassify application. The temporary file is processed by this application and a temporary ReClassified output is displayed. All user-specific DigitalDoors security measures are applied to the level of granular recovery, access controls, location,

and other limitations. The DeClassify application deletes all traces of the temporary input file as per administrative configuration. This includes overwrites per SAS, Orange Book, and other security standards for file erasure. The ReClassify application completes.

AA.2.9.0 Declassifying and ReClassifying while Browsing the web

[0910] Security for workflows, transactions, and information stores while browsing the Internet and Intranet webs require new paradigms because it is only partially addressed by current methods. The current technology fails to address workflow, convenience, and distribution issues for public and private information, the granularity of delivery for that information, or the centralized and decentralized access to that information. DigitalDoors specifically addresses the security of workflow, transmission, and browsing with methods to distribute and provide granular leveled access to information. The two control points for securing browsing with DeClassification are represented by the de facto delivery of declassified information with access to immediately reclassified information subject to secondary identification and authorization.

AA.2.9.1Generation and Distribution:

[0911] There are several issues when securing browsing and access to data (web pages, tables, data sets, and linked files). First, is the issue of sourcing information and downgrading it for general presentation. DigitalDoors implements DeClassification for all data stored directly by initiating the DeClassification process with multiple secured outputs where the declassified data is an output. Second, access to the ReClassified material is generated on-the-fly subject to identification and authorization. This is accomplished by initiating the DeClassification process with multiple secured outputs where the declassified data is an input along with the critical and important extracts.

[0912] Classified information is encrypted, MultiCrypted, and dispersed to local and/or remote

storage locations. Storage of critical and important information is configured and dispersed to different storage locations including Internet and hosted storage options. Dispersal is controlled by policy, configuration options, and threat mode based on users, recipients, situational needs, browsing content, and critical and important information categorization. For example, a static web page or even one generated dynamically from a database containing names and credit card numbers can be DeClassified with each of the two (and other categories) dispersed to separate storage locations.

[0913] Distribution represents no changes to standard browsing process in any form except for dynamically-generated pages require a dynamic recovery file. Distribution is nominally assumed to exist within a decentralized, distributed, and non-secured environment. There is no special need, effort, or consideration for distributing DeClassified browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. Because all static pages are stored in the DeClassified format, access to all pages in a ReClassified format is on a per page basis.

[0914] ReClassification is a function that is controlled and limited by the standard identification and authorization process. This includes the use of agents, and leveled access through secret sharing, alerts, threats, location, timing, policy, access to remote storage, and dispersive storage.

AA.2.9.2 Receiving and ReClassification

[0915] On the receiving side, DeClassified browsing is received as a standard data that required no special treatment, process, or interface. DeClassified browsing is the same as standard browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration.

[0916] The ReClassification process requires a downloadable user interface in the form of a browsing client reader add-in or the use of our separate browsing reader. Selection of a DeClassified data initiates a workflow process consisting of:

Identification and Authorization (either local or remote)

Lookup of recovery map information

Access control by time, location, threat mode, username, category, time of day,

level hierarchy

Access to the recovery map (granular access controlled at all time)

Access control to dispersed storage locations

Recovery of dispersed, Classified critical and important information

Decryption of encrypted or MultiCrypted critical and important information

Reconsitition (ReClassification of DeClassified data)

Display within standard user interface of ReClassified data

(such as within browsing client, PDA, etc.)

Display within separate browsing reader/ReClassifier application

[0917] Controlled release for ReClassification purposes can be limited such that different browsing users will have leveled access to the data, for example, senior users can have access to the names and credit numbers on a granular basis whereas junior users cannot.

[0918] Workflow and transactional messaging activities that require dynamic browsing with ad hoc generated reports or other similar flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways, with generation of a temporary recovery file.

AA.2.10.0 Creating multiple levels of categorization

[0919] DigitalDoors resolves conflicts in multitier overlapping hierarchical categorization to the highest, most risk adverse level, in order to create generic redaction. Granular content control with reconstitution provides future-proof "put the cat back-in-the-bag control" over repurposing while enabling preservation of ongoing critical missions.

[0920] Categorization is reflected by the end results of tagging (also known as labeling). You interpret the information and traditionally select one available category and attach the associated tag to the range over which the categorization applies. The problem is you attach one tag to a range, often a paragraph of many sentences. This one tag has to cover all possible needs because you cannot have two or more tags, and you cannot assign a separate tag to each sentence within the paragraph. In some instances, this categorization is limited to a sensitivity level, or at least what you perceive the sensitivity to be given your analysis of the circumstances, á priori knowledge, and anticipated usage and risk factors. Risk factors are often the result of technical issues of sources and methods.

[0921] Even were the information is unimportant, is reflected in the daily public news, or is general knowledge, the source or method from which this fragment of information was obtained in this specific instance, defines a specific sensitivity level. It results from organizational agenda, human limitations in processing a fire hose of information input, and legitimate perspectives. Consider the redaction of the Valerie Plame Wilson manuscript, "Fair Game," where the published addendum restored the missing information, both in terms of simple content and more complex context.

[0922] Such one-dimensional categorization defines the many problems of categorization. The entire knowledge environment is reduced to a lowest common denominator of utility or expectation. It is indicative of the most pressing needs or the most immediate interpretation. In effect, the tag is saying

that the range so defined is intended for this purpose, contains this level of information, and overlooks any possible reuse, repurposing, or the potential for inference. When there is obvious potential for inference risk, the tag is assigned that reflects that immediate understanding of this potential increased risk, frequently pushing the information into a black hole where it never gets released. It also never gets reassessed because of workflow, work overloads, and the increasing information input.

[0923] Categorization is really about creating multiple levels. These include intended purpose, possible purposes, a time line of value, and assessment of potential inference. Categorization should allow overlapping because people write sentences that are disjoint and have different ideas, reflect competing ideas, match the well-known with new fact or allegation, and create a hodge-podge of conflict. In fact, categorization should support the arbitrary selection of words, phrases, parts of pictures, parts of an audio track, and individual symbols. Categorization should be granular. That categorization becomes hierarchical and tangled. One of the reasons such categorization is rare is that because it is messy, does not lend itself to simple solutions, and does not simplify. It adds complexity. It becomes metadata, an additional in-channel and out-of-channel risk to data streams, adds knew knowledge through aggregation, and opens the doors to inference.

[0924] However, hierarchical overlapping classification enables a number of information processing solutions that otherwise are too complex to consider. Specifically, we refer to the measurement of information, the security of information, the discovery of new knowledge, and the distribution of it all.

[0925] The time component to information is frequently ignored. Most people think about information decaying and becoming less important over time because we do not need it anymore or newer information supplants it and makes it useless. However, the information stores are frequently

updated, augmented, and linked with other information to reflect a newer understanding. This might be newer information or just older information seen in a new light. The aggregation, inference, and interaction enable new opportunities, what we call the unanticipated repurposing of information. Single categorization is oblivious to the unanticipated repurposing of information, to the time value of information, to vagaries of semantic and semiotic meaning of words and phrases. It is not even sufficient for single known, preplanned purposes. It is certainly insufficient in a networked information-sharing ecosystem.

AA.2.11.0 Out-of-Channel Multitier Tagging

[0926] Tagging and metadata provides content meaning, context to the content, and a concept of utility, purpose, and applicability to data stream. It enables association of the data to a schema. It enables the transformation of data stored within one data structure into a different structure, into reports, into XML or even into hierarchical markup. Tagging technically refers to the assignment of a taxionomatic assignment to describe the topic or purpose of a unit of information, such as <Top Secret> to a tear line about the source and veracity of a terrorism threat, or <SSN> to a column of personally identifiable information (PII) like social security numbers.

[0927] Tagging has traditionally been applied to database columns as a schema definition. However, the growth in so-called unstructured documents is creating a predominance where tagging is generally applied such that the tags are in-channel with the data; in other words, the tags are interspersed within the data stream. Representations have evolved to break the tagged units into differentiable groupings with indentation often seen with HTML and XML. This has advanced such that the indentation is also viewable within a tree-displayed hierarchy. Tagging architectures have evolved to support meaning, structure, sensitivity-level marking, segmentation of data for legal compliance, and

many user- or application-specific requirements.

[0928] Formally and mathematically, tagging is defined by how Iidentification is used to repurpose resources in ways different from or incompatible to prior representations, thus requiring flexible and overlapping identification systems. Tags are not only about meaning, but about utility, purpose, ownership, source, methods, applicability, distribution, notes, security, sharing, information discovery, resource management, custodianship, and other workflow-specific requirements. The benefit of XML and metadata is the ability to create an out-of-channel description for known and as-yet-unknown requirements. Tagging is not static and must be flexible.

[0929] The Dublin Core Architecture is one attempt to standardize tagging styles with industry-wide uniformity. Although data standardization is essential for progress, obsolete tagging methods are not. In-channel tagging exacerbates security problems by increasing the content of data and its potential for risk. The DigitalDoors technology for creating security for content by granularization also extends to tags and metadata. Granularization is not just about the content, but also about the tags and metadata establishing context and concept. This drives the impetus for out-of-channel multitier tagging.

[0930] Data lives and changes and dies. It is a lifecycle. It is acquired through some collection methods, is typically stored, is processed for one or more purposes, is updated and augmented, is distributed to many other locations (in whole, in part, or transformed), underlies the basis for reports and outcomes, and eventually (or not) decays to uselessness. Data can retain value past its anticipated half-life if references are maintained for distribution as the data evolves. This data lifecycle determines the value of the data and its concurrent security risks as well. Tags and metadata improve the quality, utility, value, half-life, security, and sharing benefits by retaining context and concept.

[0931] There are three significant problems with conventional tagging methods. First, the

addition of new tags generally means the exchange of existing tags, which changes meaning. Second, the tags can be stacked linearly but cannot be multi-dimensional or multi-tiered, which precludes multiple, derivative, or evolving meaning. Third, the tags add, change, or otherwise alter the semantic content, context, and concept of the preexisting or evolving information. This paper will address the ramifications of these three problems and why out-of-channel multitier tagging resolves them.

[0932] Exchange of tags accelerates data closer to its half-life and uselessness. When raw data is tagged, the tags assert a specific meaning and a context for usage. This is usually an improvement to the data because it adds organization, applies a uniform format for consistency and validation of integrity, and it implies or defines a purpose. When data already contains tags but is stripped of the existing tags with insertion of new tags, this also adds organization, applies a uniform format for consistency and validation of integrity, but it degrades prior information content and structure. This is similar to the effects of data normalization within a relational database format. Although some process encourages standardization, it is at the expense of source, method, purpose, context, and so forth. Reference to external tags and all tags as out-of-channel data -- including the original tags -- retains the original meaning in addition to any evolutionary meanings. Security is also enhanced because context and concept of the original and additional tags can be released with granular controlled as needed to establish usage, purpose, distribution, and meaning. This enhances semantic meaning, the application of security, and the control over information sharing.

[0933] Uni-dimensional tags are sufficiently flexible to replicate a database schema and all indexing, references, and relationships. This functionality comes at the cost of verbosity, security, and inflexibility. Although tags and metadata are part of the schemas for structured data like databases, they are rapidly becoming incorporated into the data content channel itself through XML and derivative

hierarchical structures. DigitalDoors white papers have detailed the security risks of compound document formats like XPS and MS Office document formats elsewhere. This significantly degrades data security and the privacy of PII. Inflexibility prevents the reassessment of the actual content, how the context colors that content, and how concept can establish purpose and utility. As such, inline tags support only a single dimension of aggregation. It is hard to overlap them, assign tags in parallel, or establish conditional assessments. For example, a tag can describe the target content of the tag as top secret, in the French language, encoded in Unicode characters, and as important. It does not efficiently support situational meanings and purpose-driven taxonomies, for example, "wedding" as a normal noun and event in a conventional usage but also could reflect a keyword for an event of terrorism depending on context when the source is a known terrorist. Often such contextual evolutions require changes to templates, specific explanations of the intent of the tag additions and metadata solely for these extensions. This has additional ramifications in terms of how data and the tags on the data are upgraded or downgraded.

[0934] The assumption that data has a single or a few controllable purposes misunderstands the knowledge management changes brought by the Internet and the widespread integration and distribution of workflow. This undermines information evolution, data repurposing, and the verbosity tends to accelerate data towards its half-life of uselessness. This rigid structure also precludes access to the full value of convergence and divergence search technology, an assessment of security risk through aggregation, inference and interaction, and the appropriate ecosystem for inter-domain information sharing.

[0935] Tags add, change, or otherwise alter the semantic content, context, and concept of the preexisting and evolving information. While this extends the half-life of data because it enhances the

value of the data in terms of its actual content, context, and concept, it also in conjunction extends the risk because of the enhancement to the value. The permanence of tagging systems and sensitivity-level labeling, like the Ball-LaPadula model (BLP or BLM), do not support the change in data or its risk and do not enable efficient upgrading and downgrading, sometimes precluding it entirely. BLM supports only the upgrading of data, but this exacerbates the so-called "black-hole" where data disappears in higher and higher sensitivity levels and locks down potential information sharing. Inline and single dimensional tags fail to enable data repurposing as previously stated, but they also preclude the evolution in the knowledge management and semantic understanding of the data. Old data, even data seemingly well past its useful half-life, might contain seeds of new discovery, divergence search potential, and transformation into new formats if the legacy of the old meaning is augment through new meaning as represented by new tags and new metadata.

[0936] Although realistically the difference between in-channel and out-of-channel tagging is a matter of packaging and delivery, separation, extraction, redaction, and dispersion is necessary to establish a security ecosystem supporting knowledge discovery, security, and sharing. In-channel tagging recalls the shortcomings in security based on perimeter methods and point solutions. Out-of-channel tagging disperses the risk and the possible attack vectors and reduces the value and hence risk of content through granularization. Granularization is not just about the content, but also about the tags and metadata that establish context and concept. The separation of the data content from context and concept, semantic assessment of content within the data, and a functional redaction enables BLM compliance with the association of data evolution and tag enhancement by reference rather than by massive file revision and reassessment. Granular content control on-the-fly enables the security ecosystem otherwise impossible.

[0937] This dynamic change comes with the risk to security and information sharing, so the risk too must be dynamic and its measurement dynamic as well. Only out-of-channel multitier tagging enables dynamic reassessment of data at the content, context, and concept level and a responsive data-level security by means of the DigitalDoors granular content control technology. Tagging architectures must further evolve to support meaning, structure, sensitivity-level marking, segmentation of data for legal compliance, and many user- or application-defined requirements through multitier tagging.

AA.2.12.0 Multitier Hierarchical Overlapping Tag Structure

[0938] Classification of information and the assignment of tags create as many problems as it solves. Generally, classing and tagging means asserting a single consolidating category with the highest or most meaningful tag that becomes part of the data stream. This alters the physical format of the information and only represents the information in the most general of ways.

[0939] Formally, this threat is defined by how identification is used to repurpose resources in ways different from or incompatible to prior representations, thus requiring flexible and overlapping identification systems. Although it is undesirable to impose what are in effect multiple overlapping coordinate systems, it is commonplace. They develop as a result of specific needs addressed by available technology and do not always become superseded by new ones. Examples include conflicting Japanese house numbering versus western style, route maps, utility layouts, geographic information systems, latitude and longitude grids, and GPS waypoints. These systems are different and sometimes incompatible, but all must be transformed from one in to the others as a practical matter.

[0940] The more robust and flexible solution is reliance on multiple out-of-channel tags, using the DigitalDoors technology of multitier hierarchical overlapping tag structure, to reflect unlimited perspectives of the information without exposing new security risks through the elaboration of tags that

denote and dictate the actual content, context, and concept of it. This is essentially metadata in the form of metatags coupled to the information but not made part of it.

[0941] Another problem of tags, specifically simple tags and more specifically a single tag defining an information unit is that the tag does not have to be correct or useful and can in fact be misleading. A safe tag can mask the danger of malware lurking in information or misrepresent the security-sensitivity. A tag is not all things to all people or all processes, and reliance on a single catch-all representation is simple but also simple-minded and deficient.

[0942] The application of multiple and overlapping tags assigns sensitivity level, compartments, categories, purpose, usage, and other pragmatic meaning to granular content. It is multitier because many tags can be applied to the same content to describe different aspects of the content, and it is hierarchical because tags can overlap to define words within phrases within paragraphs within documents or other compound data streams. The basis for success with granular content control is to establish a comprehensive review of data content, context, and concept as a precursor to tagging and all subsequent filtering and assessment of purpose. Traditional DOD tags indicate a non-reflexive sensitivity level (and/or compartment) but do not in any way describe the potential for inter-domain information sharing except perhaps to preclude it, whereas multitier hierarchical tagging enables security compliance with granular information sharing to multiple domains. This specific implementation resolves issues that plague cross-domain semantic interoperability and the complexity if not impossibility for mapping a consistent dictionary and data model for communities of interest (COI) by reducing the problem of information tagging of the gross document to a granular one.

[0943] Multitier hierarchical overlapping tag structure undermines one single information classification, information sharing, and information distribution method by removing the dependencies

from a single centralized process to a distributed and functionally-independent environment. Rather than assessing the limitations of acquisition and usage to the owner or a managing process, information can have multiple competing and tag uses where the distribution is controlled ad hoc subject to situation, evolving requirements or risks, and distribution workflow.

AA.2.13.0 Dictionary Encoding

[0944] DigitalDoors supports a large number of dictionary encoding options for semantic, semiotic, classification, categorization for utility, perspective, repurposing, and information sharing. These are configurable by the administrator and in parts modifiable by individual users.

[0945] Dictionary entries can be augmented and overridden by functional group administrators or individual users to compensate for the differing security requirements facing a user, the group containing that user, or the larger organization. Users could be members of multiple groups with sensitivity level settings inherited by the appropriate role, group, mission, or situation. As noted in other documentation, sensitivity level domination is the overriding principle to assure granular content control. In other words, it is easier to upcode and than downcode, and it is also automated so that higher sensitivity levels override lesser ones. However, downcoding through manual selection is fully supported to override all prior and automated selections.

[0946] Sensitivity levels, compartments, and other categorization assignments are fully configurable. Names, abbreviations, numerical coding, and color assignments can be changed. All dictionary entries must conform to these coding configurations, but are also fully configurable.

[0947] Dictionary entries include acronyms, words, and phrases to support natural language syntax. Acronyms generally are single word units all in upper case. Words are single word units and can represent all dictionary words, grammar/form variants, common words, proper names, etc. and

capitalization is useful but not critical to recognition. Phrases are a string of unit words. Extension to content meaning is supported by supplemental dictionaries that map synonyms, keys into phrases, partial matches, soundex matches, and inter-language or multiple-language substitutions. User word lists and automatically-generated linear adjacent composites connect into My Group Lists, as do all the basic dictionary entries. This range is large, but by no means extensive or complete, but it can be extended as new content/context/concept interpretation can be implemented. Such flexibility is useful for multiple overlapping categorization or the manual selection or automation of a dictionary entry where multiple categorizations are available.

[0948] Matching also includes word form variants, antonyms, synonyms, and soundex encoded sound-alike words. The overhead required to match form variants and synonyms is substantial, both in upfront administrative time to establish these links (specific to each production environment) and the additional processing time to search for synonyms. Soundex is automatic, fast, but results are many. Soundex compresses every word into a non-unique 4-character representation.

[0949] The optional DigitalDoors symbol permutation engine can process these symbolic substitutions for letters into literal words and phrases. Changes include emphasis on letter substitution to convert strings into meaningful words or more complex phrases. Symbol substitution is not simple because the 2-character ΠΠ (note Pi, Π, and Π adjacent) becomes the 1-character M. The conversion is fast and usually results in a single dictionary match. Symbolic representations of misspelled words is somewhat more complicated because once the alphabetic substitutions are performed, resulting words must be matched by soundex or by a synonym substitution. In that indirect way, 'UR' becomes 'you're' or 'you are.'

[0950] DigitalDoors also enables monitoring and automated dictionary recode throughout the

database to handle changes across all prior processing, recoding through recovery files, redaction files, etc. (not done but would be a good utility).

AA.2.14.0 Declassification and scrubbing of maps

[0951] The DigitalDoors DeClassification process occurs on the end user's desktop, any server, or transactional process. The source document or data stream (to be declassified) need not reside on the user's or server computer but could be anywhere attached through a network. The recovery file and associated recovery map can be created anywhere attached through a network.

[0952] The recovery map is not a file per se, but an computerized abstraction that associates the source file with the declassified file and to the recovery file. Actually, the source file can be safety deleted and all references can be purged to maintain top security. There are no forward or backward links between these files and the recovery map. Loss of the recovery map renders recovery mathematically unlikely. Granularized and dispersed copies of the recovery map guarantee reconstitution without compromising risk and security. Association between declassified material and the source, the recovery file, or the recovery map is one way only, in the recovery map. The recovery map is functionally embodied in one of many implementations, as configured by the administrator:

[0953] The map is referenced by a prefix and the initial source filename.

[0954] The map is a database record keyed by declassified file name with link to recovery file dispersed to one of many storage locations.

[0955] The map is a two-part number [organization reference] [document ID] contained as an e-mail header used to reference a database record dispersed to one of many storage locations.

[0956] The map is known only to a person(s), requiring specific selection by the person, persons, or persons using secret sharing schemes.

[0957] The association between source document or data stream, recovery file, and recovery map is broken by the DigitalDoors process. Filenames or data object references can be randomly generated with unlimited path names with 46-factor independence; this exceeds RSA 1024 bit security levels with only 3 character filenames, and MS Windows easily supports 278 byte file paths.

[0958] The first pass of the process targets potential extractions, performs the extractions, and inserts placeholders. The second optional pass of the process generates a two-step placeholder insertion, such as numbered tags, barcodes, URLs, or pay-for-view recoveries. Numbered tag or two-pass tags (such as barcodes or URLs) allow ongoing editing, as with MS Word, wherein placement and even the existence of extractions could be altered in processes subsequent to DigitalDoors DeClassification, thus allowing alteration and functional ReClassification of the DeClassified document at any future time. A recovery file is generated during the first step of the process. The extracts can reside within this recovery file in the format of plain text, encrypted, plain text links, or encrypted links.

[0959] The format is determined by administrator configuration. The recovery file contains offset to each extraction, the length of each extraction, the length of each placeholder, the category type of each extraction, and the extract in one of the configured formats. The recovery file contains no forward or backward information linking it to the source document, to a recovery map. The declassified document contains no forward or backward information linking it to the recovery file or to the recovery map. When extractions are contained with the recovery file as links, whether in plain text or encrypted, the secondary recovery files for each extraction contain no forward or backward linking information. These secondary, subordinate recovery files can be infinitely chained through one-level forward pointing links. Each such link can be plain-text or encrypted.

AA.2.14.1 Channel Transmission Protection:

[0960] The transmission channel(s) for source data to the user's desktop, server, or transaction process, external generation of the recovery file, and external generation of a recovery map are all authenticated with multi-factor authentication, such as PKI certificates, tokens, or other methods, and encrypted with point-to-point encryption of the transmission channel with SSL or in hardware. This process can be controlled by DigitalDoors or be part of the core networking environment and independent from DigitalDoors functionality.

AA.2.15.0 Distribution of MS word

[0961] Distribution of MS Word or any other MS Office document is always potentially insecure. The .doc binary file security risks with MS Word are consistent across the MS Office tools in all releases of the software. When a MS Word document is saved in ASCII format to force the extraction of metadata, downgraded as an ASCII document, and then imported as an MS Word document, metadata will be reintroduced. It is possible to import the ASCII into the MS Word document on a standalone desktop computer with carefully masked user, organization, and other MS Office attributes. However, that raises the workflow issue of how you get the downgraded ASCII document onto that standalone desktop computer and how you get the downgraded MS Word document from the computer. Importation to MS Word also bloats the file size, generally on at a ratio of 80:1 with ASCII. Furthermore, all formatting markup from headings, to italics, footnotes, links, references, and other features that add value to the MS Office tools will be lost during the process and not regained without additional effort. This effort will tend to reintroduce the thirteen categories of dirty metadata—detailed in Microsoft security briefs—removed in the process of downgrading to an ASCII document in the first place. This is not a viable approach to MS Office security.

[0962] MS Office 2003 introduces a file storage format based on XML. Although the MS Office

metadata is still saved to the file, at least this metadata is stored in ASCII format. The Word file binary format contains a mix of ASCII, binary data, bewildering proprietary data formats with a hierarchical tree structure. Although the XML format retains the same hierarchical structure, the dirty metadata is tagged and can be processed with DigitalDoors to retain functional MS Office operability and secured portability. Complexities abound—particularly with linked and embedded MS Office files—but a secured XML path represents the means to validate the security of a document where the .doc-to-ASCII-to-.doc process is not.

AA.2.16.0 Filters, sieves and capabilities

[0963] Filters and sieves provide data stream classification necessary for the information ecosystem. It embodies knowledge discovery, security, and information sharing. Some interesting principles include:

[0964] Up front classification reduces the security footprint for universal redaction. It is a matter of parsing the data stream for content and context of potential pertinence to establish a thorough and practical classification.

[0965] Comprehensive classification is important for long-term information management.

[0966] Precision classification is not sufficiently future proof since it details existing needs only.

[0967] Precision classification, while useful in some venues, is not specifically valuable as a thorough classification for risk assessment and security because your content of risk does not reflect a creative attacker with a different view of reuse and abuse.

[0968] Initial classification does not preclude improved reclassification in the future or the value of reclassification for granular content control, specifically reversible reconstitution.

[0969] Reuse, repurposing, time factors, plus aggregation, inference, and interaction tend to

undermine initial classification.

[0970] The DigitalDoors classification metadata does not have to be included with the data stream; it can be linked to the redactions, recovery files, and processes or lost after use.

AA.2.16.1 The DigitalDoors classification metadata can be transitory.

[0971] The DigitalDoors classification metadata does not really have to be retained at all, nor in some circumstances should it even be retained as it represents a synopsis of the data stream, a bloat to storage, a risk of equal or greater concern than the original data stream itself, and a drain on system and bandwidth management.

[0972] Classification can always be reproduced, and each pass, in particular with automated classification, is repeatable and consistent yielding identical results or better topical ones with greater reflection of risk.

[0973] The classification can improve over time as we learn more about the universe, the meaning of everything, and our place in that universe. Risk profile change. Some things could well decay in importance while others can become more sensitive, more important, in need of more privacy or greater protections. As such a new classification would reassess prior items, categories, policies, and reflect the current understanding of those issues and how we now comply with them. Note that the data stream, in particular the universal redaction and the recovery files, do not need to be repeated since they already exist in a form sufficient for any future scenario.

[0974] The meanings of the extracted items and the recovery items specifically can be reassessed in light of future utility, reuses, repurpose, risk, and process requirements without any additional parsing and data stream reclassification.

AA.2.16.2 Applicability of Sieves and Filters

[0975] DigitalDoors bypasses most artificial intelligence and other high-energy efforts to assess risk and assign classification. It is one of the primary breakthroughs of this paradigm shift. Absolute classification accuracy is not required. Recognition of false-positive and false negative categorization rates is not required. It is not about accuracy but coverage. The issue really is that all probable classification potentials are recognized up front in order to create the universal redaction. Universal redaction is about denying the value of the data stream until such time as it is precisely understood what coverage is required for functional reconstitution.

[0976] In fact, efforts to most accurately redact or control information flow up front are falsely predicated on presumptions, expectations, comforting but imperfect understandings of the potential value of the data stream. As a result, content recognition is more likely to go astray than provide effective risk management. Instead, this broader classification agenda, as advocated and implemented by DigitalDoors, assures broader coverage. It might be counterintuitive, but perhaps this very broad-spectrum approach is the reason more precise but intensive approaches have proven to be ineffective and exposed to an unending arms race.

AA.2.16.3 Manual, Assisted, an Automated Modes

[0977] DigitalDoors supports data stream content and context modes where users do it themselves, seek computer-assisted support, or are aided completely by automation. Manual is just that, a do-it-yourself process. Assisted is a computerized suggestion process that is surprisingly accurate but very configurable to include cultural and technical variety and the changes that occur in policy over time. Automated mode combines manual mode, assisted mode, and all organization additions and changes to replicate those modes in a computerized hands-free process. Note that manual selections can be incorporated into assisted or automated modes before or after automation. This improves accuracy and

consistency, but streamlines the process.

[0978] You do not have to pick or settle for any single mode. These modes are interchangeable, can be intermixed, repeated endlessly in the style of a word processor with spelling and grammar checking tools. This analogy is appropriate, since a final required manual mode adheres to military two-man review mandates. Administrator configuration can allow overrides or enforce COMSEC policies where the conflict resolution rules are fully automated.

AA.2.16.4 Dictionaries

[0979] Dictionaries are databases of acronym, word, and phrase items packaged with access keys, partial keys (part of the whole), and multidimensional classifications. There are also dictionaries of semantic groupings too with groups, categories, and policies. Dictionaries include any mix of acronyms, words, phrases, and patterns that are bound by category, sensitivity level, user, and other attributes. Patterns include configurable regular expressions, bit patterns, multimedia, metadata, audio, and video. Multiple entries are allowed for the same item, although the attributes must vary in some way, either by categorization, definition, user, etc. These entries are differentiated by a 'coordinate' mapping system where each coordinate expresses a different classification, categorization, or other purpose. Multiple entries could carry different connotations, different categorization, different sensitivity levels, and different categories top reflect the evolving content and contextual value. DigitalDoors supports unlimited contexts.

[0980] The dominant context is not necessarily the accurate context. Neither entries nor definitions are static or being perceived as static. These values should be flexible and should evolve. Technically, each item within a set of multiple entries must have some significant attribute to create a sorting difference. In actual use, the dictionaries are accessed through multiple keys for different multiple

purposes and applications. This enables multitier hierarchical overlapping classification. Although the dictionaries seem to be focused to classification of content they also address the context and concept issues too. It is multidimensional.

[0981] Entries can characterize differences through lower case, upper case, and mixed case, such that JON would generally be interpreted as an acronym, jon as a common word for a toilet, while Jon would be perceived as a name. Each variant and spelling or presentation variant has value in initial classification for determining complex context assessment and more value for compliance with reconstitution and sharing. Actual attribute differences determine categorization, and application is determined by a hierarchy of format, format lengths, priorities, and contexts. For example, XXX > (dominates) Xxx > (dominates) xxxx > (dominates) xxx. This is actually efficient since the hierarchy is generally consistent with traditional dictionary sort orders. Longer phrases and longer words have precedence over shorter ones, and inevitable classification conflicts are resolved where possible in advance; otherwise, the dominance policy is applied automatically, assisted, or manually if so chosen, to rationalize overlapping and competing classifications. A simple set of less than a dozen rules, detailed in another DigitalDoors white paper, seem to fully describe and implement a simple and efficient mechanism for classification dominance conflict resolution.

AA.2.16.5 My Word List

[0982] This is a user-created list of acronyms, words, and phrases of particular importance either in terms of recognition or security. User-created lists exist for each user and can be shared or distributed. Application of My Word List is by user choice. The list can include 0, 1, or any number of items. Profiles for each user or multiple profiles for a single user establish flexibility for different assignments.

AA.2.16.6 My Group List

[0983] This is a user created list of named importance or sensitivity targets represented by the inclusion of specifically-included acronyms, words, or phrases within a defined appearance range with a sensitivity level assignment that dominates all individually included entries. User-created lists exist for each user and can be shared or distributed. When the group is triggered, that range or a larger defined range is categorized. This feature is useful for assessing context where content is not sufficiently clear for word-level or even phrase-level classification. Group list is effective for locating risky contexts or downgrading really innocuous ranges that include one or more traditional triggers because it transcends the limitations of dictionary entry categorization.

[0984] Each group list can include 0, 1, or any number of items. Profiles for each user or multiple profiles for a single user establish flexibility for different assignments. This user created list is equally useful to white list complex contextual ideas as it is to black list them. However, realize that dominance and conflict resolution policy must be tuned and process sensitivity-order might need to include manual or sequenced aspects to preserve white listing assertions through to universal redaction. Dominance for lowering sensitivity level and classification is still fairly represented and enabled with the built-in conflict resolution tools, you just have to be sensitive to the anticipated and initially unanticipated effects of the Administrator taxonomy policy configuration.

AA.2.16.7 My Category List

[0985] This is a selection of the available categories and corresponding sensitivity levels with a sensitivity level that dominates all individually included dictionary entries. User-created lists exist for each user and can be shared or distributed. This list is applied to classify only those items in the data stream that are specifically represented by the included categories. For example, all names can be located and globally tagged as "Top Secret" or "GLBA-sensitive" depending on the classification,

categorization, sensitivity level configurations. This user created list is equally useful to white list complex contextual ideas as it is to black list them. However, realize that dominance and conflict resolution policy must be tuned and process order might need to include manual or sequenced aspects to preserve white listing assertions through to universal redaction.

[0986] Note that the categories of error, unknown, and unidentified are special in that sensitivity level coding is always raised to the highest available level and included in categorization classification.

AA.2.16.8 My Policy List

[0987] This is a meta-selection of categories implementing safe harbor compliance rules with a sensitivity level assignment that dominates all individually included categories. User-created lists exist for each user and can be shared or distributed. A collection of categories is a very powerful classification tool that transcends dictionary entry categorization because the categorization is applied to encompassing sets rather than to entries within sets. This is very similar to the notion of attribute inheritance.

[0988] Consider for example, the HIPAA compliance regulation that is generally implemented so as to include 18 designated categories of information that cannot be released. This includes names, addresses, ages, social security numbers, and the policy would classify all such items in the data stream that are represented specifically by these categories as a HIPAA restriction. DigitalDoors is applying classifications to sets of items rather than to items individually. It is more accurate because classes of items receive similar treatment through the uniform application of inheritable attributes. While it is conceptually more efficient, processing overhead is comparable to entry or category assignment. However, defining and maintaining a policy requires less labor and ongoing effort than the maintenance for the relevant dictionary entries. A policy inherits all the configurations associated with existing categories and all existing dictionary entries in each category. You create a set single policy variant

definition rather than creating dictionary entry variants that number in the tens or hundreds of thousands of entries.

[0989] User-created policy is equally useful to whitelisting complex contextual ideas as it is to blacklist them. However, realize that dominance and conflict resolution policy must be tuned and process order might need to include manual or sequenced aspects to preserve white listing assertions through to universal redaction.

[0990] Policies include categories, and for example, the HIPAA policy could warrant an Administrator configured category called HIPAA that could include specific dictionary items with specialized attributes like sensitivity level or compartment that will likely conflict with last name, name, or medical terminology and will automatically resolve through the conflict resolution dominance rules. As a result, special items specifically included into this hypothetical HIPAA category are subject to the same consistent rules as any other duplicate or multiple dictionary items. Upcoding or downcoding results are contingent on the Administrator-set taxonomy configurations or by user range override choice, if that is enabled.

[0991] However, before creating a HIPAA policy with an included HIPAA category, consider that the functionality of a meta-selection applies a different categorization standard through the subtle application of utility and purpose to classification rather than the black list/white list simplicity. Also consider exactly what dictionary entries you might include into such a HIPAA policy-category extension that would not be reflected by the HIPAA policy itself. Also consider the considerable management overhead for dictionary entries, and in particular complex and frequently duplicated entries. Some organizations will dump patient and staff databases for uploading, reformatting, and row-level categorization into the DigitalDoors dictionaries to populate the dictionary with HIPAA-specific

category designations. This does work and enables savvy organizations to tailor DigitalDoors to specific and/or unusual functional requirements. While this makes some sense for optimizing functionality and processing, such wholesale information is really not inheritable meta-selection material but rather better served as specialized organizational categories and dictionary entries. But, there are better ways to implement, and this is part of requirement definition, installation, tuning, and optimizing for operational efficiencies and compliance.

[0992] Note that the categories of error, unknown, and unidentified are special in that sensitivity level coding is always raised to the highest available level and included in categorization classification and in every policy definition.

AA.2.16.9 AutoSEC

[0993] This represents application of minimum categorization of an entire data stream to baseline sensitivity level.

AA.2.16.10 OPSEC

[0994] An assisted application of dictionary-based white list patterns and black list patterns with gray list handling to preclude gaming against the system to assist classification. It is a classification assistance and suggestion process.

AA.2.16.11 COMSEC

[0995] An automated and enforced compliant application of dictionary-based white list patterns and black list patterns with gray list handling to preclude gaming against the system to assist classification. COMSEC is a mandatory enforcement that can be automatically applied to any and all data streams or applied as an automated final pass after manual and assisted modes. COMSEC is final and allows no additional manual editing and reclassification.

AA.2.16.12 Manual Selection

[0996] This is a simple and intuitive mouse- and key-directed process to select bits, characters, words, phrases, lines, sentences, paragraphs, or entire documents for classification.

AA.2.16.13 Manual Mouseover Selection

[0997] This is an assisted mouse- and key-directed process to select a semantic range of bits, characters, words, phrases, lines, sentences, paragraphs, or entire documents for classification. This is a rapid method to locate units of meaning for classification or establish tear line selections.

AA.2.16.14 Tearline Creation

[0998] DigitalDoors enables roundtrip recognition, removal, and creation of inline tagged tear lines. The process can be manual, assisted, or fully automated such that selected ranges are upcoded to the logical end sensitivity level and classification necessary for effective redaction and distribution. Dictionaries with OPSEC and COMSEC can drive a hands-free automated process that rivals extensive and time-consuming manual efforts.

AA.2.16.15 Other configurations

[0999] A number of Administrator and user configurations enable or disable features that enhance the productivity of the classification process. This includes manual, assisted, interrogated, and automated process that add unknowns to dictionaries for future use or automation, that fine-tune the definition of selections with multiple classifications, that enhance risk assessment and management, or add to legal compliance.

AA.2.16.16 Accuracy and Granularity

[1000] A number of configurations are available in DigitalDoors to control classification accuracy and the level of granularity desired. These configurations also provide flexibility for different

environments, different risk or security requirements, and varying types of data streams. These configurations seek to address any open-ended regulatory compliance or organizationally-mandated security requirements. They are part of the hierarchy or rheostats. Administrator configuration provides Dictionary Management for acronyms, words, and phrases.

[1001] Select a single element even when multiple appear, as with bin Laden. Multiples reflect different categories, sensitivity levels, definitions, or other differentiation. This activates the detail information for the selected dictionary entry. Note that you do have access to each of the multiples; just select the drop-down selection for the multiple number desired.

[1002] The checkboxes can include or exclude each dictionary entry, even one of multiples from different granular processes. Not checking the include box leaves the entry in the dictionary but precludes its general use. In this way, the entry exists and is documented, but does not apply for the organization. By specifying a username, you can limit the applicability for an entry to one or more users rather than a global default. The checkboxes include in small dictionary, include in large dictionary, OPSEC, and COMSEC include or exclude correspond to activation for each of the specific filters and sieves.

[1003] Compliance filters, specifically Category and Policy, can be turned on or off independently. When the My Category List filter is not enabled, classification is performed in the formal hierarchy defined previously. When My Category List filter is enabled, the entry category overrides the individual entry. For example, if bin Laden the family is unclassified, but bin Laden matches with the terrorist category, and the terrorist category is top secret, the category sensitivity level will control. The same hierarchy exists for My Policy List also, and it overrides all category classifications.

[1004] My Policy List is not enabled in the next image. This disables all policy-related

classification.

AA.2.17.0 Automating filters and sieves

[1005] DigitalDoors provides classification in a number of discrete methods using filters and sieves. It also enables fully manual user selection and classification with automated methods that are then rationalized and resolved for conflicts. DigitalDoors enforces explicit categorizations by matching data stream content to dictionary items. Unknowns are also explicit categorizations to defeat plays against white lists, black lists, and Bayesian inference techniques. DigitalDoors locates content and context by aggregations, inference, and interactions. These resulting secondary semantic and semiotic associations often require dominance over simpler explicit categorization. These classifications are rationalized and resolved for conflicts at a complex level. Classification conflict is perceived as normal and expected rather than the exception. DigitalDoors handles conflict automatically. It is the foundation for hands-free information classification, a means to prevent against the system, and basis for many of the breakthroughs in DigitalDoors information technology.

AA.2.17.1 Classification Conflict Resolution

[1006] The basis for classification conflict resolution is coordinating the superset of all legitimate constraints to initially fulfill but maintaining the constraints individually for controlled reconstitution. This balancing act includes the initial creation of a universal redaction that represents a secured data stream footprint. This contrasts with the eventual mediated reconstitution of the data stream driven by utility, security, purpose, process, and legal compliance. In addition, all efforts for rationalization and conflict resolution are retained through to granular content control to minimize typical classification losses into the proverbial black hole.

[1007] Granularity tools allow for harmonizing competing constraints, so as to account for:

1. Sensitivity level overrides

2. Conflicts between security and operations

3. Organization, Group, User and COMSEC overrides

4. Intercession of Automated with Manual overrides

[1008] DigitalDoors advances classification automation with sieves and filters. Automation enables greater speed, integrity, and efficiency. Automation bypasses problems of prejudice, personal preferences, judgment, failing attention-span, and human accuracy. It is all about data preparation as the ecosystem foundation for knowledge discovery, security, and information sharing. This technology is especially relevant to automated content analysis, risk assessment, compliance. It is a core DigitalDoors enabling technology because it generates a universal redaction and enforces compliance with future on-the-fly secondary distribution and reconstitution stipulations through granular content control.

[1009] It is noteworthy to recognize that many potential classification conflicts are resolved through Administrator maintenance and user configuration of special sieves and filters, such as My Word List, My Category List, My Group List, My Policy List, and the various dictionary entries. Dominance and associations are automatically checked for conflicts that can be prevented ahead of time to avert automation disruptions. Examples include a policy with a dependent category of higher sensitivity level or group conflicts with the inherited dictionary entries that force attribute interdependencies.

[1010] Dictionary entries form the basis for black lists and white lists. DigitalDoors classification is not based on mere binary allowance or disallowance. Classification is affected at category-level, in fact multiple and overlapping category levels, and at sensitivity level and compartment. Filters and sieves expose context sensitivities but also downcode meaningless ranges peppered with misleading use of key words and phrases. One of the traditional criticisms of existing classification policies and processes is

the number of mutually-conflicting agency-centric tags and the emphasis to tag everything into a black hole where information becomes inaccessible. The counterbalance to this is the ability to rightcode and even downcode where appropriate. This also allows future access to universally-redacted documents -- even ones ostensibly lost into a black hole -- with information sharing through granular reconstitution. Initial data stream classification sets the basis for universal redaction.

[1011] Recovery and future dictionary reassessment of items, categories, sensitivity levels, users, situations, and sharing requirements additionally determines the reconstitution on-the-fly subject to either the initial, later release configuration, or combinations of both. Initial selection, extraction and redaction creates a distributable basis for information sharing that does not abend the workflow but reconstitution really enables the flexible functionality that maintains the value of the work in the flow.

[1012] Classification (also called categorization) as realized in general usage today usually focuses on the creation of a single dimensional classification, or sometimes several related in detail or subordinated to the major one. However, such classification is driven by the presumption of a clear and precise purpose for that index. Single dimensional classification has no degrees of freedom for hidden or complex messages, information extraneous or ancillary to the core message and support for hidden agendas and sophisticated word drama. It creates one index of focused purpose only. It is a presumption of intended utility. For example, consider "Marketing Literature: Product A" rather than "Product A, marketing, sales, technical details, process features, utility, usage, overview." This example is focused to content, rather than the overarching meaning added with context and concept. Consider which one classification you would rather explore and which one would generate the more effective information management infrastructure for knowledge discovery, security, and sharing.

AA.2.18.0 Processing sets

[1013] Counting in the computer world is through the set collection of {0, 1, many}. Zero and one are just regular items in this set, but are often are reflected by complex or convoluted programming logic. Zero is represented by an initialization without processing, and one is represented by the same setup but a single controlled process. Two and more are repeated instances.

[1014] Many includes 2 and more, but specifically is an unlimited repeated collection, but it also includes 0 and 1. The special cases of 0 and 1 really should not be handled as special cases, but sometimes are because computer programming code is not written flexibly enough to stop a process after no passes or only one pass, and often reflects 0 and 1 only as exceptions. That is not an inherent programming problem but merely an implementation defect that should be corrected.

[1015] Computer processing is good at repetition, so the repetition should handle processing of 0 or the empty set as a situation where nothing needs to be done. The processing of 1 really means that the process handles that instance and nevertheless at the conclusion still checks for the need to repeat the process again. Process of 2 or more is many and should never need special case handling. Relationships of 1-to-1 or 1-to-many sometimes create special cases only because the design is simply not robust enough to handle the situation where there is no match or only one match. It should not be handled as a special case, such that master-detail table relationships should resolve master records lacking detail records gracefully.

AA.2.18.1Collections

[1016] A collection is any grouping of objects that can be seen, bundled, or put together. It is a container with any number of like objects or a container of containers, where the 0 or 1 count is really nothing special and just an instance of what is allowed.

AA.2.18.2 Sets

[1017] A set is a collection of objects called elements. This includes the empty set. In this scenario, a set is a collection of objects that really makes no distinction between 0 elements, 1 element, or many elements. Sets are used in computer processing but also bridge the gap between logical process and math theory.

[1018] Unless otherwise explicitly defined one should always presume that a reference is explicit and simple (a count of one or zero) rather than a collection of many.

AA.2.19.0 Failure of best of breed security

[1019] The "Best Breed in security does not deliver the needed solutions. The best of breed chips away at the background noise but does little to protect against evolving attacks and intelligent adversaries. It is a rationalized effort based on inertia, customary roles, continuing practices, systemic reinforcement by internal supervisors and external vendors that assert it makes sense and that it has always been done that way.

[1020] The best we have stops the loss of the lowest hanging fruit. Best of breed does nothing to divert the motivated, incessant, and evolving threats. If an attack fails but the prize is worth winning, the attackers return with revised efforts. There is a known the maxim, "if at first you don't succeed, try, try again."(Teacher's Manual (1840), Palmer, Thomas H. "'Tis a lesson you should heed, Try, try again. If at first you don't succeed, Try, try again.")

[1021] There is no reward for security attack failure, but no punishment either. There is so much noise that even law enforcement is interested in their lowest hanging fruit. Most failed attempts go unnoticed. Even the successful breaches of security rarely garner reproach, arrest, or repercussion. Botnet attacks hide the source and shield the attackers from any and all repercussions, and perhaps even generate opportunities to make new enemies where none existed, diverting suspicion to innocent parties.

[1022] The traditional best of breed for security is an aggregation of perimeter-based security methods and point solutions designed against specific attack types. Perimeter security is sufficient when the perimeter is actually sacrosanct and not distended to incorporate workflow integration and the ambiguous pel-mel of incremental process evolution. Calling a weakness in the perimeter an 'endpoint' is merely another means to hide the insecurity of touch point(s). Point solutions are sufficient when the security mechanism matches the risk profile one-for-one and covers multiphasic and evolving risk. Unfortunately, both perimeter and point solution sufficiency is unlikely. Perimeters get distended, and point risks tend to have multiple attack vectors but generally only one designated security fix, not the many discovered over the long tails. The application of some or even all of these approaches does repel some attacks and maybe even the majority of attacks, but not all of the attacks and allows more of the PII and trade secrets to leak out. Attacks that get through are increasingly damaging. The lowest hanging fruit has either been protected or picked clean. The more valuable and more difficult fruit is under assault. Data fields and structured tables will give way to unstructured intangible property because the risk is dispersed and the monetary value is so much greater.

[1023] The best of efforts is forming panels to study security problems and pick the best of breed to avoid the mistake of doing something different. The best of breed is at best a faux security. Both effort and breed provide some rationalization that security does something; the incident logs prove that. It provides rationalization that incremental improvements will address new venues of attacks and patch existing gaps, so perhaps the environment will improve. Ultimately that is self-deception, a rationalization that somehow, something will change. But it has not and is unlikely to change within the traditional security venue. The best of breed is ultimately based on a broken security metaphor than cannot be patched, cannot be augmented, and can only be replaced by a new paradigm. Meanwhile, the

best of breed is faux security.

[1024] Best of breed falls into two categories. The first is the application of tools that are acknowledged to be provide the best techniques understood and available providing really nothing but the protection of lowest hanging fruit. That is faux security. It just feels good. The second involves a dramatic shift to a new paradigm filled with changes, new methods, and changes to the ways business function and the shift to a production model based on the information revolution. Such change is difficult, requiring agents of change, and an openness to concepts of asymmetry, knowledge-based production, and formlessness that describe the future of both security and wealth creation.

AA.2.20.0 Integrated Single Files, Attachments, and Links

[1025] The technology enables security without disrupting the normal operational workflow as is typical with encrypted documents. DigitalDoors technology supplies readable and workflow-compatible single files and E-mail messages with recoverable attachments. The technology enables security, privacy, anonymity, survivability, and confidentiality integrating with document management, indexing, data mining, database computation, and transaction processing.

AA.2.20.1 Stream Outputs

[1026] The focus here is on the creation of a single production declassified file with integrated recovery information or recovery information as a file attachment typical with E-mail process. This adaptation is primarily a repackaging issue that streamlines workflow and simplifies management. One or more recovery files are encrypted or multicrypted and appended, merged, attached, or linked to the declassified source stream (transaction, file, or document). This forms an integrated distribution file, a single object, or a local connection to the recovery information, requiring no remote network connections as part of the authorization process for partial or full reclassification.

[1027] The single production declassified file is structured in two parts. The first part represents a typical declassified file with extractions, in situ encryption, or misinformation. The presentation options are a function of administrative and user configuration and also a function of workflow requirements. The second part of the file is encrypted or multicrypted recovery information. When the file is processed through the reclassification process, the second part is stripped from the declassified part of the file, decrypted, and processed according to standard granular access permissions. The reclassified output will appear just as any other partial or fully reconstituted data stream or file. This packaging enhancement precludes the need for corporate network, Internet, or VPN facilities to reach the recovery file.

[1028] The attachment feature is simple the recognition that an encrypted or multicrypted recovery file can be delivered as part of a standard POP-format e-mail message. The attachment (or a hierarchy of attachments) contains the message subject and text recovery information and also recovery information for any user-attached files, which might have been declassified as part of the message delivery workflow. This packaging enhancement precludes the need for corporate network, Internet, or VPN facilities to reach the recovery file. In the case of E-mail, the message could have been downloaded at a prior time and read in part of in full with a remote connection option.

[1029] Links (URL, x-pointer, and other formats) can be local or remote, or a hierarchy of local and remote, thereby supporting granular levels of access without remote access control.

[1030] Note that single file packaging, attachments, and imbedded or appended links alter the structure of the source data stream. This change in data structure may require minimal process changes for non-structured documents, a reconfiguration of processing for structured documents if only to ignore the integrated recovery information, and no workflow affects for database transactions. Database

transactions potentially have field additions with file regeneration for the database administrators, but are unlikely to create any workflow alterations. However, the single packaging as a file or object structure simplifies and streamlines many aspects in continuous workflows with providing cleartext rather than unitary encryption.

AA.2.21.0 Hierarchical Security Sanitization and Reconstitution

[1031] The sanitization and processing of a data stream into non-sensitive and sensitive items is unlimited in that the original data stream can be decomposed into an arbitrary and unlimited number of subordinated representational streams. DigitalDoors can create an unlimited number of transformations of the original data stream and also an unlimited number of aggregations of sensitive items. Security is created through the process and method of dispersing content and granularizing items that define context so as to preclude information knowledge, aggregation, and inference.

[1032] Maps and metadata for mapping, reconstitution, and (none-to-partial-to-complete) recovery of the original data stream can also be decomposed into an arbitrary and unlimited number of streams and dispersed to an arbitrary and unlimited number of local or remote storage locations. DigitalDoors can create an unlimited number of transformations of the original data stream. The purpose for the DigitalDoors data stream transformation is: to secure content; to secure context; to prevent data aggregation and inference and to preclude data repurposing by insiders and attackers.

[1033] DigitalDoors can create an unlimited number of aggregations of sensitive items for knowledge management. Reconstitution, therefore, is controllable from none-to-partial-to-complete contingent upon a role-based multi-level security identification and authorization hierarchy. As a result, security is created by minimizing the access to the content in the data stream and thereafter enabling hierarchical reconstitution of granular pieces into the data stream in a controlled environment.

[1034] Vetting through identification and authorization enables (none-to-partial-to-complete) recovery of the original data stream or a step-wise defense-in-depth reconstitution in one, two, or simultaneous multiple levels of hierarchy. In other words, reconstitution with semblance to the source data stream can be recomposited in situ with the redacted sensitive items, misinformation, or other tags, tokens, and labels. Filters enable this multi-dimensional simultaneous hierarchical redaction and reconstitution into one, two, or many simultaneous data streams with controlled multi-level versions, or the exact reconstitution of the original data stream.

[1035] A 'sensitive' items is any type of bit pattern within any type of original data stream which is identified by the DigitalDoors processes and methods. A sensitive item can include characters, words, audio segments, image ranges, synchronization symbols, phrases, paragraphs, markers, structures, references, metadata, formats, indexes, tags, labels, tokens, URLs, embedded constructs, metadata, or pointers to external structures. Where a sensitive item creates a dissonance in the functional flow or contextual meaning, the range for the sensitive item can be extended to include punctuation, markers, additional words, and linguistic or data processing formats to preclude inference to this sensitive item. A sensitive item is identified at any arbitrary level or simultaneous multiple levels of hierarchy, such as sensitivity level or levels, compartment or compartments, parts of speech, potential overlaps into various purposes, etc. Filters enable this multi-dimensional simultaneous hierarchical characterization of the item.

[1036] Substitution is used to replace the sensitive item in the original data stream. The substitution is usually represented by a null set (that is, nothing) to preclude inference of the missing sensitive information. The substitution can also include a non-null set, represented by metadata tags, standard DoD labels, referential linkages, or placeholders sufficient to maintain functional computerized

processing. Placeholders can include tokens, misinformation, or other hierarchical structures that are arbitrary or unlimited.

[1037] It is critically important to recognize that the fundamental 'replacement' for the sensitive item is out-of-band to the source data stream; in other words, the replacement is the null set but could be a hidden marker, visible marker, or tag presence. Such in-band replacement of the sensitive item with a 'token' is contingent only on conformance with needs to strictly maintain computerized or human activity workflow without disruption. Specifically, many operations necessitate a fixed-sized and precise field structure, column structure, or hierarchical organization and disruption with high-byte toggling, tokenized tagging, or alteration of the data stream with in-band substitutions will cause operations to crash.

[1038] The sensitive items are discarded, or dispersed in an unlimited number of local or remote storage locations. Items can be saved each within its own unitary streams (item-wise), or aggregated together. Aggregated streams o items can include all items of like sensitivity level or levels, compartment or compartments, parts of speech, or coordinated for a specific purpose or security requirement. In other words, a sensitive item might be stored as a single element, combined with others all conforming to a single sensitivity level, or combined into a hierarchy with a minimum and a maximum sensitivity level.

AA.2.22.0 Software implementation

[1039] The purposefully-simple user interface hides sophisticated processes. DigitalDoors masks the process with an intentionally pristine single document window. Other subordinated windows are hidden to reinforce the simplicity and focus emphasis on results.

[1040] The application suite is a series of standalone and interelated modules that address

security, information sharing, and discovery. Behind the windowed facade include integrated processes for encryption, key management, semantic analysis, classification, risk assessment and management, intangible asset monetization, partial reconstitution for sharing, and dispersion to the cloud. This DigitalDoors semantic granular content control is based on the unified premise of information inventory, classification, selection, and dispersion to the cloud. This implementation serves competing and contracting functions at the same time. There are many masters, many outcomes, many instances of formlessness that are necessarily important although rarely necessary for the user to see. These processes are based on deep technical analysis that usually goes unexplored.

[1041] Because the Windows user environment abstracts things to a simple view, DigitalDoors also abstract these to a simple user interface (UI). The goal is minimal user interaction with maximum results. Automation is available to further reduce user interaction and increase overall productivity. The enormous number of options behind the scenes address the complexity of semantic content. They can be tuned for explicit environments but are not necessary in the ordinary. The user interface hides these options in favor of generic defaults and workflow simplicity.

[1042] Loading and process time reflects on the user interface. It is a function of dictionary size. The more entries in either the small or large dictionary and the more additional specialized dictionaries used, the slower Secure and View run. This is a factor of white lists, black lists, gray lists, and semantic content information used to classify complex and overlapping selections with superior accuracy.

[1043] The options support a complex semantic analysis and production process. There are options for inputs, format and presentation alternatives, outputs, logs, and reports. The logs and reports categorize the ordinary, the exceptions, the default presumed from the many available choices, the errors, the unknowns, the white and black list elements, the classification decisions, and the dispersions. The

user interface and the process is planned for low user investment in time and training and low administrative burden, but supports customization for more difficult classification and security situations. Rest assured, the breath and depth are available when needed, but simplicity of the user interface is a predominant concern.

AA.2.23.0 Software operations: Sensitivity Level Hierarchy and tear line

[1044] Redaction and tear line operations are typically conducted such that the results are specific to the requirement and the intended distribution. In other words, if a document is released to [S] level, the document is redacted to remove all higher [TS] material. This creates a duplication of material that increases the leakage risk. It is also disruptive to process continuity and workflows, increases the overall storage requirements for multiple versions, and encourages the overcoding of material to retain power, security, control, and curtail leakage.

[1045] DigitalDoors provides redaction and tear line through the traditional versions but also by creating a master template and separating the crown jewels individually. This enables reconstitition of information for innumerable purposes and distribution as needed through granular content control. Reconstitition is driven by any one or number of contexts, from sensitivity level to role to mission to user or to group. Primarily, reconstruction is driving by sensitivity level. The sensitivity level hierarchy can be defined so as to include all granular content (raw data and metadata imbedded or extant) up to and including the defined target sensitivity level. Therefore [TS] would include [TS], [S], [C], and [U]. That is defined as inclusionary and is the typical delivery. However, information can be minimized and focused to delivery only [TS] for [TS] and exclude [S], [C], and [U]. That is defined as exclusionary and is the optional delivery.

[1046] This sensitivity level hierarchy is defined by the administrator as either inclusionary or

exclusionary.

AA.2.23.1 Secure Editor application

[1047] The DigitalDoors Secure Editor is a standalone application or a module to add into other applications for plain text and media creation, editing, and sensitivity level tagging. Other types of tagging, including categorization, taxonomy classification, privacy, security, compliance, and semiotic meaning are also available. This technology closes the loop for DigitalDoors such that the full range of document management can be supported into a unified infrastructure, from creation, editing, document markup, tagging, tag conversion, tag removal, context sensitivity level redaction, context reconstitution, and support for complex process workflows. The architecture assures separation of data from metadata so that no security lapses are introduced into the traditional word processing and document management cycle.

[1048] From the user standpoint, the DigitalDoors Secure Editor is not much different from vi, Word, Notepad, and other desktop tools. However, behind the scenes, this application separates the data stream from all markup and tagging for security purposes. The interlacing of user content with metadata creates significant process, storage, distribution, and workflow security failures that are not resolved with current technologies. Current technologies include encryption, firewalls, intrusion detection, perimeter guards, and locked distribution packages.

[1049] The DigitalDoors Secure Editor enables text and media creation. However, all additions, deletions, changes, insertions, and reorganizations and reordering are tracked as metadata that does not become part of the document. The document as seen and shown represents the deliverable format. Since formatting is metadata, it is not included in the representation. Formatting, such font sizing, colors, font selection, footnotes, headers, subscripts, superscripts, line numbering, indexing, and other features

characteristic of standard document preparation can be supported but are represented only as metadata. Tagging, including sensitivity level, categorization, taxonomy classification, privacy, security, compliance, and semiotic meaning are also represented only as metadata. This separation of representation from meta-representation is critical for creating the infrastructure for secure information sharing, privacy, security, and compliance.

AA.2.23.2 DeClassifying and ReClassifying while Browsing the web

[1050] Security for workflows, transactions, and information stores while browsing the Internet and Intranet webs require new paradigms because it is only partially addressed by current methods. The current technology fails to address workflow, convenience, and distribution issues for public and private information, the granularity of delivery for that information, or the centralized and decentralized access to that information. DigitalDoors specifically addresses the security of workflow, transmission, and browsing with methods to distribute and provide granular leveled access to information. The two control points for securing browsing with DeClassification are represented by the de facto delivery of declassified information with access to immediately reclassified information subject to secondary identification and authorization.

AA.2.23.3 Generation and Distribution:

[1051] There are several issues when securing browsing and access to data (web pages, tables, data sets, and linked files). First, is the issue of sourcing information and downgrading it for general presentation. DigitalDoors implements DeClassification for all data stored directly by initiating the DeClassification process with multiple secured outputs where the declassified data is an output. Second, access to the ReClassified material is generated on-the-fly subject to identification and authorization. This is accomplished by initiating the DeClassification process with multiple secured outputs where the

declassified data is an input along with the critical and important extracts. Classified information is encrypted, MultiCrypted, and dispersed to local and/or remote storage locations. Storage of critical and important information is configured and dispersed to different storage locations including Internet and hosted storage options. Dispersal is controlled by policy, configuration options, and threat mode based on users, recipients, situational needs, browsing content, and critical and important information categorization. For example, a static web page or even one generated dynamically from a database containing names and credit card numbers can be DeClassified with each of the two (and other categories) dispersed to separate storage locations.

[1052] Distribution represents no changes to standard browsing process in any form except for dynamically-generated pages require a dynamic recovery file. Distribution is nominally assumed to exist within a decentralized, distributed, and non-secured environment. There is no special need, effort, or consideration for distributing DeClassified browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. Because all static pages are stored in the DeClassified format, access to all pages in a ReClassified format is on a per page basis.

[1053] ReClassification is a function that is controlled and limited by the standard identification and authorization process. This includes the use of agents, and leveled access through secret sharing, alerts, threats, location, timing, policy, access to remote storage, and dispersive storage.

AA.2.23.4 Receiving and ReClassification

[1054] On the receiving side, DeClassified browsing is received as a standard data that required no special treatment, process, or interface. DeClassified browsing is the same as standard browsing as there are no changes within the workflow as perceived by users, servers, or other intermediate delivery processes. The software for ReClassification is itself non-secured and contains no secrets. It is possible

to key each ReClassification tool for unique access by specific user, location, document, timeframe, time expiration limitation, and other configuration.

[1055] The ReClassification process requires a downloadable user interface in the form of a browsing client reader add-in or the use of our separate browsing reader. Selection of a DeClassified data initiates a workflow process consisting of:

Table Declassification Workflow process

Identification and Authorization (either local or remote)

Lookup of recovery map information

Access control by time, location, threat mode, username, category, time of day,

level hierarchy

Access to the recovery map (granular access controlled at all time)

Access control to dispersed storage locations

Recovery of dispersed, Classified critical and important information

Decryption of encrypted or MultiCrypted critical and important information

Reconsitition (ReClassification of DeClassified data)

Display within standard user interface of ReClassified data

(such as within browsing client, PDA, etc.)

Display within separate browsing reader/ReClassifier application

[1056] Controlled release for ReClassification purposes can be limited such that different browsing users will have leveled access to the data, for example, senior users can have access to the names and credit numbers on a granular basis whereas junior users cannot.

[1057] Workflow and transactional messaging activities that require dynamic browsing with

ad hoc generated reports or other similar flow can also be enabled to employ the remote distributed DeClassification and ReClassification processes in similar ways, with generation of a temporary recovery file.

AA.2.24.0 DeClassification Mapping - Channels mapping

[1058] The DigitalDoors DeClassification process occurs on the end user's desktop, any server, or transactional process. The source document or data stream (to be declassified) need not reside on the user's or server computer but could be anywhere attached through a network. The recovery file and associated recovery map can be created anywhere attached through a network.

[1059] The recovery map is not a file per se, but an computerized abstraction that associates the source file with the declassified file and to the recovery file. Actually, the source file can be safety deleted and all references can be purged to maintain top security. There are no forward or backward links between these files and the recovery map. Loss of the recovery map renders recovery mathematically unlikely. Granularized and dispersed copies of the recovery map guarantee reconstitution without compromising risk and security. Association between declassified material and the source, the recovery file, or the recovery map is one way only, in the recovery map. The recovery map is functionally embodied in one of many implementations, as configured by the administrator:

[1060] The map is referenced by a prefix and the initial source filename

[1061] The map is a database record keyed by declassified file name with link to recovery file dispersed to one of many storage locations

[1062] The map is a two-part number [organization reference] [document ID] contained as an e-mail header used to reference a database record dispersed to one of many storage locations

[1063] The map is known only to a person(s), requiring specific selection by the person, persons,

or persons using secret sharing schemes

[1064] The association between source document or data stream, recovery file, and recovery map is broken by the DigitalDoors process. Filenames or data object references can be randomly generated with unlimited path names with 46-factor independence; this exceeds RSA 1024 bit security levels with only 3 character filenames, and MS Windows easily supports 278 byte file paths.

[1065] The first pass of the process targets potential extractions, performs the extractions, and inserts placeholders. The second optional pass of the process generates a two-step placeholder insertion, such as numbered tags, barcodes, URLs, or pay-for-view recoveries. Numbered tag or two-pass tags (such as barcodes or URLs) allow ongoing editing, as with MS Word, wherein placement and even the existence of extractions could be altered in processes subsequent to DigitalDoors DeClassification, thus allowing alteration and functional ReClassification of the DeClassified document at any future time. A recovery file is generated during the first step of the process. The extracts can reside within this recovery file in the format of plain text, encrypted, plain text links, or encrypted links. The format is determined by administrator configuration. The recovery file contains offset to each extraction, the length of each extraction, the length of each placeholder, the category type of each extraction, and the extract in one of the configured formats. The recovery file contains no forward or backward information linking it to the source document, to a recovery map. The declassified document contains no forward or backward information linking it to the recovery file or to the recovery map. When extractions are contained with the recovery file as links, whether in plain text or encrypted, the secondary recovery files for each extraction contain no forward or backward linking information. These secondary, subordinate recovery files can be infinitely chained through one-level forward pointing links. Each such link can be plain-text or encrypted.

AA.2.24.1 Channel Transmission Protection:

[1066] The transmission channel(s) for source data to the user's desktop, server, or transaction process, external generation of the recovery file, and external generation of a recovery map are all authenticated with multi-factor authentication, such as PKI certificates, tokens, or other methods, and encrypted with point-to-point encryption of the transmission channel with SSL or in hardware. This process can be controlled by DigitalDoors or be part of the core networking environment and independent from DigitalDoors functionality.

AA.2.25.0 Security by dispersion – the cloud

[1067] The DigitalDoors selection, classification, and dispersion technology was designed primarily to address a distributed universe of processing and data storage. The cloud is a composite of data, process, and workflow where boundaries represent a distinction no longer viable. The cloud is a single integrated ecosystem that ignores firewalls, perimeters, guards, and security point solutions.

[1068] The first DigitalDoors illustrations defined a network within the traditional framework of a cloud, even multiple clouds reflecting multiple layers of security both isolated from each other and interacting together through DigitalDoors technology. Implementation was applied to enable and support traditional data processing operations within a structured, field-oriented, free-form, unstructured, and database-oriented architecture. Data structure (the lack thereof or complexity) was not perceived as a design constraint or limiting factor. Processing technology for this new paradigm, the generations of information processing equipment, and networking interconnections, were not constrained by available concepts. DigitalDoors anticipated information as the primary resource and the expression of both process and distribution within its own framework as an extension of an evolutionary information and production model.

[1069] Furthermore, the basic dispersion technologies were invented to address the security requirements of a distributed agency and military global information grid (GIG) supporting centralized and distributed processing resources, multihomed delivery channels, uncertain security at the edge, edges under constant attack, silos and stovepipes, uneven methods to interconnect high and low risk resources (MLS interconnectivity), and the realization that warfare had evolved to a point where attacks are amorphous, asymmetric, easily involving both friend and foe alike at the same time.

[1070] Of course, the concept of the 'cloud' that we know today is not the same cloud described by the mainframe and terminal architecture, the network of peer servers, clients and services, remote procedure calls, or a distributed LAN and WAN network architecture that comprises the Internet. It is possible to think of these subsets within the cloud as spaces delineated by perimeters, with the cloud comprising the superset of all possible sets and subsets. The cloud we recognize today is a ecosystem cloud comprising all possibilities, with seamless interconnection and integration, not the specialized VPN generally presumed. Although the cloud of today actually incorporates VPNs and specialized links for silos and stovepipes, any presumption of security microperimeterization is based on archaic views of trustworthiness.

[1071] The cloud today attempts to describe an uniform environment of process, storage, and utilization not defined by any where or any time, and not static in any terms familiar to us now. This definition is still evolving as information and process are mashed together in interrelated but dependent forms. This is the concept foreseen by the GIG and what DigitalDoors anticipated would provide security through dispersion as a fundamental paradigm shift. The cloud transcends perimeters and point solutions and any expectations of when and what order, so solutions that presume a defined location or definitive risk are inadequate for this cloud paradigm shift. The cloud becomes our functional universe where

perimeters and microperimeters are archaic delineations lacking force of security or control.

[1072] DigitalDoors sees information as the primal resource, reflected in data, application definitions, workflows, and expressions of design blueprints that become programs, process, products, services, intangibles, and intellectual property. DigitalDoors sees most resources as a reflection of information, and the enablement of it cannot be based in place or in time. The DigitalDoors technology addresses this cloud as it is exists today, as it develops in new ways, as the distribution of information evolves into new forms, formats, structures, composite processes, and purposes where the primary production model inputs and outputs are data and the production model is an omnipresent cloud.

AA.2.26.0 Semantics dictionaries

[1073] The DigitalDoors dictionaries contain some one million English language entries. Entries are distinguished by type as acronyms, words, and phrases but stored together. These entries are used for white listing, black listing, gray listing and also contextual classification and weighting. It is applied for automated selection, contextual or conceptual recognition, classification and tagging, and also monetary valuation for intangibles. The dictionary includes common entries borrowed from other languages. However, it is not exhaustive. The dictionary is extensible, but it is also supports word-format generation with prefixes and suffixes. In other words, "kind" can become "unkindest".

[1074] Prefixes and suffixes are applicable to words, but not processed for acronyms and phrases. This enhancement is effective for categorizing non-standard or rare forms of common words if the prefix and/or suffix is recognized as a valid modifier in the dictionaries. This functionality is particular relevant for Germanic languages as many words are conjoined.

[1075] Conjoined words Consider, the word "fremdenverkeresburo" which loosely translates as the "travelers information office". Consider conjoined words in other languages, such as Swedish, where

"kommunalskolsflaggstångsknoppuppsättaringenjörslärare" translates quite literally as the "supervisor who teaches engineers working at communal schools to put the top cap piece on flagpoles." This serial conjoining is supported only if the prefix and/or suffix is recognized as a valid modifier in the dictionaries. Complex compound words like these two examples represent an exhaustive parsing process that is not practical at this time. However, common compounds can be added as prefixes and suffixes to enhance basic dictionary matching.

[1076] Root matches provide the classification inherited for all compounded word forms. The classifications includes sensitivity level, word type, user, group, mission, role, and becomes the key for all subsequent My Group List and contextual matches too. Note, however, that precedent assures that any compounded dictionary entry will take precedence over conjoined form variants. This is important and logical since many compound words have cultural meanings no longer rooted in the root word or even modified by the prefix or suffix.

[1077] For example, consider the root word "sight". The prefix and suffix lists allow for compounding and recognition of derivative words that include:

Table of derivative words

Insight	Insightful
Sightwise	Sights
Sighted	Sighting
Sightless	Sightfully
Sightlike	Metasight
Unsighty	Unsighted

[1078] This dictionary enhancement facility is actually useful for mixed language environments

because it lightens the load when looking up words and when trying to ascertain which language is most likely to pertain.

AA.3.0.0 Privacy and security:

AA.3.1.0 Context: The difference between Security and Privacy

[1079] The difference between security and privacy is context with its material effects to specific involved parties. The difference might seem to be nothing, but it becomes something when perceived within appropriate contexts. The application of security is the same as the application of privacy but the effects of a breach in security is contextually different from a breach of privacy based on its effects to specific involved parties.

[1080] Consider PII (personally identifiable information) in the form of a report about a person, one of thousands stored to enable business operations for its custodian, generally a business-oriented institution. The PII includes name, address, telephone number, account number or purpose-specific-index, surrogate keyword, social security number, and account information. The destruction of that PII is a loss of institutional functionality. Whereas, the leakage or complete exposure of that information, how it is being used, and the results of its usage represents a security breach for its custodian, the institution. The leakage or complete exposure of that information, how it is being used, and the results of its usage represent a security breach for its owner or the reference to which it applies.

[1081] Owner and reference are the same in effect but different in terms of legality at this time. Both the security and privacy breach are one and the same events, but the ramifications are different depending on perspective. Context is the only difference between a security breach and a privacy breach. It is a security breach to those responsible for the integrity and storage of the information but a privacy breach when the information refers to someone personal. Piracy, the theft of information and/or services,

is a breach of security and also privacy when the impact includes PII.

[1082] If you think of "security" and "privacy" as the actions to prevent a breach, there is no difference between these implementations. Both are one and the same events. Only the consequences are different, and those consequences are a factor of context and context alone. Any breach of PII will have components of a security loss and a privacy breach. The ramifications will depend on context, and any specific applicable laws or regulations violated.

[1083] More specifically, security is the protection of information about a person, related people, an organization, operations, talent, and processes for which you are responsible through custodianship or stewardship. In contrast, privacy is the protection of information about you, your related people, your organization, your operations, your talent, and your processes. That responsibility might indicate purely business requirements, moral, or even legal ones. The difference between security and privacy is only context.

AA.3.2.0 Security and privacy: protecting the crown jewels

[1084] Data security and privacy requires a controlled barrier to information. Security is the protection and continuity of the process and integrity of the information. Privacy is the control over ownership and use of information. Security and privacy are not antagonistic but rather complementary. Security is continuity and integrity. Confidentiality and privacy is the control of usage and distribution in both the aggregate and details.

[1085] Crown Jewels, as in the national treasure of the Royal House of the United Kingdom, are one-of-a-kind heirlooms assessed at a very high value. This collection denotes the regalia and vestments worn by the sovereign during ceremony and at various other state functions. These items are the most single valuable jewelry collection in the world. The jewels are protected by forts, moats, guards, alarms,

and other perimeter defenses. Some of the Crown Jewels are even dispersed to vaults at other locations. Some of the Crown Jewels are shown in public with copies. The security of the Crown Jewels is based on maintaining the continuity and integrity of this unique resource. The physical copies are not valuable. The value of the original Crown Jewels and the copies together are equivalent to the value of the originals. The loss of continuity and integrity of a Crown Jewel, should it be fractured into pieces, will retain little original value.

[1086] The security of data processing operations have also been based on maintaining the continuity and integrity. Data diddling, mutating viruses, and other intrusions cross the line of control and allow undermining absolute continuity and integrity; this is not a small problem, but is proven to be a resolvable operational issue. The multiplier effects of manufacturing, distribution, and marketing created a world-wide increase in wealth; this has been driven by data processing. The process and collections of information making such growth possible are based on the "crown jewels" embodied by computers and databases. Data has become the commodity of business in this information age. Most data can be replicated by matching the original collection methods, so the value of data is perceived to be the discounted time value of replicating it.

[1087] However, a subtle but significant shift has transformed security into the age of privacy. Copies in data processing are every bit as valuable. The value of data is also through the covert misuse of it. Virtuality has made copies indistinguishable from these original "crown jewels." Thefts by way of copies are not always noticed because the original maintains its continuity and integrity despite the obvious or hidden intrusion. Furthermore, copies of data, subsets of data, reports of data, inferences of data, and aggregations of subsets of data are frequently distributed as part normal operations. Virtuality is not just about theft, cover misuse, but also about the unintended aggregation and inference of freely

distributed data copies and subsets.

[1088] The original data retains its full value. The copies, as many as they are, frequently have as much value as the original data or can complete devalue the original data. How is that again? The original with ten copies is worth eleven times the originals. The original with hundred copies is something less than a hundred and one times the originals. However, a million copies become worthless and debase the original too by forcing its cancellation and deniability. A glutinous distribution conceivably could undermine the entire economy by devaluing the data because this commodity glut results in the loss of its one-of-a-kindness. Many millions of copies can clog the system and curtail all data processing services and data access. In contrast to data over-replication and Crown Jewels, the fragmentation of the data crown data jewels is very valuable as reassembly in context and concept restores all of the original value and perhaps even more through aggregation, inference, and mining. Search engines like Google and Yahoo, establish heretofore unknown patterns in data. These patterns generate new data and hence the need for new types of security.

[1089] This is the mathematics of Virtuality, the loss of distinction between the original and the copy. Even physical commodity jewels suffer loss in value due to virtuality. DeBeers Diamonds understands that the true cost of a jewel pried from the ground is the sum of license fees, taxes, labor, manufacturing, and distribution costs. However, they maximize value through limited editions of this commodity by artificially depressing supply below demand with a controlled release. Is there a lesson here in privacy? Absolutely. Data has no value other than a necessary replacement commodity cost unless its distribution can be restricted. This is where of the DigitalDoors doctrine adds value through controlled release.

[1090] The true value of data is its context, purpose, and availability. Because data mining,

aggregation, and inference provide an effective method to replicate the full productive value and perhaps additional value over the original data, granular control is essential for artificially limiting supply. Restricting access to parts of it retains its continuity and integrity, but controlled release retains the high value and privacy. In terms of data, the limited and artificial release of granular parts of the data maintains the uniqueness of the data virtuality rather than it commodity value.

AA.3.3.0 Surrogate Keywords as unique identifiers for privacy and security

[1091] Social security number (SSN) is a nine-digit key or index used to identify a resource, in this case a person. The social security number was invented to create a purpose-specific index. Uniqueness is not assured because of the methods of assignment because of key reuse and the lack of forethought into the difficulty of process conversion and data translation over the long term. The distribution of the SSN in usage has created a privacy and security problem. There are legally-driven efforts to replace the SSN with another globally unique identifier (GUID). This GUID is likely to have a bigger field size for uniqueness and to support a wider audience containing no partial intelligent values that can expose PII or an associated individual.

[1092] This effort to create an alternative surrogate keyword creates as many problems as exists with SSN plus new ones. In the end, any surrogate key does not address the fundamental issues of privacy and security. Any surrogate key or keyword is merely another purpose-specific index attached to the same or similar personally identifiable information (PII) and all the corresponding risks inherent in the storage of that PII.

[1093] The object for a GUID is to represent the intersection of name, address, phone number, weight, height, sex, and identity of a person (or other resource) with some sort of unique nomenclature. The aggregation of name, address, phone number, weight, height, sex, identity of a person, and other

personally identifiable information (PII) does indeed create a unique nomenclature, but it is not necessarily a unique result. The set could contain 1 or more references. That is a failure. Confirmation of a GUID will require secondary confirmation by means of the PII or SSN. This creates the chicken or egg problem for any assignment of an universal identifier for a electronic health care record system. Composite keywords might seem like a good idea, especially when they are translated from PII, but the reality, like any hash, is that the composition can be recreated and often re-identified through forced methods, coercion, or statistics.

[1094] SSN originally started out with no meaning. Usage and reliance created meaning as the de facto GUID for each person. Meaning was not in the content but in the context and concept. The substitution of serialized GUIDs for insurance policies, account numbers, and transactional references recreates the SSN problem. Eventually, and that could be years or instantaneously, any GUID is referentially linked to PII so any surrogate keyword itself becomes a comparable risk factor because it references context and concept. It becomes yet another de facto ID. The distribution of these quasi-randomized GUIDs and the lack of any as-yet centralized lookup system is not security but only a level of obscurity with an added level of management and operational complexity. Obscurity is not security but merely a whiff of security.

[1095] As a result, the creation of a surrogate keyword only inserts a level of misdirection to the PII, but no protection for it. Use of biometric and N-factor identifiers as keys or ID are wracked with these same problems. Biometric information, albeit physical, is converted into digital representations that are equally liable to fraud, widespread dissemination, and overexposure (lack of scarcity), thus defeating its security. Even encrypted GUIDs are not immune to these problems. If the basic GUID is unique, the encrypted GUID will be unique. The mapping between them will be unique. One of the

fundamental requirements of encryption is the lack of clashes. However, by virtue of the usage of even an encrypted GUID and the likely waterfall effects of distribution, even the encrypted GUID loses its uniqueness and security through usage.

[1096] Consider how banks used mother's maiden name as a key and validation of identity since that name usually has no connection to any account PII. Mother's maiden name was often linked to a Caller ID. Caller ID could be situationally unique, but the ease to fraudulently create whatever one is needed undermines its integrity and usefulness. Its uniqueness is not, and the widespread use of this ID or verifier has been compromised by its very success. However, mother's maiden name has morphed into the "verbal passcode" and this too is neither unique, secure, or free from inference or exposure. This example is typical and prescient of any attempts to create a long-lasting surrogate ID. In the end, any surrogate keyword does not address the fundamental issues of privacy and security.

AA.3.4.0 The Metrics of Privacy: Identification, De-Identification, and Permitted versus Forced Re-Identification

[1097] The following describes metrics to measure privacy and the necessary structures to monitor, alert, and secure privacy. This paper also specifies the corresponding methods for proving the social and personal losses of privacy through associative inference, data repurposing, and data mining with undisclosed identification. It outlines types of identifiers, categories for the same, and how association with public, private, personal (extra-systemic) knowledge, and a circle of association or inference can enable seemingly insignificant details to expose individuals; this yields a definition of what is critical and important to uniquely and sufficiently identify an individual. In addition, it defines de-identification and re-identification processes to set the bar for minimum implementation performance in order to forestall adverse legal encounters. Finally, this paper establishes the outline, design, and

implementational requirements for conforming to data privacy laws in an efficient, necessary, sufficient, and comprehensive process.

AA.3.4.1 Privacy overview

[1098] E-commerce initiatives, online financial services, customer demands, and worldwide privacy legislation drive the cultural response to and need for identity safety, privacy protection, and the defense of personal information. By far the most compelling impetus for implementing these privacy protections is the legal enforcement threat granted by recent laws. HIPAA (Health Information Portability and Accountability Act signed into law in 1996 but implementation delayed until March 2003) privacy rules are the most impressive of these with fines ranging upwards from $250,000 per instance. In most instances of abuse or violation, the laws remain unclear, as is any standing for a remedy unclear. I would suspect that local, regional, and innocuous tort liability law and prior court precedent will be extended so that individuals will be have standing and remedy. If past case law is any indication, privacy and security failures will be prosecuted under the concepts of contributory negligence. However, the potential costs for litigation, the inability for smaller parties to access the court system, and pressure for out-of-court settlements nonetheless cause sub-optimal stress and economic pressures. As a result, physicians, hospitals, pathology laboratories, insurance companies, financial institutions, and payment processors are scrambling to implement complex and vague privacy rules. The key issue is not so much about correctly identifying a specific individual per se but rather not being wrong about statistically placing an individual within an operative and functional grouping.

[1099] All these privacy laws inherently undermine the customary need for information sharing and disclosure. This endangers traditional mechanisms for data processing. Business practices inherently will shift to comply with these laws to enable ongoing business workflow while balancing these goals

without creating new opportunities for justifiable claims of data suppression, insurance or financial frauds, legal malpractice, and misinformation.

[1100] The primary privacy technology is called de-identification, a process where identity and potentially identifying details are expunged from records, reports, and correspondence. Other synonyms include declassification, expunging identity, creating micro-data, data blanking, data blurring, and dis-information. Accordingly, re-identification is the reciprocal process applied to reconstruct the de-identified information and make it whole again. Access control, by way of passwords and biometrics, are anticipated to limit re-identification and/or multi-level access to the original undiminished data—should any be retained in violation of many privacy laws.

[1101] However, the methods proposed to achieve legal compliance raise significant technical uncertainties. These include component implementation and data suppression during medical diagnosis, insurance payment approval, banking activity, financial transfers, clinical activity, and traditional business workflow. Also, the implementation necessary to retain fundamental workflow consistency as a design and technical concern is in direct conflict with workflow. This holds true not only for medical care (directly addressed by HIPAA) but also for a wider range of commercial activities (under the EU Data Directives, Patriot's Act, Financial Modernization Act, Data Protection Act 1998, and many conflicting global and local ordinances). SEC Sarbanes-Oxley regulation and forthcoming congressional law extend the responsibilities of chief financial officers to monitor and report privacy violations and a wider range of indiscretions. This indicates an attorney full-employment boondoggle for years to come.

[1102] The concept of minimum essential identity information—in other words, de-identification of information to achieve a level of information release corresponding to a need-to-know basis—is inherently flawed. Other papers, academic research reports, and the Bayesian statistics herein

demonstrate that a de-identified document is insufficient to preclude inferential and hence forced re-identification of individual(s). Other, additional technologies are needed to curtail this, technologies including leveled reconstitution (re-identification in whole or in part driven by a hierarchy of access protocols), information dispersion, and granular access control. This problem, as well as the previously stated workflow integration, can be easily solved with innovative products and services, such as DigitalDoors' eponymously-named software, based on traditional and logical security methods not yet widely implemented in automated data processing or Internet. (Note that development of the patented DigitalDoors software and privacy framework has driven the academic research and scientific results outlined in this paper.)

[1103] One serious dilemma with de-identification is that the Safe Harbor (specifically under HIPAA meanings) identifier categories and other entity-specific (e.g. "sensitive") information is commonly used as indexes for databases. Removing such information breaks database operations. The process of de-identification renders indexing useless, thereby undermining data collection, data integration, and merging operations. Note that indexing technology is unfortunately sufficient with inferential methods to re-identify people—unless the indexes are also de-identified, an unlikely scenario because it undermines traditional data processing. The typical methods used for government de-identification with tags and XML-like paragraph markup leave category tags in the resulting declassifications. This information is sufficient for human intelligent operations to infer or completely understand the contextual meaning; this is a vastly inferior method, as extrapolated from work by Bell and LaPadula. The generation of taxonomies through rules, phrase extraction, sample and training sets, or other topic-oriented subject categorization by means of intelligent systems with replacement of those taxonomies with labels or tags allows significant information within the context; this is not overcome

unless the labels or tags themselves are completely removed and substituted placeholders are statistically random in designation and size.

[1104] Data entry errors, which typically start at 4% for names and addresses, increase to as high as 20% for long numerical strings. Long numerical strings include assigned ID numbers, social security numbers, and lab results. Other problems, such as name aliases, misspellings, and the lack of name entry consistency, such as Tim for Timothy or Bob, Bert, Rob, Bobby, Robbie, and Robby for Robert, create core data consistency errors even prior to any de-identification issues. One-time pads, encryption, and serialized numbers—the backbone of operation security—also break merging and indexing, data mining, and reporting operations too. These are either/or scenarios with limited partial controls.

[1105] The remainder of this paper defines the core objective and philosophical issues that underlie identity management and information privacy science. This includes the privacy workflow, mathematics for defining identity and privacy as a prelude to building a measurement and tracking system, informational theory for understanding identity, and one possible framework for successful automation of privacy within data processing. This corresponding sections comprise:

data ownership

definition of identity

uniqueness and sufficiency

direct and indirect identification

unique identification and subset identification

indirect identity information sources

definition of privacy

implementing privacy policy

privacy measurements

measuring uniqueness of identification

density of identification

matrix of privacy information association

[1106] The concept of privacy in the information age hampers implementation of privacy. We need an understanding of the etiology of critical and essential identifying information. We need a formal mathematical theory for identity and privacy. A number of recent legislative efforts, including the Privacy Protection Act, the EU Communications Data Privacy Directive and corresponding U.S. Safe Harbor provisions, Patriots' Act, DOT E.O. 12958, HIPAA, and GLB{A} (Gramm-Leach-Bliley Financial Modernization Act), either protect or encroach on privacy. Even legislation, such as the Family Rights and Privacy Act, better known as the Buckley Amendment, represent significant legal hazards. We need practical and constitutional limits to protect abuse. The authors of such legislation sought to address identity theft and breaches of privacy by imposing financial penalties for non-compliance. However, significant gaps in definition, logic, statistics, and implementation undermine this legislative activity. There is also lack of standing for the individual injured by an identity theft to seek damages from those who abetted that crime under current statutes and case law. The next few section outline the definitional, logical, and mathematical infrastructure for defining personal identity, and hence through inversion, the concept of privacy necessary to draft precise laws and realistic cultural and software implementations.

AA.3.4.2 Data Ownership

[1107] There is a serious legal uncertainty as to who owns data, identity data, and any information about a person. If you state categorically that the individual owns that personal information,

soon afterward many individuals would assert rights to take it back with full data erasure. Clearly that is not true because it has not happened and probably would entail an interesting legal battle. Data ownership is not clearly defined by law, statute, or court opinion. Perhaps, more to the point, is the question of who owns the right to correct data and what are the rights of ownership over flawed data?

[1108] Rather, the true issue if one of custodianship, the functional control over the collection, storage, processing, and application of privacy data. Until ownership (of identity and personal information) is defined by law, statute, and/or legal opinion, data ownership remains a gray area with enormous ramifications. For example, GLB speaks about violations of privacy on "non-public" data; this creates a legal quagmire for data exploits because the burden of proving privacy falls to the individual infringed, whereas penalties are assessed only by the Department of Justice. Before ownership or more detailed rules for custodianship are refined, secondary issues such as data value, costs, utility, access to, reuse of, misuse of, identity, and privacy remain in limbo. These are important unanswered philosophical, legal, and commercial issues. Such philosophical, legal, and commercial issues are not the substance in this paper. We need to be able to measure privacy specifically and differentials driven by events. Identity differentials are defined as any change (loss or gain) in identity through any event. Privacy differentials are defined as any change (loss or gain) in privacy through any event. Differentials (with calculus inflection points and derivatives) are important in measuring identity and privacy in practice.

AA.3.4.3 Definition of Identity

[1109] Identity is the condition of being a specified person, not to be confused with another person, or the process used to define an individual from the set of all existing and possible others. The set of all existing and possible others logically includes living, deceased, future, literary, and fictitious

persons. Limited sets create insurmountable problems because it introduces uncertainty of missing data and false matches. In functional terms, the condition of being a specified person also means the ability to identity a person, qualify the asserted identity of a person, and the ability to link all information of ownership, political, and freedoms to that particular identity and physical persona. In other words, identification requires the ability to assign or validate an identification against a physical persona, and subsequently link medical, legal, banking, property, ownership, descriptive, relationship, political, and financial records to the identity and also to the physical persona.

[1110] Note that identity is not limited to people. Identity can as easily defined and measured for any object or set of objects, ranging from chattel, to processes, to documents, to specific data elements. Certainly we can talk about the identity and authenticity of a collection King Louis 15th antiques, a Picasso painting, or war orders by George Washington. Replicas, copies, forgeries, and other representations can be thus identified and measured for authenticity. For example, it becomes technically viable and commercial valuable to assess the identity of a counterfeit $100 bill as authentic or perhaps 52% authentic, just as it is to say that we are certain to 52% that the person before us is the credit card holder. Many current identity and authentication processes currently in effect are binary, providing either a truth or fraud, and not a partial statistical assessment that is useful as the technology of identity is expanded by new developments and assaulted by increased levels of frauds and possible thefts.

[1111] An identity is provided through entity-specific information, generally respecting to a format, purpose, or structure, which is called in this paper an identity category or just an identifier. The entity-specific information corresponds to a data field, columns within a database, or tags in a hierarchical and object-oriented data structure. While identity might well represent a single concept, the identifiers used to construct identity are many and often used in conjunction with other identifiers to

construct a single identity. Names, assigned numbers (generated randomly, sequentially, or loosely), fingerprints, biometrics, and descriptions are only a few of many possible identifiers.

[1112] This paper illustrates that no single identifier is unique and sufficient. Uniqueness is a mathematical term describing the situation where one identifier is known from all others and uniquely maps one-to-one to a physical persona. Sufficiency is also a mathematical term describing the condition where identifier(s) uniquely define and describe that identifier(s) from the set of all possible others

[1113] Unfortunately, unique identity is a Platonic concept, an ideal that can never be established in certainty because of limited sets and overloaded and complicated use of identifiers. We can approach identity but not obtain it. At some point, identity assignment is held by a trust holder or by a custodian. This trust holder or custodian must apply the very same types of identity assignment to provide access to and authentication of the initial identity assignment. In other words, some identity assignment and validation must be established for the trust holder or custodian that asserts the unique identity and its validity. It is a circular logic that unravels when any trust holder or custodian is attacked. The logic undermines this circular chain. For example, the Microsoft .NET Passport and Wallet or Sun Microsystems Java-based One Platform both create a trusted custodianship for identity and personal information for which the ultimate security is controlled by these organizations applying traditional security identification and authorization. How secure and accurate are their data entry, accuracy, and security systems? Reliance on a custodian merely moves the validity issue to that other party without actually resolving it.

[1114] Some might argue that gene typing will define identity with absolute certainty, although realize that technological implementations will always link bank accounts, property, and licenses with that gene typing. I assert that even explicit gene typing requires a trust relationship to assert linkage

between the profile, the individual, and any associated logical relationships. A trusted proxy relationship is merely a practical solution to the ideal, which can never be achieved, since the gene type and the individual are indeed one and the same, but the functional representation of the gene type becomes a practical but inherently flawed identifier with trust limitations. Privacy too, is a Platonic concept achieved by association to identity. This has bearing in both legal and commercial endeavors; the definition accommodates social, religious, and superficial components. Thus, applicability becomes an issue of what is practical and what technology will work, trading on pragmatic elements of uniqueness and sufficiency.

AA.3.4.4 Uniqueness

[1115] Uniqueness defines the mathematical condition where an identifier delimits a one-to-one match with a specific individual persona. It is an abstraction that has not been achieved in any form nor with any technology yet. In point of fact, names, social security numbers, and even a representation of gene typing are not unique. In some instances, these identifiers are unique for a subset of individuals, but not sufficient to distinguish one from another within an accidental or purposely-constructed subset. An additional issue with uniqueness is non-repudiation, authentication, and confirmation, as identity theft becomes a significant economic crime (cf. "Vicomte de Bragalonne", Sommersby, and "Le Retour de Martin Guerre", and many less fictional people, such as those who were holding wills claiming inheritable rights to the intestate Howard Hughes.)

AA.3.4.5 Sufficiency

[1116] Sufficiency is the condition that several single but not necessarily unique entity-specific facts define an individual from all others through the process of minimizing the set of possible matches. This is the Venn Diagram approach to identification, a useful metaphor when I later explain the

implications of Bayesian inference on forced re-identification. Sufficiency is also the condition whereby a selection of several aggregated but not necessarily unique identifiers are necessary to define an individual from all others. The first condition is necessary to modern economic and social activity. The second condition is an unfortunate fact of data mining and aggregation technology that undermines personal privacy in modern society; it is perhaps the underlying social concern driving new privacy legislation. Bayesian inferential statistics later in the paper show that forced re-identification asymptotically approaches 100%, even with some of the least significant and seemingly most innocuous personal identifiers. Most research efforts have stressed data mining and data aggregation driven by exact identifier matches. However, a few limited efforts explore inference, proximal-match accuracy, and a preponderance of circumstantial evidence insinuating that privacy might be unobtainable, as Latanya Sweeney suggests in "Privacy and Confidentiality" and her other related papers. Expect ongoing research to explore the possibilities of and effects from unanticipated forced re-identification, how possible it is to achieve, and its prevalence or desirability. This paper shows that re-identification is a statistical certainty.

[1117] The indirect implication is that a de-identified document could be statistically sufficient to uniquely re-identify or misidentify an individual. This has importance to decision-making based on re-identified documents, decision-making which could be incorrect. This also important in terms of damage caused by immediate or other tertiary parties applying or misapplying that information. Identification must be asserted and confirmed in life-threatening situations and many commercial settings, as there is certain to be legal tort, liability, and malpractice liabilities, as well as increased in the incidence of fraud.

AA.3.4.6 Direct and Indirect Identification

[1118] Direct identifiers represent a theoretic concept but not a reality. Even within modern societies names are not unique, functional identification numbers are not completely unique even though serialized, and competing and conflicting systems create failures in distribution, control, and enforcement of uniqueness. At best identification is at best a proxy for the ideal. Specifically, some people have duplicate social security numbers; many are recycled after the death of an initial holder, an important reason for assessing data entry errors and including deceased and fictitious persons in the set of identifiable persons. Even national identity papers or cards pose serious logistical and technical problems well known within the allied technologies of biometrics. As a result, direct identification defers to a composite process applying an acceptable level of overlapping partial identifications, in a Venn Diagram style of increasing the statistical level of identification accuracy. Fraud, forgery, theft, counterfeiting, errors, and other attacks on the identification systems potentially undermine direct and indirect identification systems and create a need for better economic and technical solutions.

[1119] As a result, direct identification is defined four necessary and sufficient processes. These are:

identification

authentication

authorization

confirmation

[1120] Identification is the unique and accurate definition of an individual. The technology does not exist today. Within a technical environment, identification becomes a passive technology where the individual asserts his own designation, a designation, which might or might not be valid or validated. Authentication is the process of validating the asserted identification with variable levels of integrity and

accuracy, often referring to a trusted third party or custodian to assert that (quasi-) validity. Authorization is a functional designation of what an individual is allowed to do after identity is asserted and trust established for any authentication. Confirmation is the process of validating that bound information matches the identity assertion.

[1121] Note that authentication and confirmation are the same processes practiced in different situations or as different sequential steps in a process. Authentication is about validating the assertion of the identity of an individual. This is a serious issue in electronic commerce, particular as a means to reduce credit card fraud. Confirmation has more to do with the de-identified or re-identified information and validating its accuracy, integrity, and relevance to the authenticated individual. For example, confirmation is very important to consumers vis-à-vis false, incorrect, or mismatched credit report line items. The most disquieting aspect of de-identification and re-identification is the situation where incomplete or specious information will cause a surgeon to operate on the wrong patient, an insurance company to deny coverage, or create a situation of data suppression with indeterminate assignment of responsibility. Confirmation has been not described nor is traditionally part of the biometric literature; however, confirmation is a critical and necessary step overlooked in commercial, medical, and legal workflows as more products and services contain significant informational content with data inaccuracies and base decisions on inaccurate data.

[1122] Confirmation establishes the accurate binding of de-identified and re-identified data to the individual in order to minimize errors through the integrity and relevance of the re-asserted information. Since de-identification and other technologies, such as information dispersion and granular access control, destroy the context and relationships of identifiers within a data set or record, confirmation becomes crucial under conformance to the new privacy legislation. Note that formal

confirmation has not been part of the economic or online processes. Consumers typically react against errors only after the errors cause obvious harm, a situation that will change as information becomes more of a critical component within products and services. Generally a proxy serves instead, such as credit card possession with presentation of the credit card, or a hospital bracelet on patient's wrist. Most hospitals, banks, and financial institutions validate the identity of the individual but unfortunately make the false-logical assumption that their own internal data is correct and correctly established to that individual. Confirmation by proxy is often correct but is not logically indisputable, thereby showing the need for a more formal confirmation process.

[1123] The overwhelming social and technical problem with authorization, and hence unique and accurate direct identification, is that ultimately some authority must be trusted for that identification and/or delivery of that identification. All such systems have flaws because they are inherently circular in logic. The authority must be granted the identity information in the first place, a circular logic flaw. As a result, authentication becomes an issue of trust, a belief in a system with axiomatic flaws, or simply a practical solution with a degree of statistical failure. Although a gene type could well represent a unique personal identifier, some authority must vouch for the integrity and applicability of a gene type, thus undermining the realization of a platonic identifier, as previous stated. In security and access control, identity and authentication are subordinate to authorization, which becomes the functional issue. In effect, identity is meaningless as nothing more than a useful designation—only as secure as any other authentication based on the inherently-compromised trust-based system. The platonic ideal of identity is something we can not as yet achieve.

[1124] In effect, every identifier is indirect, in as much as each one identifies a subset of individuals. It is not exact, not precise, not infallible, but often becomes sufficient for social, medical,

economic, and government applications. As previously stated, it is a proxy for the concept. Nonetheless, identity is a commercial and practical compromise, that we generally accept, and one for which we indeed have little alternative. Indirect identifiers include any definition, description, qualification, or physical subset of a unique individual. Statistically, no identifier is guaranteed to have 100% accuracy, but individual identifiers could statistically define a unique individual with 100% accuracy. Both the inverse and converse are not true. This does not mean that an individual identifier that statistically defines an individual with 100% accuracy is 100% accurate; this is an anomaly that cannot be mathematically or legally assumed. There is also a time component to identity, identifiers, and identity information. They tend to age and atrophy in accuracy, reliability, and utility, requiring constant maintenance to retain reasonable currency.

[1125] This logical distinction is crucial when planning and implementing de-identification systems compliance with both legislation imperatives and accommodate advances in identification and data processing technologies. Neither identification nor data processing technologies will be static, so compliance only with formal legislative laws will not achieve desired results within the legal, Fair Harbor, and "reasonable standards" even today. It is not reasonable, especially not legally responsible, to adhere to a minimum standard when this paper outlines available and effective technology to resolve many of the implementational limitations of GLBA, HIPAA, Data Privacy directives, Safe Harbor rules, and other statutory requirement.

AA.3.4.7 Unique Identification and Subset Identification

[1126] Identification is not an either/or proposition. It represents a range of additive, cumulative, and overlapping individual identity properties. I assert that no single identifier exists to uniquely define an individual and link that person to all associated property and information. I cannot prove this

assertion, only disprove its likelihood. I point to a serious and significant problem that plagues our society, that of fraud, forgery, and theft of identity. These crimes would not and could not exist with unique and sufficient identification systems. Logically, one counter example is sufficient to disprove any purported axiom.

[1127] As a result, I also assert that all identifier categories represent information reflecting a subset of individuals. That subset could be a null or empty set, contain a single entry, or numerous entries. The null entry represents a critical failure in that an identifier with no corresponding identity can statistically exist, thus allowing the criminal creation of false, secondary, or replacement identities. The single entry is a special case, a statistical aberration or accident, without guarantee of accuracy or identity, as previously stated. In other words, be aware that an identity-to-individual match is not an economic, legal, medical, financial, or commercial authorization without a corresponding confirmation.

[1128] For example, the reference to "William Jefferson Clinton" may appear to be unique, but an Internet name search uncovered at least three distinct individuals, in addition to the former president. Hence, I could assign a pointless identity accuracy of 25% to that name, based on one correct match from a potential of four. "George Bush" returned too many current and time-dependent historical matches for analysis here. Even the identity subsets, "president" and "George Bush" specifies more than one individual, or multiple references in time to a particular individual. In fact, the identifier "president" or "President" is insufficient to distinguish between the presidency of the United States, the presidency of the Dallas baseball team, the presidency of a credit union, or the presidency of a rugby club (in his prep school). Results are evidently context-sensitive. I can improve the imperfect "Clinton" or "Bush" matches by additionally specifying birth date, educational history, legal status, medical history, and other categorical identifiers to create matches with smaller subset unions. Multiple identifiers create additional

Venn subsets overlapping with smaller unions, yielding with smaller confidence intervals.

[1129] This discussion has both direct and indirect implications. The most direct implication is that a document ever after a most rigorous de-identification process must logically contain information statistically sufficient to allow forced re-identification of an individual. I can improve efficacy of de-identification by removing that document from within any contextual basis. It might appear impractical or technical infeasible to link de-identified documents from the individual, but it is possible; the risk persists unless the context is broken and access to information is controlled at an infinitely granular level rather than an aggregate level.

[1130] Granular control is a core feature of software tools, such as the previously cited DigitalDoors, that provide item-wise security rather than gross access control. Granular re-identification and granular access control supersedes existing computer security methods in several ways. First, access is not either/or as with a textual or biometric password. Second, access is controlled not merely at system, volume, disk, or file level but within the file level for each de-identified item. Each item can be controlled at category, user, group, process step, or role levels. Third, each item can be accessed or denied based on time of day, internal and external threat modes, or other arbitrary policies. Fourth, access to any particular identity information can be overridden; by owner, manager, or other need-to-know ad hoc allowance granted by a supervisor, or on a per-item payment basis or other consideration. Fifth, all granular access is logged to comply with current and foreseeable mandatory audit requirements. Sixth, de-identification also works on free form textual content without the need for a database-like or XML structure, where automatic de-identification is variously supported for missing, substituted, categorization, misinformation, blanking, blurring, and other obfuscating replacements.

AA.3.4.8 Indirect Identity Information Sources

[1131] Although oral, hearsay, non-professional, journalistic, other sources are specifically beyond the scope of minimum essential information, the reasonableness legal standard still applies. Such externally derived information is fair use to achieve forced re-identification. Some identity and privacy information is very likely to be public as a matter of practice, but linkage to a person with surety is the issue. How external information applies—or will apply—is certain to be subject to legal action in order to define the boundaries of responsibility. This includes the cumulative responsibility of parties owning external databases with information that both predates and postdates the HIPAA implementation dates and deadlines for other privacy initiatives. List 2 defines external data sources that compromise de-identification and empower forced re-identification. Nevertheless, recognize that legal responsibility has not been defined or rigorously prosecuted for loss of privacy from legal use or misuse of commercial databases.

AA.3.4.9 Definition of Privacy

[1132] Before we can do anything substantive about privacy, we need a functional definition for privacy. We also need methods for measuring privacy in order to control it and measure the success for that control. The reality today is that privacy is an ephemeral assertion which we all seem to know about and know when we have it or lose it, but are unable to define, qualify, or quantify. It is unlikely we will be able to measure it, let alone control it, without a definition, some utilitarian measurement units, and a corresponding measuring process.

[1133] "Privacy, like an elephant, is more readily recognized than defined." John B. Young. Any definition of privacy depends on perspective; therefore, no complete meaning exists for all possible situations. William Prosser proposes the most encompassing legal explanation by suggesting that privacy consists of four concepts: intrusion, public disclosure, inaccurate reporting, and false identity. These are:

[1134] Intrusion refers to the invasion of an individual's solitude or private actions or thoughts. An example would be gathering information on an individual's daily schedule or reading the diary.

[1135] Public disclosure of embarrassing information that an individual wishes to remain private is another form of violation. Informing the public of a medication a person is taking that reveals an illness may be categorized under this element.

[1136] False representation is another facet of public disclosure where incorrect information falsely represents an individual. An illustration of false light is the erroneous reporting of legal action brought against an individual because of misidentification or the surgery performed on the wrong individual.

[1137] Appropriation of another individual's identity (theft of identity) is becoming a well-known problem as people seek to secure others' financial assets.

[1138] Prosser rationalizes this complex concept as being analogous to the difficulties that courts have defining and protecting privacy. Privacy can not be guaranteed because no common laws exist to protect privacy. Actually enabling privacy, subject to laws, is yet a secondary step once removed from the primary need. The idea itself is a compilation of different interests that when brought together are called privacy but otherwise lack commonality. Despite the subjectivity and vagueness of the concept, and the inability of society and the law to fully develop a functional definition, privacy might be considered a group of rights, which only prevail with infringement upon a more fundamental right.

[1139] A somewhat more practical and commercial definition for privacy is the right to be free of unnecessary public scrutiny, or the right to be let alone. Once an individual is a "public figure" or involved in newsworthy events, the right to privacy may evaporate. Hence, more than likely, I think of privacy more as the right to privacy, which is an operative issue. This right to privacy is the possible

right to be let alone, in absence of some "reasonable" public interest in an individual's activities, like those of celebrities or participants in newsworthy events. Invasion of the right to privacy can be the basis for a lawsuit for damages against the individual or entity (such as a magazine or television show) violating the right. However, the right to privacy does not extend to prohibiting persons from taking another individual's picture on the street. An invasion of privacy is the corresponding legal loss of the right to privacy. This is intrusion into the personal life of another, without just cause, which can give the person whose privacy has been invaded a right to bring a lawsuit for damages against the entity that intruded. However, public personages are not protected in most situations, since they have placed themselves already within the public eye, and their activities (even personal and sometimes intimate) are considered newsworthy, that is of legitimate public interest. An otherwise non-public individual has a right to privacy from:

intrusion on the individual's solitude or into the individual's private affairs

public disclosure of embarrassing private information

publicity which puts the individual in a false light to the public

appropriation of the individual 's name or picture for personal or commercial advantage

[1140] Lawsuits have arisen from magazine articles on obscure geniuses, use of a wife's name on a hospital insurance form to obtain insurance payment for delivery of a mistress's baby, unauthorized use of a personal photograph to advertise a photographer, and "tabloid" journalism treatment of people as freaks. There are numerous instances of governmental invasion of privacy such as the Federal Bureau of Investigation (FBI) compiling files on people considered political opponents, partially offset since the passage of the Freedom of Information Act in 1966 (FOIA).

[1141] Privacy is more the cultural issue than any right or concept that is clearly defined;

different societies define inconsistent boundaries between private and public information. Attitudes to privacy are governed by social norms. A culture tells people how to behave, and defines standards for how individuals should interact, and by implication how groups should interact with individuals.

[1142] Thus, as a society evolves, attitudes towards privacy change. Perhaps the most fundamental evolution is that of data processing during the last 20 years, rendering privileged information readily available and easily cross-referenced to create inferences and analysis.

[1143] Privacy is often regulated legislatively. There are two broad classes of legal structures: freedom of information, and privacy protection, the former to insure public access to government process on behalf of individuals (although they may not know about it), the latter to provide enforceable rules for the collection and use of information. These two classes are synergistic but often in conflict.

[1144] Privacy concerns depend upon the parties involved, and on a trust relationship. An individual will decide to provide different information to different parties (and hence preserve different levels of privacy), depending on the trust between them and on the benefits to the individual from revealing information. Privacy is often fine-grained; individuals and organizations need to choose what information they release to each other, or to business. Trust relationships can depend on proof of identity of the parties. In complex interactions, it is important that each party be able to prove the identity of the party with whom they are communicating. Privacy must be ensured in communication, as well as storage. Privacy policies could ensure that archived information is safe, but must also ensure that communicated information is safe from interception. Furthermore, the communication mechanism must be able to prove the identity of all parties, to ensure secure custodianship. As a result, privacy policies must be well known. It is important to define (and state) clear policies regarding the use of the collected information from individuals (as patients or customers). This can be a complex problem when a web site,

service provider, or business partner offers a variety of different services, when each service requires differing levels of disclosure to process the workflow.

[1145] Just as identity is a visionary notion, privacy is an illusive concept, implying different issues for different individuals. Privacy is concerned with the fundamental right of an individual to decide about the processing of personal data as well as the protection of the intimate sphere within four functional categories:

improper usage of identity information, including its deployment and its transfer to other parties other than that for which it was explicitly collected

privacy invasion, including unwanted solicitation of personal data

improper collection, retention, association, and storage of information.

[1146] There are two forms of privacy protection in the United States. The first is Constitutional. The second is protection afforded by the common, or tort, law. Although the U.S. Constitution does not expressly protect privacy, it has been inferred by many Supreme Court decisions. In addition, many States also include express provisions within their state constitutions guaranteeing an individual's right to privacy.

[1147] Common law, as manifested in the law of Torts on the other hand, deals more with invasion of privacy issues and the actionable components thereof. There are four distinct common law privacy protections that correspond to Prosser's legal privacy definitions:

appropriation (of another's likeness or identity for financial benefit)

intrusion (upon an individual's solitude)

publicizing details about an individual's private life (not of legitimate public concern)

false light, or placing the individual before the public eye in a manner that is damaging to

reputation or inaccurate

[1148] In the U.S., the challenge is about balancing the legitimate concerns of the public against the individual's legitimate right to be left alone. This leaves endless room for legislation, definitional bickering, and case law. I think a more practical definition of privacy is as the inverse of identification. In effect, identification and privacy are the two endpoints of the same range.

[1149] By taking this liberty and viewing identity and privacy as statistical constructs, privacy becomes a statistical definition rather than an absolute certainty, inheriting all the characteristics and flaws of identification technology. Also, as a result of this, it becomes easy and practical to measure identity and privacy in substantive terms, usable both as a manual or automatic computerized basis. Identity is inversely proportional to privacy, as shown:

Equation 1. Privacy = 1 / Identity

Or: Equation 2. Privacy (represented as a percentage) = 100% - Identity (a percentage)

Also: Equation 3. Identity = Loss of Privacy or Loss of Privacy = Identity

[1150] While equation 1. and equation 2. are equivalent, each circumscribes differing functional and utilitarian characteristics. Equation 1. asserts inverse relationship, that privacy and identity are not only mutually exclusive (and granular), but also inversely proportional. Granularity is important in terms of practical implementation and could have legal bearing when determining responsibility and assessing punitive damages. Equation 2. asserts that any increase in identity decreases privacy, or that a decrease in identity increases privacy, refining the distinct reciprocity between identity and privacy. Equation 3. merely reflects the relationship between privacy and identity creating a utilitarian measurement for the loss of privacy, taken for granted by individuals, but certain to be valuable in assessing legal liability.

[1151] By the way, this identity and privacy measure has some historical and literary basis; to

know the private name of an enemy—particularly with regard to dragons and wizards—is to sap that enemy's strength. Such name-based spells have a basis in Sumarian cuneiform. It is an old, certainly intriguing, if difficult, concept to defend, particularly in light of its chiefly fictional basis.

[1152] However, many fictional concepts are brought to reality through evolving technology. This one seems very relevant for modern society in light of the exponential surge in identity theft and financial fraud. Perhaps, this fictional concept of name bound to privacy holds some validity with regard to modern society as crime and war are waged both politically and economically. For example, governments limit international travel and freeze access to accounts when they have identity information.

[1153] One limitation to this mathematics is that data retained by a custodial (trusted agent or unwanted but necessary provider) database makes that data functional for commercial purposes but limits its effect on these equations because the identity is taken or granted for limited but not unrestricted applications. In effect, privately-held knowledge is supra-equational; it is non-planar. This opens doors for advanced research in areas as diverse as risk management, competitive theory, and risk-oriented actuarial mathematics. Privately-held knowledge represents uncertainties best defined within specific types of statistical mathematics, well understood and directly applicable to identity and privacy mathematics. This can be used to qualify and quantify accuracy, confidence intervals, rates for false positive and false-negative assertions, authentication, and confirmation.

[1154] Since we have to assert and prove or disprove identity, privacy represents the de facto and the standard default situation, from which we must actively assert a loss therein. In other words, the legal burden of proof is to prove identity and prove subsequent loss of privacy. This follows from both a basis in a priore knowledge, legal theory and case law, and this logical proof. The mathematics are applicable to legal, financial, and commercial applications, and thus simple to implement in information technology

(IT) systems.

AA.3.4.10 Implementing Privacy Policy

[1155] HIPAA, GLB, EU Data Directives, Safe Harbor policy, and others are legal mandates for establishing privacy policies. Unfortunately, this legislation lacks sufficient definition for privacy, privacy measurements, and measurement methods. In fact, as explained in the introduction, unique identification is truly a statistical concept.

[1156] If you were hoping for ideal answers, there clearly are none beyond the practical implementation ideas suggested herein. Legislative directives represent minimum standards for operation. You might also note phraseology such as "reasonable" or "what is reasonably determined," or even "other identifiers as determined to be necessary" in the legislative directives. In effect, this phraseology creates a moving target of conformance, a standard that will be defined by actions, court orders, case law, judgments, settlements, additional legislative activity, and conformance costs. This paper shows clearly that the minimum standard to implement privacy security is woefully inadequate. The failure to implement more rigorous de-identification, access controls, and re-identification at this early stage in the HIPAA lifecycle even now foreshadows legal liability.

AA.3.4.11 Privacy Measurements

[1157] It is not possible to judge effectiveness of any identification, de-identification, or re-identification without a measure and a system to apply that measure. Identity is a statistical concept, and by inversion, so too is privacy. Since privacy is the access to, release of, and misuse of identity information, the measurement of privacy is the measurement of the access to, release of, or misuse of identity information. Measurement is both absolute and relative (albeit against a statistical basis that precludes absolute certainties). Specifically, this means we can measure the absolute identity or privacy

of an individual against a complete measure of the individual, or the incremental or relative identity or privacy of the individual. I mention this aspect here because it portends massive legal and economic applicability for privacy non-compliance lawsuits. Specifically, consider the situation where a hospital publishes karposi carcinoma statistics with zip code information, which an insurance company subsequently cross-references with pre-HIPAA historical internal data and purchases hospital stay records against address listings in the phone book to redline specific applicants with AIDS from obtaining life insurance. Techniques of Bayesian inference make forced re-identification a statistical certainty.

[1158] While there is most certainly economic injury to these particular redlined individuals, assessing legal responsibility and financial damages against the particular hospital is complicated by asserting partial statistical blame to the hospital as opposed to the relevance of the preexisting cross-matched private and public information databases. Although the hospital could be seen as releasing the ultimate necessary and sufficient information to force re-identification of individuals, that information is not necessary and sufficient without the existing prior knowledge. At least three fundamental legal questions arise:

[1159] Is the hospital responsible for recognizing that its incremental release of knowledge damages individuals?

[1160] How much partial or full legal and economic responsibility is the hospital to bear?

[1161] How much responsibility are the other prior parties to bear for the damage to individuals?

[1162] Although the hospital was the ultimate trigger, all parties are in effect cumulative purveyors to the damage with a share in setting the threshold. As a result, the list of plaintiffs and defendants grows long very quickly with conflicting assessments for joint negligence and/or shared

culpability.

[1163] Also, consider the possibility for substitution ciphers, dependent interaction of words or phrases, application of encoded words or word meanings (as in the Navajo code talkers of WWII), or the potential for context-specific linguistic meanings that surpass the ordinary uses of common language. This is a semiotic problem that surpasses the purpose or scope of this paper. It is important to recognize this as a method to defeat de-identification methods, but not these methods when combined with dispersion and granular access control, because context and access provide another layer of security.

AA.3.4.12.Measuring Privacy

[1164] It is not possible to measure privacy directly. However, I can measure identity and infer privacy using the previously defined mathematical relationships. This is sufficient and accurate, assuming accession to the identity and privacy definitions. As such, there are four levels of possible identity measurement. The Platonic ideal is an unobtainable concept. However, we can measure practical and technical instances of identity, internal aggregates of identity, comprehensive internal and external aggregates of identity, and differentials in identity and privacy.

[1165] An instance of identity represents the potential for a single document, patient chart, report, database record, file, or other single object to assert and confirm the identity of an individual. In other words, how much information within that single chart or note can compromise the identity of the person to whom it refers. The internal aggregates of identity represent the sum total of internal information to assert and confirm the identity of an individual. In other words, how much internal information from all possible available sources, this can include private information, public presentations at grand rounds, or other private conversations can compromise the identity of the person to whom it refers.

[1166] The comprehensive internal and external aggregates of identity represent the exposure of internal information when combined, cross-matched, data mined with external public, private, commercial, and governmental sources of information to assert and confirm the identity of an individual. In other words, can internal documents, no matter how carefully de-identified be used to compromise the identity of the person to whom it refers or even a subgroup of people to whom it refers? Statistical studies become very powerful to characterize subgroups with probabilities approaching 100% to identify specific individuals.

[1167] The differentials in identity and privacy complete the concept of a measurement system, allowing full application of curve fitting, calculus, area measurements, and comparisons. This will useful in financial and legal assessments, but also furthering development in identity and privacy mathematics.

[1168] These four distinct measurements are raised because it remains unclear where the legal responsibility will be assigned for a breach of conformance. Will courts assign blame to owners of preexisting databases because they should know nonetheless the potential for harm derived from the misuse of their information? Will courts assign blame for only the final and latest release of information most proximate to the breach of conformance? Not only are these issues unclear, it will be expensive for some parties to bear the defensive costs to define the legal responsibility parameters for the rest of us. As such, the best defense is clearly foreknowledge and anticipatory response in terms of operational responses that exceed minimum de-identification and access control standards.

[1169] A security meter becomes a relatively simple concept after discounting impractical Platonic ideals. Practical implementation is feasible. Level 2, 3, and 4 are respectively measured by processing the single document for identity identifiers, the internal database (all internal record sources, both manual and computerized), and measurement of internal database along with external data sources.

Categorical identity identifiers, previously listed; would be applied. As new identifiers are discerned or developed through technology, these identifiers would be added to increase measurement accuracy and lower statistical confidence intervals. Previously measured documents, databases, and combined internal and external databases would need to be reevaluated both after new identifiers are available or when changes are made to documents, internal, or external databases. Measurement is by no means static, as demonstrated by the previously mentioned commercial DigitalDoors system.

[1170] Finally, note that the measurement is based on assertion and confirmation of identity. This is statistical in nature, reflecting both the rankings of each identification category to define an individual uniquely, and the probable inference in data mining and aggregation. We can calculate standard statistical confidence intervals within which the certainty of identification falls based on the aggregates of identification categories and risk methodologies outlined in game theory and competitive analysis.

AA.3.4.13 Measuring Uniqueness of Identification

[1171] Uniqueness represents a quality issue for authentication or confirmation. Its inverse also represents the measure of re-identification possible through internal channels alone. Identifiers that return multiple matches clearly fail at demonstrating uniqueness. However, a response in terms of a matrix, object, or set establishes a range of possible matches that can be narrowed by additional information or physical investigation. Response with a single identification does not logically demonstrate uniqueness, only the failing of the database(s) to have sufficient information or the range of information to assert a Bayesian inference. However, a useful statistic is a match inference, a compound statistic including an entry for each identifier category. It is important to note that unlike in prior unanticipated re-identification research, the emphasis herein is placed on inferential identification. Inferential identification does not guarantee exact identification per se, as you see below, but does

demonstrate the statistical probability of re-identifying de-identified data sets using Safe Harbor de-identification categories and standard commercial data attributes for individuals.

[1172] In no cases do the unanticipated re-identification rates using associative inference and set theory fall below 95%, even for remotely indicative commercial data attributes, such as magazine subscriptions, credit reports, credit cards used, and blurred service location. By the way, associative inference and (Venn) set theory enable anyone trained in this technology to recover data broken into binary categories with or without missing categories.

[1173] Associative inference is primary implemented through Bayesian methods against large databases. Bayesian inference is a branch of statistics and predictive mathematics devised by the Reverend Thomas Bayes (1702-1761) to update a prior distribution of information and obtain a posterior distribution in order to generate an uncertain quantity or reformulate a decision tree. Bayes wanted to predict the order of a card deal given cards already dealt, and perhaps use this technique to prove the existence of God. He was a theologian after all. His philosophy categorized the physical world in terms of certainty, impossibility, and a functional and useful range in between defined by probabilities.

[1174] The Bayes Theorem provides the root for inference, simulation, and many other branches of statistical mathematics and probability research. The theorem leads to very practical methods to estimate, forecast, or model the world in ways that are more accurate than finite mathematical and physical science methods. In fact, some processes are statistical and not certain, leaving only statistical but accurate methods. These practical statistical methods are used in political forecasting, pharmacological research, finance, and atomic physics. It is presently implemented with computer technology using linear programming, regression analysis, simulation modeling, neural networks, and Monte Carlo methods with Markov Chains. Bayesian inference is most effective with sufficient

information, and data collection and organization achieved the necessary critical mass during the 1990s to make it a statistical certainty that every individual within modern society can be identified with a small (3 to 4 categorical identifications) sample set.

[1175] For example, consider that a name used as an identifier returns 18 matches from a database with 67,679 entries. The simple probability of getting the matching record is a simple 1 out of 18, or 5.266%, somewhat simplistic statistics used by others studying the privacy, anonymity, and re-identification issue. In contrast, Bayesian analysis yields an inferential accuracy of only (1 - ((67690-18)/67690)) = 0.00226, that is the "knowing" that this is the correct record. Note the importance of applying the proper statistics. While the gross chance of identifying an individual from within the database is 18/67679, which is nearly 0%, the statistical accuracy of knowing for certain that this is the correct individual is nearly 100%. Additional identifiers with date of birth return 356 entries and zip code return 8756 entries. The probability composite would be (1 - ((67690-18)/67690)) * (1 - ((67690-356)/67690)) * (1 - ((67690-8756)/67690)) = 0.000000187 certainty.

[1176] Notice that even the least significant or analytically-unrelated identifiers asymptotically drive the re-identification probability to 100%. While such identifiers as gender, blood type, aggregate 3-digit zipcode zone appear to be unrelated and insufficient to identity an individual, they are logically related within set theory with inference methods as describing groups of individuals. By applying internal and external data to forced re-identification, the cohesiveness of the information is made obvious by reducing the union of data overlaps until such time as the union set contains one or a small number of individuals. In point of fact, we can find a data needle in a haystack. By combining previously-unknown search phrases constructed from trace information with external Internet, commercial, and private or secret government and military agency information sources in conjunction with Bayesian

inference, forced re-identification is virtually guaranteed. Not only that, but the supplemental information sources provide or restore context so that the identification can be linked to all real-world applications and information utility.

[1177] For example, we can expand the previously mentioned George Bush to easily construct a biographical history from grade school, through work, to political office. Commercial, private, and secret databases—depending on access to them—would allow generation of a medical, social security, driving, criminal, financial, friendships, associations, business dealings, and political profile. Accuracy is not a guarantee, given the data sources, but would be improved with cross-referencing, reduction, and some human intelligence. Certainly, the Bush profile would be useful in a wide range of commercial, financial, law enforcement, and political contexts. Although Bush is a public figure, this task would be as easily accomplished for any other person within a modern society.

[1178] Whereas most people think in terms of damage done through exact identification and the subsequent loss of privacy, the issue is not about being right with a specific match. Instead, the identification of an individual as part of a functional and operative group, and not being wrong about that assignment, is the actionable and legal issue. It is not about right, but not being wrong. This fundamental difference changes the legal onus on identification, security, and privacy issues in order to broaden the opportunity for legal liability. Since the measurement of identity is cumulative, any release of information that refines the overlapping circles of group identification thus becomes actionable.

[1179] Realize that re-identification with a set of 2, 5, or 20 potential individuals is as useful as a unique match; for redlining mortgages to cancer patients, life insurance to undiagnosed HIV carriers, or other economic poor risks. Many business decisions are not made on certainty but actuarial risk factors. While the issue of such information in such valuable applications is not against the law, the

access to such information and particularly the release of such information is rapidly becoming illegal and actionable. The implication goes far beyond privacy into other practical application of commercial data mining.

[1180] Note that even with a relatively large number of matches from within a fair-size database, the de-identification process is not reasonable until wider range of identifiers than that specified by the Safe Harbor policy are actual removed or masked. Supplement information, aggregated but seemingly unrelated and minor identifiers are extremely effective to defeat de-identification. Specifically, the Bayesian inference of unanticipated re-identification of the individual masked within the subset of eighteen is 0.99774 or 99.774%, whereas when that individual is masked within three identifier categories, the statistical chance rises asymptotically to 100%. The probability actually increases because there are more categories to identify an individual uniquely, as it should be. Misapplication of the gross chance of identifying an individual from within the database with more categories actually decreases when applying the incorrect statistics. This should be a clue to concept flaw in simplistic statistics used by others studying the privacy, anonymity, and re-identification issue, but confirmation that I have applied the correct inferential statistics above. It confirms the relative lack of safety of de-identification by identifier categories, and the need for improved de-identification methods, contextual dispersion, and granular access control.

[1181] The most efficient method to measure the uniqueness of the identification in lieu of other tabular methods to authenticate or confirm the identity or the individual is to vary the identifier categories to test for stability of the identification. Calculating this stability is supported both in calculus derivatives and confidence intervals. This statistical method is similar to measuring experimental variance. This technology will most certainly be applied to assess contributory negligence and be applied

to determine jury awards of compensatory damages. Realize that assignment of statistical values to categories and cross-category inferences is subjective at best. Perhaps there will be a mathematical means to assess this, rather than guess and empirically improve the values over time.

AA.3.4.14 Density of Identification

[1182] The uniqueness of an identifier for positive and efficient identification of an individual from all others suggests the need to rank identifiers as to their effectiveness. This density of information transcends customary assertions about whether to blindly remove names, social security numbers, and addresses as per the HIPAA guidelines and describes why the information must be removed and the potential with information overlaps and cross-matching to re-identify individuals. Density is clearly not linear as some identifiers aggregate together or with external information to create better re-identification results. Recognizing this yields the consequence that any attempt to define a "minimum necessary data set" or a reasonable best effort to achieve anonymity through legislated de-identification procedures will surely fail as legal risk management efforts, current or future. The complexity of de-identification is far greater than first obvious, and distinctly non-linear. This makes it much harder and more expensive to achieve.

[1183] While multiple identifiers create additional subsets with smaller confidence intervals, it is also possible to review the identifier categories included to resize the subsets so as to include more individuals. This becomes critical for de-identifying research data and aggregate results to minimize the potential for unwarranted re-identification of the initial individual.

[1184] Matrix of Privacy Information Association

[1185] There are some thesis and published papers asserting information density and forced re-identification based on purely practical and iterative mechanical methods. Such results speak only to the

feasibility of forced re-identification but not to its likelihood within given circumstances. However, a number of studies on data mining and data aggregation suggest a methodology for starting such research. The non-linearity and power of aggregating identifiers with public, semi-private commercial, and private databases is suggested by the grayscale combination in the cross-tabulated table or by Bayesian inference.

[1186] This falls into information theory, logic, linguistics, game theory, and associative mathematics. Bayesian inference, linear programming, regression analysis, and other statistical methods are being implemented with computers, overlapped with data mining, extraction, and other search techniques to push privacy to the brink of social extinction. For example, consider just how few identifiers and the minimum number of identifiers overlapping subsets could be and/or must be provided to uniquely identify an individual? This is akin to the problem of determining the number of unique colors needed to color a world map such that no two adjacent countries have identical colors. If nothing else, it seems I have suggested two new sciences, identity efficiency category ranking, and the determination of minimum identifier categories for efficient or exact identity assignment.

AA.3.4.13 Inference

[1187] Inferential methods are the tools of choice for breaking encryption, de-identification, and other security techniques. It is not a question of defeating them, but rather of making these techniques as inefficient and as difficult as possible while still retaining the practicality and usability of data within a collaborative environment or workflow-enabled processing environment. The point is that a one-size fits all methods to combat inference is ineffective; rather, a variety of extraction and dispersive techniques used in conjunction are necessary.

[1188] This paper has discussed Bayesian inference in terms of re-identification with partial

personal and public data sources. Bayesian inference is a specific mathematical method for backtracking missing information. Inference corresponds to and raises the potential for other types of re-identification. Specifically, inferential re-identification does refer to:

Statistical inference

Opportunistic inference

Linguistic inference

Cultural inference

Context inference

[1189] Statistical inference does include Bayesian methods, but it also includes other statistical polling and data aggregation and mining. It also includes the concrete experimental approaches although rudimentary data diving techniques as defined by Latanya Sweeney in her research. Linguistic inference is part of human intelligence methods (HUMINT) where researchers presume the types of missing information by stylistic, lingual, or comparisons against possible data matches. Cultural inference establishes situational and application link from data sources to generate context or application in order to establish potential missing information. Context inference is similar to linguistic inference because comparisons are made against similar data sources for possible matches of missing information, but context within the data source itself are used to infer the missing information. Contextual information also includes utility, application, data structures, location of data, purpose for data, time stamps, and other clues. Context also includes such clues as source and destination, the fact that the data just exists, was created, and was sent for some purpose, and can be linked to particular people, organizations, or time frames. Inference is a strong detective method for re-identifying information, whether encrypted, de-identified, limited through access, or secured by other means. Inferential methods are the tools that

enable forced re-identification.

AA.3.4.15 Conclusion

[1190] This paper defined types of identifying information, categories for the same, and how association with public, private, personal knowledge, and a circle of association or inference can enable seemingly insignificant details to force re-identification of unique individuals. This yields a definition of what is critical and important for uniquely and sufficiently identifying an individual. The paper has presented the etiology of critical and essential identifying information and a mathematical theory for identity and privacy. It provides the philosophical concepts for identity and privacy. It provides the technology for implementing identity and privacy measurement systems. In addition, it defined the requirements for de-identification and forced re-identification processes and establishes the bar for minimum implementation performance to forestall adverse legal attacks. Finally, this paper has established the outline, design, and implementational requirements for conforming to data privacy laws in an efficient, necessary, and complete process.

AA.3.5.0 Personally identifiable information: Meta PII

[1191] Personally identifiable information (PII) is any attributes, characteristics, labels, or properties that uniquely indicate an individual person. PII is generically and sometimes legally defined (under fair harbor rules) as a set of categories for information types that uniquely identify an individual. See Categories of PII. PII usually represents a unique label that maps one-for-one to each individual. Social security numbers (SSN) are the dominate PII, but other keys are created to distinguish individuals. Telephone numbers map one-to-one but sometimes many-to-few, but it is a legally defined PII category.

[1192] MetaPII is anything external to that PII set, usually descriptive, that can create PII, or through other aggregate methods, uniquely identify an individual or approximate the identification of

an individual. An approximation of an individual could include the categorization or all dyslexic 5'2" mathematicians (no names just the characterization) or the actual list of names for possible individuals represented by a set of characteristics.

[1193] MetaPII is also anything, usually descriptive, that can uniquely indicate a location, a property, a process, or any other tangible or intangible asset, as differentiated from a person. MetaPII is any information that can create aggregation, inferential, or interaction subsets that either physically or statistically confines an individual within an identifying scope or range. Consider metaPII as direct precursor to PII. MetaPII is inclusive all information lightly termed sensitive PII, non-sensitive PII, HIPAA "quasi-identifiers," public PII, and private PII.

[1194] While PII focuses on the real physical indication of a person living or dead, the information revolution has extended the legal status of the individual to organizations and corporations, while processing has expanded the model of the individual as owner and actor to process, web sites, and workflows. Individuality is embodied by many other things, including real property, systems and methods, digital resources, processes, distributed operations, virtualized operations, intangible information, and even assetized information. It is important to identify and differentiate these individuals (person, process, and property) from one another. This is only possible with formal and informal naming conventions. This maps many-to-many , many-to-many-to-one, and one-to-many. MetaPII is an intractable metacontent problem since multiple characterizations map into the same result sets and PII maps into metaPII.

[1195] In practice, metaPII can create PII, aggregate to uniquely identify an individual person, or identify exactly or approximate the identification of an individual (person, process, and property) through aggregation, inference, and interaction.

AA.3.6.0 Categories of Personally Identifiable Information (PII)

[1196] The legal definition of the Personally Identifiable Information (PII) is a function of law and regulation. The explicit categories of information considered as PII also vary by law and regulation. Note that the prohibitions and the criminal or legal consequences imposed for the release of PII remain unclear as to what that release, usage, or loss actually means. Specifically in contrast, usage and repurposing PII within the stewardship is usually valid. Usage of PII against the interests of the individuals is not prohibited. Prohibition only defines the unlawful public release of that PII. In other words, an organization with allowable access to PII can crunch that information endlessly and combine with other available sources deny claims, cancel insurance coverage, reject loans, or require stringent limitations for service, product, inclusion, or participation. Restriction on the release of PII is not the same as restriction on the usage of PII or the repurposing of PII. None of the current PII laws and regulations have recognized that aggregation, inference, or interaction create information that is fuzzy PII, which DigitalDoors calls metaPII, and is at least as revealing as any separate or combined category entry of PII.

[1197] PII is strictly defined as any an all categories of information useful to indicate or de-identify a specific individual. The definition applies to categories of information, like social security number, account numbers, or date of birth, under HIPAA Safe Harbor regulations and other legislation. PII is a strict definition that is not uniform, but rather varies by law, regulation, or proscribed operational requirements. PII is only that which is defined as PII. On the other hand, there are fuzzy definitions for PII that are misapplied and misused in discussions of privacy, security, and workflow implementation. HIPAA requires removal of eighteen direct and other "quasi-identifiers" before information is released, as listed below:

Table of direct and other "quasi-identifiers"

Name/Initials

Street address, city, county, precinct code and equivalent geocodes

All ages over 89

Telephone Numbers

Fax Numbers

Electronic Mail Address

Social Security Number

Medical Record Number

Health Plan ID Number

Account Number

Certificate/License Number

Vehicle identifiers and serial numbers, including license plate numbers

Device Identifiers and serial numbers

Web addresses (URLs)

Internet IP Addresses

Biometric identifiers, including finger and voice prints

Full face photographic images and any comparable images

Any other unique identifying number, characteristic, or code (where a code is an identifier if the person holding the coded data can re-identify the individual)

Deployed genomic privacy technologies leave DNA susceptible to re-identification(Malin, JAMIA 2005)

[1198] DNA re-identified by automated methods, such as:

1) Genotype – Phenotype Inference (Malin & Sweeney, 2000, 2002)

2) IdentiFamily:

software program that links de-identified pedigrees to named individuals

3)Uses publicly available information, such as death records, to build genealogies

[1199] Fuzzy PII also includes physical characteristics, tattoos, histories, likely locations to find an individual or relatives of the individual, or even aggregated categorical information that can be used to expose an individual. PII includes any characteristics that can uniquely identify or force the reidentification of an individual, or statistically denote groups (open or closed sets of sets) likely to include individual(s) of interest. Realize that strict PII does not reflect or legally define the results of aggregation, inference, or interaction on lesser information pooled to identify the individual. That information separately is not PII but rather pooling information useful to establish open sets likely to include individual(s) of interest. That information combined is not PII either but pooling information useful to establish open sets likely to include individual(s) of interest. PII is only that which is defined as PII. The "any and all categories" is not clear enough for legal action until such time case law or new regulation specifically asserts the inclusion of fuzzy definitions for PII. There is more than the direct method to reidentify the individual without reliance on strict PII information with this ill-defined and misused application of fuzzy PII.

[1200] Strict versus fuzzy PII definition is actually recognized by HHS HIPAA documentation, but note the legal differentiation between PII and Safe Harbor requirements. Various research papers define forced reindentification in mathematical and statistical terms based on the aggregation, inference, or interaction of information. This includes k-Map (Sweeney, 2002), and where each shared record refers

to at least k entities in the population under k-Anonymity (Sweeney, 2002), and where each shared record is equivalent to at least k-1 other records as defined by k-Unlinkability (Malin 2006) such that each shared record links to at least k identities via its trail which satisfies the k-Map protection model.

[1201] Strict and fuzzy definitions aside, there are also attempts to distinguishing between sensitive PII, non-sensitive PII, and HIPAA "quasi-identifiers." This linguistic nomenclature is rediculous? Is such a proxy a good firewall or even possible? Probably not. There is also literature on the Internet describing other fuzzy definitions that include public PII and private PII. Insurance companies and banks are issuing privacy notices under GLB informed consent requirements with a definition of nonpublic personal information as PII available internally but not available to the public. This distinguishes the value of intangible information as applied to information production and workflow where strict PII is not used or extended by fuzzy PII. In order to encapsulate the disparity between the strict definition(s) of PII and other information that can generate PII or fulfill the same function as PII to uniquely define an individual, DigitalDoors calls all this secondary information meta PII, which is what it is. Obviously, this can expose a tort liability not expressly actionable under existing PII law. Many categories of information not perceived as sensitive or legally defined as PII can be exposed through social networking, the vast networks of email users and YouTube videos, and the generic loss privacy within modern society. For an example of this PII and secondary metaPII risk, see Exploitation of Anything to Generate PII.

[1202] PII generally represent a one-for-one matchup between the surrogate PII value and the individual. Such items which are not always defined as PII but might be considered meta PII include:

Table: Possible Meta PII

National identification number

IP address (in some cases)

Vehicle registration plate number

Driver's license number

fingerprints, handwriting, or other biometric identifier

Credit card numbers

Digital identity (ID, handle, username, etc.)

Scanned image of signature

Recorded audio of name or other phrases

Bitmap image of person

Bitmap image of identifying marks, tattoos, or scars

Digital Signature

Private encryption key

voter registration

political party affiliation

URL

IP registration

Medical records

Rx records

Billing details (like Rx purchases)

Travel itineraries

Calendar entries

Alarm services

[1203] It is important to recognize that the workflow available today and the readily-available editing tools makes it trivial to create fictitious documents or access tokens without the knowledge or permission for which this PII refers. In other words, it is possible to grab a signature and face photo of a person from the Internet to create a corporate identity badge, or drop a signature into a lease document, or composite photographs either adding, changing, or deleting a person or place from the image. There are techniques to overcome these frauds, such technical analysis of image edges that would indicate drag and drop, or requiring a signature in pen in front of a witness who personally knows the signer with corroborating documents.

[1204] MetaPII generally represent a many-for-one matchup, one-for-many matchup, or a many-for-many matchup between the value and the individual. Information that is not generally considered personally identifiable because many people share the same trait include:

Table: of Information that is not generally considered personally identifiable

First or last name, if common

Country, state, or city of residence

Age, especially if non-specific (anonymized by range or by omission of birthdate)

User name or handle (such as GoBoast@Facebook.com)

Gender or race

Name of the school attended

workplace

Grades, salary, or job position

Criminal record

Voter registration

Library card

Magazine subscriptions

Other subscriptions

Service billing information (phone, electric, cable, satellite, Internet)

Common service or product registration

Mailing inclusions (think Lists USA subject-matter lists)

Consumer click-stream data

Consumer email-open data

Cookies

Registry entry

File (cookie)

NAC settings

System settings

Screen saver settings

IP address or URL

Address lists (in Email) because they establish hierarchical, group, or role relationships

Social tags (establish non-regulated or uncontrolled contextual relationships)

Process logs

Timestamps

Phrases

Quotes

Common sayings

Unusual turn of words

Consumer click-stream data;

Consumer email-open data

Aggregate click-stream data

Alternate representations in other fonts, transformations, formats, syntaxes, and punctuation include:

Table: Alternate representations

Zip+4 or zip+4+2 is exact designator of position

Zip as barcode is also a PII

DEA number

Medicare or Medicaid number

Military ID number

Participation number

Registration number

[1205] A username or handle used as an ID at one web site can also be looked up at other websites, as the example detailed in Exploitation of Anything to Generate PII unfortunately demonstrates. Although these different web sites often protect user information robustly given the recent high-profile web site breaches, the aggregation of these like names and the related site information can also be used to paint a useful profile of an individual. Additionally, if access to ID profile information is publicly available, or you own a website and have access to ID profile information, you can link and match the PII details. So the use of the same ID and password on multiple sites might simplify the ID holder's life it creates an open security hole that might link into actual PII or cached credit card

information. Granular content control and masking (particular is uniform among sites) adds protection to the ID.

[1206] Also, information from web videos or web sites can be used to reidentify an individual. While such information is not specifically covered by the formal umbrella of PII, it most certainly is PII in the truest sense of the definition since it can be used to explicitly match an individual, as previously detailed in Exploitation of Anything to Generate PII.

AA.3.6.1 Technical Identifiers

[1207] Personal information is a vestige of agricultural and industrial revolutions. The information revolution also distinguishes between the physical person and a logical persona as well as more virtual individuals performing actions and possessing assets. Persons are differentiable from persona or process. Specifically, software application and workflow are owned by persons and sometimes only by a persona in no way tied to a person but to some other logical identity, like a corporation, role, user id, a process id, or login entity. These personas and identities can own process, assets, information, and other intangibles. These technical identifiers can be used to link assets to individual persons and expose PII through a many-for-one matchup, one-for-many matchup, or a many-for-many matchup. Some identifiers include:

Table: Technical identifiers

Vehicle ID

Real Estate Parcel ID

Check Routing Number

ABA routing number

Various service account numbers

Proxy Identifiers- pool size

[1208] A more technical set of identifiers links computer hardware to a person or persona. This includes:

Table: Identifiers linked to computers

the serial number of hard drive 2b

the bios version of motherboard

the bios manufacturer of motherboard 20

the bios element id

the windows cd key stored in registry

the windows product id stored in registry 40

the windows nt digital product id stored in registry

[1209] Some of the relevant PII laws and regulations are listed below:

Table: PII laws and regulations

California has privacy written into the state constitution Article 1, Section 1.

Online Privacy Protection Act (OPPA) of 2003

Federal Laws:

Privacy Act of 2005

Information Protection and Security Act

Identity Theft Prevention Act of 2005

Online Privacy Protection Act of 2005

Consumer Privacy Protection Act of 2005

Anti-phishing Act of 2005

Social Security Number Protection Act of 2005

Wireless 411 Privacy Act

US 'Safe Harbor' Rules (EU Harmonisation)

Title 18 of the United States Code, section 1028d(7)

European Union (member states)

Article 8 of the European Convention on Human Rights

Directive 95/46/EC (Data Protection Directive)

Directive 2002/58/EC (the E-Privacy Directive)

Directive 2006/24/EC Article 5 (The Data Retention Directive)

Further examples can be found on the EU privacy website.

United Kingdom & Ireland

The UK Data Protection Act 1998

The Irish Data Protection Acts 1998 and 2003

Article 8 of the European Convention on Human Rights

The UK Regulation of Investigatory Powers Act 2000

Relevant Case Law

Employers' Data Protection Code of Practice

Model Contracts for Data Exports

The necessary content of Privacy Policies

The Privacy and Electronic Communications (EC Directive) Regulations 2003

The UK Interception of Communications (Lawful Business Practice) Regulations 2000

AA.3.7.0 Toxic spill of personally identifiable information (PII)

[1210] When personally identifiable information (PII) spills, the problems are potentially magnitudes greater than simple PII loss. PII loss includes misplacement, damage, inability to access and use, destruction, or erasure. Spills include information leaked, reused against authorization, reused violating published acquisition parameters, duplicated, or stolen with loss. Legislation and compliance laws do little to prevent loss and spills but say a lot about assessing the ultimate toxicity of these events

[1211] PII loss is an internal problem. PII spills are an external problem. Internal problems incur business expenses but external problems involve all that and more. Externals problems entail compliance, legal, remediation, and reparations. Data spills can be toxic and often are.

[1212] There is also the question of ownership or custodianship. When the information belongs solely to the organization, any damage through spill or loss goes against the organization. When the information represents other people, other organizations, others' trade secrets, or is PII, then the damage might go against your own organization and there is also likely to be a spillover of effects on these others that will likely come back as a responsibility to your own organization too. The legality of ownership of PII is not clear so the responsibility is often dropped.

[1213] The thought of a data spill -- bits and characters all over the place -- seems like a unreasonable joke at first. However, with the thought that all that data represents business methods, trade secrets, intangible property, and PII, that this data represents some 95% of the value of a typical organization, the joke ends. A data spill is not a joke. It is perhaps the most serious complication in business today, and it really is toxic.

[1214] The PII toxicity is a measure of demonstrable spillage and the actual damage or potential for damage. Other DigitalDoors whitepapers blur the line for exposure because aggregation, inference, and interaction expands the damage caused by all types and categories of PII and even other categories of information that do not specifically indentify or force reidentifiation directly. Statistical and inferential methods are very powerful. As such, PII should be presumed to be highly toxic. Also consider its half-life decay.

[1215] There is the issue of scope. How much data about a particular individual was spilled? How many individuals were compromised? How pertinent is the information? Was it social security numbers or only internal account numbers? How useful is this information for compromising the source of the PII or assets associated with PII? Could this PII be used in other unconnected activities to create fraudulent checks, credit card accounts, or house loans? Could this information be extended -- like a stolen email list -- to contact related people using the credentials from the exposed individuals?

[1216] Context adds much the concept of the information. It is not only about the amount or density of the information, the category of information, it is also about the where it came from and how it is used. Subtleties have profound impact. Consider a list of library loans. Consider now a list of videos borrowed or bought from the movie store. Consider next a list of videos bought from the adult movie store. Each source raises the credentials about the accuracy of the info and its specificity.

[1217] Instead, consider medical information from an Internet survey. Consider now medical information from an employer's files. Consider next medical information from a cancer specialist. Each source raises the credentials about the accuracy of the info and its specificity.

[1218] Data spills are specifically not about the loss of a business asset but rather about the contamination peripheral people and resources and organizations. Shareholder lawsuits against

executives defocus business and chill people. The legal and criminal ramifications could dwarf the original economic benefits, much like the uses of asbestos, PCB, and radioactive enrichment have created toxic sites the world over.

AA.3.8.0 From aggregation, inference, reidentification to Anonymity

[1219] Medical studies published in journals often contain aggregations of participating patient information. The unspoken presumption or even specific designation in the participant agreements is that compiled statistics protect the privacy of individual patients. While this is true generally, it is not true specifically. As a result, anonymity fails to provide privacy protection, both specifically to these medical studies, and generally to online Internet activity and integrated workflows. The implications are far darker because the reidentification of the participants is very likely. Although the information as published does not undermine the privacy of these patients, other available public or even private information within the insurance network, hospital systems, or physician practice can be used to statistically reindentify individuals and exploit personally identifiable information (PII). The reindentification process advances from this presumptive anonymity to specific reidentified patient names. The technology of inference, both human intelligence and statistical, are powerful tools used to expose the privacy of patients, in spite of the protections taken, the redaction of sensitive information, the security of patient records, other fair harbor efforts under HIPAA, and even more presumably stringent containment under the study procedures.

[1220] A journal article exposes a medical specialty, a medical practice, the umbrella organizations funding the activities, the physicians and other named facilitators to the study, and a specific range of medical cases. Also, this other information can be inferred specifically if not generally from the study details: the patient cachement area, the age range, the patient race and cultural

backgrounds, and the outcomes. If the study includes drug trials, this useful information can be matched against chain-pharmacy databases. Pharmacy records are available for payment, incentive, and performance monitoring. Even body part photographs are identifiable through social networks, as in Exploitation of Anything to Generate PII.

[1221] Information can be bought, exposed, or inferred. Specifically, access to patient lists, hospital service lists by patient, any information matching patients to conditions, diagnoses, patients to service dates, patient admission information, patient prescriptions, patient demographics, and per patient PII is the fodder for reindentification. Many studies provide patients with services, freebees, and even payments. Accounting records for these activities represent PII when organized by patient, but when organized by study, by expense accounts, or other normally obtainable business aggregations are not perceived at risk. They are often exposed in audits, FCC filings, and other documentation that is published, subject to less stringent controls, or available through extreme measures. Aggregation of disparate sources compromises PII.

[1222] Population PII is available from many sources. Of course, this is a huge set, but techniques of inference winnow out the chaff to expose exacting or probable sets of participants. In many cases, as explained in other papers, the statistics of inclusion versus exclusion are merely a monetary issue. Percentage risk factors can be applied for business activities, such as redlining an individual as a medically bad loan risk unless the loan is more expensively collaterized by assets or loan insurance. This is one of the reasons why genetic databanks and medical expense profiling is a hot issue (see the Genetic Information Nondiscrimination Act of 2008) because it creates more risk pools to the detriment of many individuals.

[1223] The inferential information from the actual journal article, lists of patients, doctors,

services, PII databanks, and other sources aggregates into a master PII list. Broad swaths of people are excluded by demographic non-matches. Each attribute is described and used to exclude more and more people. While a study is likely to represent tens to hundreds of patients, the set exclusions could be as large as thousands. That might be sufficient to expose study participants. Additional lists, pretext calls to providers, or even direct calls to participating patients are subtle ways to validate or deny study participation. Legality might not be an issue at all, as offshore service brokers shield liability or are not practical to prosecute. Furthermore, use of PII by an insurance company or other owner does not violate HIPAA and other compliance regulations which do not prohibit the reuse of this information, only the exposure of it to undermine privacy. Organizations seeking the participant PII, the unrelenting effort often pinpoint exactly the original patients to their own detriment and exposure. Aggregation is no shield for anonymity.

AA.39.0 Control personally identifiable information (PII) through granular data control

[1224] User identifiers -- indexes or keys -- get the bad rap of being personally identifiable information (PII) that exposes security and privacy risks. This is not precisely true. It is the legacy that creates this risk through a generation of mediocre security and the lack of understanding the ramifications of data capture, distribution, and sharing. There are ways to create indexes that are not PII, and perhaps even universal identifiers that are not PII if referential information and other PII are protected through granular protection.

[1225] User identifiers are common indexes or keys used to find a specific record or subset of records. Common identifiers include name, addresses, telephone number, social security number (SSN), account numbers, and other values. User identifiers are not personally identifiable information that breaches privacy per se because they often lack uniqueness, although an address or telephone number

might represent an immediate security risk because of its specificity. External facilities to lookup or perform a reverse lookup to match these index or key values to PII enhances immediate risk because the measure of uniqueness is increased. Most identifiers are quasi–random numbers or non-unique values. Because these do not uniquely identify an individual, they do not create as much risk as more precise and unique ones, like SSN, which are normally a one-for-one match to a unique individual. Numbers map to metadata, and it is this metadata that exposes actual PII. However, names are statistically personally identifiable even for large sets of surnames, like Ng, Smith, or Wu, because there are ways to increase the measure of uniqueness even for large sets.

[1226] However, in and of itself, a SSN and other similar serialized or generated identifiers really contain no personally identifiable information. They are content and context neutral. The information that they reference or link into actually provides the security risk. This metadata creates the risk. The identifier itself is not the risk, only the linked data represents risks. Widespread use of SSN for unanticipated uses has created easily accessible references, and pragmatically exposes SSN as PII, but only because of history.

[1227] In order words, if SSN could not be used as a search entry point for personal information about a specific person, it would not be a risk factor. It is only when the referential information is readily available and not locked down that the identifier compromises security and privacy. If all the public databases lacked references to SSN and all the private databases were secured properly, SSN would not represent a risk. However, given its legacy and the widespread misuse, it is now and will likely remain a damaged index value. Account numbers for bank accounts, brokerage accounts, credit cards, debit cards, and gift cards at present do not represent a PII risk. That would change if readily-accessible references cross-match accounts and people.

[1228] If user identifiers are generated mathematically from damaged identifier values, like SSN, they can be regenerated easily from known SSNs, and these new identifiers are damaged at first use. The ASTM (formerly known as the American Society for Testing and Materials) UHID (universal healthcare ID) is just such an example of a flawed privacy identifier because it is a transformation of SSN. Indexes generated as hashes from names, masked SSN and zipcode, or other exacting values are likewise flawed. Any surrogate keys present complications as they represent a transformation from existing process and eventually these surrogates become the same PII compromise for which the surrogates were designed to avoid.

[1229] If user identifiers are generated from unique biometric factors, it should be assumed that the user identifier can be reversed to expose the PII. Even salted generation does not guarantee uniqueness or protection against reverse engineering the index. The value of generation from unique biometric factors is perceived as means to prevent repudiation because the person and index are uniquely bound together. However, such schemes saddle this identifier literally with a built-in PII limitation. DNA -- when it becomes fast and cheap to map -- presents most of the same problems of user identifiers and PII but is also subject to data diddling. DNA is flawed surrogate identifier.

[1230] The simplest way to create a PII-safe index is to generate a unique random key for assignment as a system-specific index and relate this to the system-specific PII. Serialized keys are not random. System-specific PII must be secured by encryption or other comparable security methods. Any association between the index and the related information must be rigorously protected. In other words, the index is only an index and cannot assume content or context value. It is best if it is not displayed. This is a significant limitation because it means that account numbers, user health identifiers, and the like can be used for lookup, quick reference, or substitution in whole or part of the other PII.

Alternatively, if the other PII is controlled through granular content control, then the identifier can be used as a normal index because the linkage between the index and the data is protected.

AA.4.0.0 Classification and categorization:

AA.4.1.0 Multiple Hierarchical Tagging and Data Extraction

[1231] Data security has been traditionally based on the metaphor of perimeter-based defenses. This is a historical fact. Castle and moats, locked strong rooms, vanadium steel vaults, and mathematical encryption have protected money, assets, and information well until this age of networks, the Internet, and distributed processes. However, as identity theft, fraud, and other swindles have struck the bottom line such protection at the edge is no longer sufficient. The paradigm shift herein is that data security must be internal in order to travel with the data throughout all distributed processes even over networks to thwart data reuse, data repurposing, and replication, and reformation. DigitalDoors enables this dispersive subtractive technology in order to enable granularized content security.

[1232] One of the extraordinary data security risks is that data collected for a specific purpose today is distributed as a report and then recycled for new purposes tomorrow, purposes that are unanticipated and even contrary with the original data collection and edge security. Formally, this risk is defined by how iidentification is used to repurpose data resources in ways different from or incompatible to prior representations, thus requiring flexible and overlapping identification systems to reflect that unlimited risk.

[1233] Edge security, such as encryption, fails when this data is released in clear form or when the encrypted data is released with the keys as part of the distributed operation. Data is often decrypted, consolidated, and reported in transformational formats for distribution regardless of ownership, stewardship, or the long view security. Clear form is necessary for the purpose of reporting but is now

exploitable by others; distribution of encrypted data (complete or even partial) for processing requires deliver of the keys, but the resulting decrypted clear form is also exploitable by others.

[1234] Specifically to combat such security realizations -- it is really no longer a credible risk but rather an actualization or potentialization -- DigitalDoors recognizes the content of items. DigitalDoors classifies those items in terms of absolute security risk directly or indirectly through aggregation and inference. DigitalDoors assigns multiple tiers of in-channel or out-of-channel tags describing the content and purpose of the classified items. DigitalDoors extracts those items too risky to travel within the data stream. Finally, DigitalDoors reconstitutes the data stream, in whole or in part, as necessary, to enable data processing subject to a defenses-in-depth identification and authorization.

[1235] So that we can explain why this subtractive internalized security works, we present you with a deceptively simple sample that has little content but packs a lot of context and concept into this two-sentence paragraph. The meaning and importance of this sample is contingent on external data aggregation, inference, or á prioré knowledge. In other words, this is a sample with aggregation or inference exploitation value through actualization, realization, and potentialization. What might appear superfluous is realistically exploitable by others. The sample:

[1236] Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight

[1237] This sample demonstrates the functional difficulty in proper tagging because it contains names with both clean and dirty words within 1) a defined content, but 2) a mixed context, 3) an ambiguous concept, and 4) an unknown applicability. Tagging, the assignment of contextual information tags to a defined information item, is applied to ascertain what information items should be extracted

from the information stream in order to apply security or protect privacy. Tagging is neither obviously definitive nor linear. Recognition of items alone does not convey the necessity of redaction (extraction of information items) or assignment of a classification. A single tagging, such as the sensitivity level [confidential], is insufficient because it is lacks directives for its confidentiality with functional security reference as to what group of people and for what purpose, and for how long. Also, the very inclusion of the tag as an annotation alters the sensitivity of the sample itself because it adds more information. A tag or multitier tags are in fact metatags, a subset of metadata, and metadata is well-known for in-channel and out-of-channel security leakage. A sufficient sensitivity tagging contains the out-of-channel sensitivity level, compartment, category of item type; these are additionally extended by metatagging with user, usage, mission requirements, time-sensitivity, and other dimensions of information, as well as distributed in unlimited multiples to constrain the unwarranted inclusion of contextual annotations.

[1238] Clearly, "Robert Jones" and "Jane Doe" are categorically names and need to be tagged—but at what sensitivity level? Should they be unremarkable, [top secret], or [sensitive but unclassified]? It is not obvious to tag them with multiple levels of sensitivity, as DigitalDoors does, because the functions for discerning distribution and purpose are not in place for most user requirements and are alien to the existing computerized markup technology and presumption of the single assignment. The applicability of free-form documents with database field assignment and application of non-database hierarchical metatagging is neither obvious nor well-understood in the current state of the art. We are applying data analysis technology to a security containment operation, a new concept both for data and for security. You might ask what data analysis and semiotic meaning has in common with security; the answer is everything for subtractive internalized security where content, context, concept, and applicability define the actualization, realization, and potentialization for the data.

[1239] The sample contains contraction titles of "Mr." and "Ms." but could easily contain "Captain" or "General" alternately indicating a military context. "Mr." and "Ms." could establish a military content, but not necessarily or evidently so. The sample also contains "ballistic missile submarine" although dispersed throughout the single sample sentence in the correct and usual order of the words when describing [top secret] military equipment. However, the contextual usage is clearly different, as DigitalDoors discovers. The content is clear but ambiguous. The context is not about activity on the seas but conference room antics. The concept is not about military-sensitive equipment but rather boorish interpersonal behavior.

[1240] First, DigitalDoors tags the names categorically as names. Second, DigitalDoors assesses "Mr." and "Ms." categorically as titles with EEOC-contextual meaning, that is, specifically in this case a woman physically assaulted by a man. Third, DigitalDoors assesses those dirty words separately or the dirty word phrase "ballistic missile submarine" occurs within the sample document, but more specifically within a single sentence. Fourth, DigitalDoors assesses that these dirty words are non-contiguous and unrelated. The differentiation between document, paragraph, and sentence range, and the aspect of contiguity is important for tagging applicability. There is no obvious or overt military or agency context. In addition, the statistical analysis of "Mr." versus "Captain" or even case-specific "captain" within the contextual sample also is important for applicability. In some cases, the statistical preponderance of clean versus dirty words changes the sensitivity level (and other multiple hierarchy tagging instances).

[1241] The two sentences in the sample are disjoint. The content and context are at odds and seem unrelated. That should really be a distinct clue of exploitable external information aggregation and semiotic value. Language and dialect also pertain to the assignment of tags because specific words and

phrases often convey context- or concept-specific meaning that alters the applicability of the information. Specifically, the included cliché "it's not this dog's fight" contains no dirty words but likely establishes a indirect reference to West Point Military Academy because the speaker was likely trained there or associated extensively with someone trained there. This is a common phrase in WPMA classroom case-study methods. This specific external reference is important for information aggregation and retagging, as discussed later in this paper. Nevertheless, the analytical reader can see the semiotic value and, as is said, connect the dots.

[1242] Sensitivity-level classification in neither simple nor linear. The limitations of current DoD specification presume the simplicity and linearity of tagging. As such, the stretch to non-linear tagging is not obvious since it breaks with cultural conventions and existing technical conventions. It is technically complex with no direct transition to implement because the implementation combines unconstrained information formats with technology that is most often constrained by single field assignments with the expanded requirement to apply an unlimited hierarchy. Support for unlimited tags and unlimited dimensions of tags is a shift from current data tagging concepts and application of most data processing technologies where assignment is traditionally one item from an available lookup list of items. Additionally, new legislation by way of executive orders and congressional bills have altered the landscape for information sharing and agency security by creating conflicting tension of what must be shared and what must not be divulged. Current specification defends the tagging like this:

[1243] <TS>Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1244] However, new requirements drive the tagging in terms of sensitivity level and

compartments to be something like this, a format not even obvious in current implementations:

[1245] <SBU NATO, TS PERSONNEL, U FBI, STONEGHOST ALL>Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1246] This demonstrates two-dimensional tagging at best, although the two-dimensional-format tag is repeated in four instances. A fundamental paradigm shift is required to redress mixed context and applicability. First, DigitalDoors tags are multidimensional and hierarchical, potentially without limit. Second, distribution of the sample with tagging annotation is specific and unlimited. In other words, the sample could be tagged like this:

[1247] <SBU NATO until 2006-03-22 relating to Project 13467 re: cancellation, TS PERSONNEL for all records relating to Robert Jones only, U FBI for criminal arrests reports, STONEGHOST ALL irrelevant>Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1248] In addition, distribution of the tagging sample is potentially without limit, and the inclusion of the complete hierarchical tagging is inadvisable since it contains contextually unwarranted annotations. Therefore, DigitalDoors controls the distribution to dispersed storage with a controlled redaction and tagging annotation. Five examples from the unlimited tearline possibilities include:

[1249] <SBU NATO until 2006-03-22 relating to Project 13467 re: cancellation>Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1250] <TS PERSONNEL for all records relating to Robert Jones only>Mr. Robert Jones went

ballistic, launching the paper stapler as a missile for a direct hit, because Ms. Jane Doe planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1251] <U FBI for criminal arrests reports, STONEGHOST ALL irrelevant>Mr. Robert Jones went ballistic, launching the paper stapler as a missile for a direct hit, because <removed> planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1252] <STONEGHOST ALL irrelevant> I did not involve myself since it's not this dog's fight. [redacted name] went ballistic, launching the paper stapler as a missile for a direct hit, because [redacted name] planned to submarine his proposal. I did not involve myself since it's not this dog's fight.

[1253] Notice that the final example contains no imbedded tagging annotation to minimize source and method information. In addition, tagging and redaction does not address the issues of data mining, information aggregation, and inference when documents or single informational items are combined together from multiple sources. Combinations from within even single sensitivity levels can alter (increase) the resulting sensitivity level. A sustained multiple hierarchy of tags would allow for future reclassification when information is combined. The notional shift from fixed single sensitivity level tags or sensitivity level/compartment tags to on-the-fly reassessment and retagging is not obvious; it presumes inter-disciplinary knowledge -- of semiotics, of information theory of statistics and combinatorics, of Bayesian inference, of DoD policy. This notional shift also requires an action plan encompassing the willingness to break the existing cultural methods of information sharing, ingrained cultural attitudes toward intramural agency security, perimeter-based defenses for information security, and privacy enforced by third-party custody.

[1254] For example, even when two intramural documents are combined -- herein when the first sentence in the original sample is combined with comment in the form of the West Point cliché -- the

context is no longer just any conference room but specifically a military one. You can correctly conclude that two [unclassified] items potentially infer a [sensitive but unclassified], [secret], or even higher combination. DigitalDoors retains the multiple hierarchical tagging for reassessment and reapplication of the tags creating new scenarios on-the-fly.

[1255] Finally, in addition, the assignment of tagging is constrained by applicability. Although it is conceivable to generate many or even all of the possible permutations of these multiple hierarchical tags, multiple redaction, and multiple storage dispersal, information overload and effective utilization encourage minimal distribution. In addition, this technology retains the ability to call-back information and put the cat back into the bag, so to speak, by applying DigitalDoors reconstitution to minimize the reuse, repurposing, and redistribution of data stores of security-sensitive or privacy-privileged personal data. Whereas, one tag in one dimension does not meet requirements even today nor with a full dump of tags within a multiple hierarchy either, DigitalDoors creates subsets on the fly as dictated by configuration and needs subject to content, contexts, concepts, and purpose. This tagging methodology is based on a multitier hierarchical overlapping tag structure. Formally, as described, this threat is defined by how iidentification is used to repurpose resources in ways different from or incompatible to prior representations, thus requiring flexible and overlapping identification systems. This shifts the traditional security from the broken perimeter metaphor to the subtractive internalized security paradigm.

[1256] AA 4.2.0 Commonality of Data: categorization, a multitier hierarchical overlapping tag structure.

[1257] The requirements within an ecosystem to search, categorize, security, and share data have different requirements to be sure, but all have a common base. The base is categorization, preferably implemented as a multitier hierarchical overlapping tag structure. The reason for this commonality is

that search, security, sharing all require an assessment of the meaning of the data, best performed with a categorization of obvious, potential, and hidden value.

[1258] In terms of search, you want to establish a context for the information you want, and the context is usually ad hoc when Internet search engines are used. However, more effective search tools establish parameters of interest in the form of categories to isolate and focus the search. In some cases, you want to establish a search specifically for content that is new and does not match existing content to widen the knowledge base. In other situations, search overlaps existing content in order to strengthen the deep and detail of the information are assert integrity, accuracy, and veracity. Success with search to increase width or strength depth is driven by understanding of the taxonomy of relevant information.

[1259] Effective security is no longer possible unless the data is fully understood. This includes content, context, and concept so as to anticipate the potential to aggregate, infer, and interact with the data when security fails to contain it. It is not enough to identify sensitive selections and personally identifiable information (PII), segregate specifically that, and preclude access to the rest of the data columns, specific rows, or documents and data streams of apparently lesser value. This all or nothing security precludes operational workflow, the reason in the first place that the data was acquired, and to a fair degree the security is good but you cannot do anything. This all or nothing security misses the value of data with apparently lesser value when it is pooled with other like sources or public databases. It often is not so depersonalized. Separation of only safe-harbor categories of data does not prevent a statistical re-assemblage of the sensitive components or PII. Just because a designation of blood group, physician, and illness seems generic or aggregated beyond useful recognition does not preclude the likely inferential re-identification through other data sources. Interaction can expose sensitive components or PII even when it is carefully released into a secure data processing operation because reports are

generated that can re-include what should be secured or be passed through waterfall of processes that leak results outside a contained environment. Another risk of interaction occurs with distributed processes, SaaS, and SOA because the new concept engenders a continuous processing cycle where data literally might never be at rest; the PII, even in small pieces at any instance, still aggregates with the potential for leakage over longer periods of exposure.

[1260] The DigitalDoors paradigm addresses all aspects of data categorization, search, security, and sharing within a commonality of granular data control. While these different functions have different requirements, they nonetheless share a common basis.

AA.4.3.0 Categorization structure tagging, risk measurement and context

[1261] The measurement of risk begins with assessment of the value of information, both structured, unstructured, anything in between, and all explicitly or implicitly tagged data. That notion that structured data is implicitly tagged with the inherent schema and that tagged data is structured by the tagging is only partially true. Databases are structured by tables, rows (indexes), relationships, and fields but only "tagged" specifically for an express purpose. The applied data structure is but one possibility of many that imparts a concept and context to the data; it is easily transformed for reuse and repurpose. Although database design is often generic enough to make data mining possible, the veridical value of the database will not be realized until the aggregation, inference, and interaction is fully explored. Likewise, the veridical value of the unstructured document will not be realized without tagging to explore the obvious, hidden, and potential value within the information or transformation into a structured format or processing within any application that matches to unstated data structure.

[1262] Structure and tagging are not synonyms, functional equivalencies, nor analogs. Structure provides accessibility within the specific construct of a purpose or process. Tagging provides content-

related and contextual meaning as well as the possibility for accessibility, as with XML. Structure does not replace tagging, nor does tagging replace structure. They coexist independently for different reasons. Structure is likely just the application of a single instance of categorization.

[1263] The dependencies and interdependencies, relationships (known and hidden), between data and the structure appear only within a context and concept. Meaning is contextual at all times, and although we try to categorize data in databases, mark content with XML, or create complex hierarchies with tags, each view is just that, a view among many possible. Consider for example the carton of eggs seen in Security for the Information Matrix wherein we showed that the information container characterizes the content even when separated. However, context and concept play a role as well.

[1264] As a result, there is a problem with the definitions of structured and unstructured documents. See Degrees of Structure for more information. Premise of structure in field or metadata tags define range of data elements and provide directed input in a defined process. Such a structure provides flexibility for redirecting the input into another compatible process. Clearly, structure is purpose-driven and defined vis-à-vis the process. There can be other purposes, those known by you, antithetical to you, and unknown to you.

[1265] Nevertheless, that does not mean that your structure is wrong, just incomplete in that other structures can augment the preexisting structure. Unstructured documents can also be augmented by structures. Structures can overlap or coexist without limit. Furthermore, software, style sheets, queries convert between structured and tagged and visa-versa with a mapping. Although a database structure is limiting and inherently vulnerable, it hints at aggregation, inference, and interaction. Structure and flexible tagging without limits drive the categorical imperative to explore the full meaning, potential, risk, and utility through information sharing.

AA.4.4.0 Data and Metadata Granularity for decomposition to break context

[1266] Metadata contains more risk than plain text or simple content alone because the metadata adds attributes to the basic data and furthermore metadata can include all manner of things that are not pertinent to that basic data. Although it is generally true that data with imbedded or linked metadata represents a greater likelihood of security risk, it is also important to realize that no simple data item, data file, or even a flat data set is independent from implicit metadata.

[1267] Data about data -- the definition of metadata -- is inherent in every reality of data. Data does not exist in a vacuum. Each data item is intended for some purpose, and that purpose adds to the conceptual meaning of the data. A flat file has a binary encoding, a language designation, and perhaps even an application this is automatically invoked to open that file. That is metadata, whether incorporated into the file itself, vaguely understood to map to the file by convention or configuration, or whether explicit system settings and processes presume and apply that metadata as a fact of operations. Therefore, the absence of explicit metadata does not preclude the existence of implicit metadata. Implicit and explicit metadata both imparts security risk. It does not matter if you recognize the implicit metadata or not; it is still a risk and still a responsibility to recognize and mitigate this risk.

[1268] Metadata is not a file storage format, as most people think when they classify XML as metadata. XML is really a protocol expressed within a file format that most often contains metadata, but is not metadata specifically. Metadata is useful as a structure for the simplification of process, the codification of information relationships, and a mixed packaging system that combines content, context, and concept. A database is a metadata structure just like any MS Office document or XML package. Metadata does not per se include or reference a structural template, a style sheet, or a transformation, but the legitimate reality is that metadata is at risk because of that implicit or explicit linkage. Explicit

references impart risk, but so implicit ones too, and the implicit are somewhat more dangerous because they are so often missed.

[1269] A template, a macro, or a transformation that acts on a metadata file adds risk because it binds metadata to data. Even the MS Word normal.dot contains toolbar buttons, add-in code, macros, and implicit formats that alter the concept of any seemingly innocuous document. Template changes with the application of an alternative .dot template file, an XML style sheet, or dynamic .css tools that reflow the presentation of the content do indeed alter meaning. As funny as Punctuation: Context alters Concept is, realize that presentation change can be manipulated by document templates and style sheets resulting in even subtle alterations that dramatically effect meaning. When metadata results in more profound workflow variations, the output and thus the risk of information alone as part of a process or product or service output is even greater. The transformation is not separate from the content, but it needs to be separated from it in any effort to secure content. This explains the risk factors from data format, structure, and other metadata.

[1270] Consider how grammar, spell-checking, syntax, and formatting can "correct" a document to change meaning. Any such inadvertent corrections can be explicitly exploited. Consider how a template can infect an innocuous document with new meaning, functions and features and even virus. DigitalDoors makes light of some situations where punctuation, capitalization, and word spacing altered meaning dramatically. While this is not the norm, the metadata infrastructure is part of the metadata universe, and both the metadata infrastructure and the metadata itself impart security risk.

[1271] The template need not be the only implicit risk. When metadata 'touches' other metadata, a macro virus or other embedded feature can spread infection. Such a touch can clone or insert templates, styles, interpretations, macro code, entire embedded baggage, even the file version information can assert

source, ownership, methods for production, time frames, and distribute security risks. The issue is that not all macros include a virus or even bad, but the metadata infrastructure open venues for risk that cannot be secured because the context for the metadata is open ended and the social, cultural, and functional references of the metadata on the data content represent complex semiotic expansion, as described in Metadata in Pictures..

[1272] Metadata does not exist only as a secondary byproduct of process and purpose. In many cases, metadata is specifically created to enhance or enable complex data interactions. Databases are of course complex structures of metadata. The table, row, and columns are defined by metadata. Keys, indexes, and relationships, joins, and views are also metadata. Furthermore, SQL schemas, clustering, and partitioning are all forms of association-related info through contextual relationship metadata. These types of metadata are established before the fact to set the usage and purpose for complex data, but also are created as the performance optimization requirements or business intelligence mining distinguishes additional utility in the core data. It is important to recognize that this metadata creates both implicit and explicit risk that transcends the basic content of the tabular data. Metadata, particularly in database structures, defines the context of purpose, utility, and applicability that is not clear from the context alone.

[1273] Digital signatures on macros and documents do not vet the integrity of the metadata. A macro could have been previously infected or recently infected a document. A digital signature designates security on a narrow range and even that narrow range is subject to internal undermining. Furthermore, a digitally signed document is no guarantee of the quality of the digital signature or that it has not been compromised. The chain-of-custody is linear process that is reliant on prior events in the chain. There is limited guarantee of such security.

[1274] The only legitimate method to secure the metadata infrastructure, style sheets and templates that define the metadata, composite metadata files, and files with content but stripped of actual metadata but acted on by implicit or explicit metadata relationships is to break the context. Any implicit or explicit metadata should be decomposed into parts small enough to erase any evidence of relationship and utility. Redact the content (data) and corresponding metadata to break any connection between them. The parts can be stored in separate files and even widely dispersed. It is possible to create parts so small that they lose any integrity of meaning or purpose, but this is not always necessary. Although the parts might be small enough so that meaning and relationship to the whole is disruptive, be aware that the collection of whole provides context of its own, and therefore the impetus for the disruption and dispersion. This enables security through formlessness, and it is as effective for compound metadata as any other type of content.

[1275] Metadata must be decomposed into elemental data streams where format matches content, for format is singular and uniquely describes a single contextual purpose, and where content represents a singular context. In other words, when metadata describes the context for content, a structure specific enough to infer purpose and utility, and the descriptive metadata is still bound to the data, formlessness is not sufficient for security. The alternative – which is incomplete formlessness -- is multiple renderings that still retain semantic meaning therefore useful for aggregation, inference, or interaction.

AA.4.5.0 In-Channel and Out-of-Channel Tags

[1276] The placement or visual inclusion of tags is purely a procedural issue. Although there are differences, these are immaterial with one exception defined below. Examples of tagging for the original content This is top secret include:

[TS] This is top secret.

<comment = TS>This is top secret.

Tag file: [TS]→This is top secret with actual separate data file: This is top secret.

[1277] Too many other styles exist to list including database and linked references. One style is easily transformed into any other. In effect, the format of tags is a metadata issue and the actual storage location is a procedural issue. Reformat and changing the presentation of a tag is a metadata transformation often performed with a style sheet. Albeit, there are content issues whether tags are in-channel or hidden in out-of-channel overlays, nevertheless, presentation controls can relay the same results regardless of the initial or transformed style.

[1278] Style 1 betrays more information that original because it includes an assessment in stream. Style 2 also betrays more information but it is accessible only under certain displays or with a source view. Style 3 makes a physical abstraction between the tagging and the content requiring an explicit effort to simultaneously show both pieces of information.

[1279] Controlled presentation with overlay or integration or filtered list display could still reveal the tagging or explicitly excluded from the display. It becomes a matter of control or deliver of raw data streams to processes or the users. Ultimately the issue can be reduced to how information is released under granular content control and how the control is varied for content, context, and concept.

[1280] The sole exception to in-channel and out-of-channel tagging is that out-of-channel supports a depth a range of tags including overlaps that is not easily represented in-channel. It is not flexible to show that 'This is top secret' is [TS] under certain scenarios but [U] under others, or that secret is word that in general should not be exposed for its connotations. Multitier tagging is easier to support out-of-channel where rules can be assessed and applied dynamically.

AA.4.6.0 Capitalization is content-, context-, and concept-driven

[1281] Capitalization classification accuracy has direct applicability for risk measurement, security application, knowledge management, discovery, and information sharing. Classification changes as the context and concepts changes, and the failure to accept and support this undermines risk management. Risk management (and subsequent security application) requires completeness and flexibility in classification, including classification driven by capitalization and the meaning variability driven by word and phrase case sensitivity. All capitalization, tagging, and classification is contingent on content, context, and concept. Results are neither static nor absolute, but color the implementation of classification and tagging.

[1282] Note that capitalization does not drive classification or provide any primary value; it merely augments clues to yield more accurate classification. Capitalization is necessary but not sufficient to drive classification. Capitalization is a modifier but does not provide a standalone meaning. Capitalization is neither a constant, a fixture, a given, nor even necessarily accurate within a data stream. Capitalization can be random, erroneous, incidental, applied for haphazard purposes, or used to game against the system. However, capitalization adds to the effectiveness when automating classification.

[1283] Capitalization is content-, context-, and concept-driven. It both reduces and increases classification ambiguity. Traditional classification methods do not reflect this complication because they are lookup-driven within a limited contextually-driven expectation. The differentiation that capitalization can add into classification is blurred by position, usage, and randomness. While often not obvious, it is very important to apply exact classification to acronyms, words, and phrases. While manual classification is generally very effective, it is not efficient and requires highly trained people with a special aptitude for conceptualization. On the other hand, automated classification is essential to create efficiency, to support classification required in high-volume data flows, and to reflect other requirements

including validation, authentication, and reproducibility. Capitalization is one of the clues useful to enabling automation in classification.

[1284] The problem with automated classification is that it is not simple and generally not accurate. Accuracy is increased when classification transcends content list matching and includes semantic factors found within context and concept. In other words, intent and purpose drives classification too. Successful semantic capitalization classification is not driven by white list, black list, or gray list categorization. That fails miserably. Capitalization is often arbitrary, incorrect, or just applied for convenience. Letterhead, advertising material, documentation, and slides often arbitrarily apply capitalization. Sometimes, capitalization can be used to play against the system, for example to thwart spam and Bayesian filters. However, dictionary categorization when combined with semantic parsing rules and pattern matching significantly increases the accuracy and enables full automation. DigitalDoors implements fully automated semantic capitalization classification with a level of accuracy approaching 100%. DigitalDoors supports inline and post-production manual examination with overriding revision options and reporting.

[1285] Most languages, and in particular all European languages, apply capitalization as semantic clues to meaning. The statistically most common capitalization is driven by punctuation rules, specifically, the initial letter in a heading, title, new paragraph, new sentence, line, bullet point, caption, and footnote. This is a telltale of a contextual idea demarcation. The second most common application of capitalization is for distinction of proper nouns from common words. Proper nouns are represented by names, places, and specialized things. For example, Jon is a name, but jon is a common word. However, capitalization is not distinctive because semantic rules override capitalization and blur the answer.

[1286] Furthermore, even meaning is fuzzy, and this has direct implications on correct classification. When jon is a common word, it can be a functional place containing multiple toilets, a single toilet, a customer, or a boat. JON could be an acronym, just a capitalization error, or a functional convenience. The ambiguous initial placement and capitalization of what can be either a proper noun or common word as part of a heading, new paragraph, new sentence, line, bullet point, caption, and footnote is resolved only by context.

[1287] Dictionary matching of acronyms, words, and phrases provides an unclear range of classification matches. In many cases, as with SQL, matching is not configured as case-sensitive, or even case-aware. There can be and often are multiple matches. Elsewhere, in DigitalDoors whitepapers on filters and sieves, we have shown this ambiguity in matching even simple names. For example, reconsider bin Laden as a family name, a terrorist, an industrialist, a reference to any number of legitimate businesses, a clan, a number of different families, a particular person, an antecedent in a complex narrative, and a reference to specific history events. Capitalization does not differentiate the meaning or the classification, whereas only semantic context and concept do.

[1288] Frankly, DigitalDoors does not want to just minimize the range of potential matches. Sometimes a single best classification is necessary, other times, multiple possible, extended and expanded, overlapping, and potential classifications are necessary for the knowledge management ecosystem. DigitalDoors actually wants to expand this range to the extreme because it increases the accuracy in risk measurement by recognizing the potential for oversights due in large measure to classification bias. Multiple hierarchical overlapping categorization enables granular risk measurement and assessment for repurposed value beyond initial and obvious intent. This is important in the extreme for balancing risks of exposure and downstream information.

[1289] The failure to recognize and allow for the full potential in semantic meaning creates security risks within workflow and information sharing. If these additional potential meanings are not identified with initial accuracy, then information that should not be shared is not correctly controlled in secondary downstream sharing or in ownership chaining enabled through database workflow. Security is context-specific. Likewise, classification is context-specific. Context isn't always specific, and it can change, and often does. Situational change alters classification. Time generally causes information to decay, but new understanding can assert new meaning and value to old information too. Both security and classification change as the context changes, and the failure to accept this and support this means that risk management requires completeness and flexibility in classification, including classification driven by capitalization and the meaning variability driven by capitalization.

AA.4.6.1 Categorization Specialization

[1290] Categorization specialization usually occurs when there are multiple dictionary or regular expression matches with multiple possible or conflicting classifications. Capitalization is one automated method to minimize the number of potential categories and establish an unambiguous classification. There are five categories of capitalization categorization:

Table: categories of capitalization

Exact match

All L/C

All U/

Regular case application

Mixed case

[1291] These are implemented as part of all the DigitalDoors dictionaries, as acronyms, words,

and phrases. It is also implemented as part of the inherited dictionaries, such as My Word List, My Group List, and categories and policies. The same rules of capitalization and modification of meaning that apply to words apply to more complex dictionary entries too.

[1292] Realize that the five categories of capitalization categorization do not provide a standalone meaning. Capitalization is necessary but not sufficient to drive classification.

Exactformat example IBM, bin Laden, or BitLocker

Acronym

Trademark

Special form

All L/Cformat example xxxxxx or xxx xxxxxxx

common word

Accident/Error

Convenience

Processing and sorting stability

Special form

All U/Cformat example XXXXX or XXX XXXXXX

Acronym

Accident/Error

Convenience

Processing and sorting stability

Special form

Regularformat example Xxxxxx

Semantic punctuation capitalization

Proper noun

Intent

Accident/Error

Mixedformat example xXXXx

Acronym

Label

Special meaning

Special form

Accident/Error

AA.4.6.2 Semantic Parsing Rules

[1293] The first step is to ascertain the capitalization pattern. The second step is to ascertain the purpose within each capitalization category.

[1294] Capitalization specificity…important to recognize that most databases sort and match with capitalization specificity… need to validate and check for exact database capitalization match.

Statistical (frequency) assessment.

Organizational environment.

My Word List, My Group List, My Category List, My Policy List hierarchies.

Punctuation

Position.

Modifiers (Dr., Ms., Captain, etc.)

Data stream format.

[1295] Data stream structure (field-oriented, database row-column entries, metadata)change the meaning.

Manual Override always last resort (subject to COMSEC override rules)

Playing against the system

Preceding punctuation

Preceding markers or special characters like bullets

AA.4.6.3 Process

[1296] DigitalDoors processes capitalization first by structure, then by syntax. Because structures and syntax provide context to the potential meaning. Structure and syntax provides parallelism that defines the concepts in play. Specifically, structures in databases, XML, data sets with column-oriented fields or positional fields establish relationships that important to meaning and capitalization. It is not sufficient to view the data stream as unstructured when it is clearly structures. Additionally, so-called unstructured data, such as email and documents are not without useful organization, including linguistic syntax. Specifically, data streams are searched for:

preceding quotes

Punctuation

Line feeds or line breaks

Bullet or special characters

[1297] These clues are used to statistically assert whether capitalization is a positional construct or is used to set apart words as different content.

AA.4.6.4 Conclusion

[1298] Capitalization classification accuracy has direct applicability for risk measurement,

security application, knowledge management, discovery, and information sharing. Classification changes as the context and concepts changes, and the failure to accept this and support this means that risk management is inaccurate.

AA.4.7.0 Data classification – the leakage threat

[1299] Classification is not a neutral activity. It is a powerful tool both for good and for bad. It can undermine security methods based on classification as the determining quality for access and usage. Classification is necessary to enable the security paradigm shift for granular content control and it shines from transparency. The more transparency, the better the classification results and the better the resulting security. However, there are limits to this transparency because of the fundamental nature of the structure of information.

[1300] Classification is a management technique to control the usage and distribution of information through the use of labels, tags, classes, metadata, or physical or logical placement. It relies on a taxonomy or ontology for an assignment, but for practical purposes, these assignments are too small, just right, or too big. Too small and too big are wrong. Any assumptions that most assignments will be just right is naïve. The reality is that most things are assigned into these categories rather than just right. It is a matter or power and control. Whoever controls this assignment controls access and utility of the information.

[1301] The manipulation of classification represents the future of security leakage. If information is a high-value or high-risk asset and is classified too small, it leaks past the guards to the benefit of whomever or whatever controlled that classification. When high-value assets are classified as too big, this over classification defeats information sharing. When low-value information is classified as too big, it distorts the perceived value of the information and overgrandizes the source, custodian, or processes

involved to the detriment of more important business activities. Consider the naming of an executable virus as WinWord.EXE. That is the effect of misclassification of process, but it is analogous to the misclassification effects for information.

The classification itself adds context through the label, tag, placement, or metadata. Thus, while the classification can be transparent, this does not mean the information or the classification itself can be transparent. Sometimes the classification itself becomes classified. Content, context, and concept moderate the fundamental nature of the structure of information, so transparency itself must be part of any classification so that the classification itself does create unintentional leaks.

[1302] It is just as easy to buy a control over the classification process as it is to buy an insider. The effect will be similar, although any review of classification is far more useful than exposing spies and nefarious insiders. Classification is far more transparent than the hidden agendas and masked motives.

AA.4.8.0 Semantics dictionaries

[1303] The DigitalDoors dictionaries contain some one million English language entries. Entries are distinguished by type as acronyms, words, and phrases but stored together. These entries are used for white listing, black listing, gray listing and also contextual classification and weighting. It is applied for automated selection, contextual or conceptual recognition, classification and tagging, and also monetary valuation for intangibles. The dictionary includes common entries borrowed from other languages. However, it is not exhaustive. The dictionary is extensible, but it is also supports word-format generation with prefixes and suffixes. In other words, "kind" can become "unkindest".

Prefixes and suffixes are applicable to words, but not processed for acronyms and phrases. This enhancement is effective for categorizing non-standard or rare forms of common words if the prefix

and/or suffix is recognized as a valid modifier in the dictionaries. This functionality is particular relevant for Germanic languages as many words are conjoined.

[1304] Conjoined words Consider, the word "fremdenverkeresburo" which loosely translates as the "travelers information office". Consider conjoined words in other languages, such as Swedish, where "kommunalskolsflaggstångsknoppuppsättaringenjörslärare" translates quite literally as the "supervisor who teaches engineers working at communal schools to put the top cap piece on flagpoles." This serial conjoining is supported only if the prefix and/or suffix is recognized as a valid modifier in the dictionaries. Complex compound words like these two examples represent an exhaustive parsing process that is not practical at this time. However, common compounds can be added as prefixes and suffixes to enhance basic dictionary matching.

[1305] Root matches provide the classification inherited for all compounded word forms. The classifications includes sensitivity level, word type, user, group, mission, role, and becomes the key for all subsequent My Group List and contextual matches too. Note, however, that precedent assures that any compounded dictionary entry will take precedence over conjoined form variants. This is important and logical since many compound words have cultural meanings no longer rooted in the root word or even modified by the prefix or suffix.

[1306] For example, consider the root word "sight". The prefix and suffix lists allow for compounding and recognition of derivative words that include:

Table: prefix and suffix lists

Insight	insightful
Sightwise	sights
Sighted	sighting

Sightless Sightfully

Sightlike Metasight

Unsighty Unsighted

[1307] This dictionary enhancement facility is actually useful for mixed language environments because it lightens the load when looking up words and when trying to ascertain which language is most likely to pertain.

AA.4.9.0 Difference between Content, Context, and Concept

[1308] Content is data, symbol, information, a data stream, or static fact. It is anything that fills a newspaper, magazine, data stream, signal, message, missive, document, video, audio, multimedia presentation, or web site. Content is synonymous with data, but content also includes material represented by labels, tags, structure, metadata, references, links, applets, processes, notations, and out-of-band controls.

[1309] A "ball" is content in that it is four characters or 32 bits of a basic signal. It represents any one of a hundred different things by definitions, nearly an unlimited number of symbolic possibilities disregarded by the Shannon Theory, and points to nearly a hundred different concepts. Among other things, a ball is a good time, a dance, an event without dancing, a toy, a person's name, a game, an activity, and a bullet.

[1310] Context is an abstraction of information to include discourse that surrounds a language unit and helps to determine its interpretation. It is an abstraction of circumstance that sets the facts that surround a situation or event. It is also the environment, purpose, usage, or application of information to a particular scenario or problem set. It is the non-entropic value of data within a signal, message, file, or other data structure. It is the linkage of relevance of information to a particular biological system,

cultural understanding, or specialized workflow. This is frequently presumed, but any framework or perspective is not absolute.

[1311] The content "ball" takes on a more specific meaning when a context is specified, such as a baseball stadium. Nevertheless, a Darwinian context abstraction is still not specific enough to determine value. For example, if we hear the announcement that "The pitcher has thrown the ball, an inside curve ball, to number 27, Red Ball, and the umpire has called it a ball." Three references do not have substantive indication, but each instance in turn, defines a toy, a person, and a measurement. The two contexts, location, and distribution of the message, narrow content meaning, but is not definitive as to meaning.

[1312] This wiggle-room provides unlimited private communication within public venue as long as the private context is not exposed or decoded. Al Qaeda published messages as wedding announcements in metropolitan newspapers that disguised more menacing information. This was a low risk and consistent context. Note that posting a wedding announcement within a category of vehicles for sale could either represent a regular mistake or a specific intention to create a semaphore about the context of the message. Such semaphores are contextually dependent and represent a method to encode information in plain site.

[1313] Concept is an abstraction inferred from specific instances of information based on a cognitive knowledge of the context and content. This can include culture, predetermined interpretations, business rules, formats, constructs, structures, abstractions, attributes, properties, categories, labels and tags, notations, metadata, and/or the definition of a process or computer application to define usage and intent. A concept can also be a partially-formed idea that is still incomplete. Using the same sentence, "The pitcher has thrown the ball, an inside curve ball, to number 27, Red Ball, and the umpire has called

it a ball," we have a concept given sufficient cultural acclimation that one player on the other team has thrown a ball at another player named Red Ball that is trying to hit or otherwise accumulate enough points to get to first base or beyond inside a large presentation venue.

[1314] The concept is a broader but narrowing constraint on the content. In normal situations, context and concept together establish cultural meaning, but as stated above, a secret context can override meaning for establishing a private communication within a broader public context. Concepts that can be masked, purposefully perverted, or those at odds with the general context distort meaning for normal literary expression, to joke, or to nefariously mask content.

[1315] The complicating thing about concept is that it is modified by the more definitive context, but can be used to establish a set range for meaning. For example, if the context where defined as a secret message passed in the clear at a particular time, then a special dictionary would provide lookup on actual meaning instead of the more obvious conceptual indication. The concept of "I had a ball." depends on a defining context, so that the concept is transcendent of both content and context.

AA.4.10.0 Classification as Proxy for Scarcity

[1316] Secrecy classification is based on the loose assignment of fuzzy-meaning definitions that often include words like 'top secret,' 'secret,' 'confidential,' and 'unclassified.' Other tags exist to fill the gaps of a complex taxonomy, such as 'sensitive but unclassified,' with many others based on utilitarian functions, features, ownership, stewardship, source, methods, power, or organizational control. In practice, this system is used to limit access to information, thereby creating artificial scarcity of information.

[1317] In other words, the higher the classification, the more limiting the distribution for that information. Intermediate tags and other compartments are applied to balance the functional needs of

information access against the absolute needs for security. Since absolute security precludes workflow, security is relaxed in order to ease that functionality. These intermediate and confusing tags establish touch points for alternate accessibility while legally complying with the more defined and restrictive secrecy classification system. The intermediate tags blur the lines for strict assessment, without real compliance.

[1318] A simpler measurement schema is the scarcity of information. Whenever information is accessed and routed through a processing system, the information is exposed. The duration and frequency of that exposure detracts from that scarcity. Scarcity itself is not measurable, except as the absolute of unobtainability or logical uniqueness. On the other hand, the absence of scarcity is measurable. As such, we can measure the exposure of information, how it is shared, under what cloak of control (limiting secrecy), and how the information is redistributed downstream in alternate forms or formats. Therefore, a range of scarcity is reductive from absolute unobtainability. At some point, and this point is both flexible and difficult to define, the distribution of information is sufficiently widespread as to undermine that secrecy to make the information public knowledge or generally available.

[1319] Public knowledge has limited value. It is not scarce. It is not really useful. Therefore, security can be defined as the artificial control and enforcement of scarcity. Secrecy classification can thus be better defined in terms of numbers of access or rigorous limits for access. In this way, Top Secret is not just a tag with some judgmental meaning or defined in terms of people, roles, or missions, but rather in numerical terms of views seen versus view restrictions enforced. Intermediate tags assume more useful meanings because they reflect intermediate values. In fact, even top secret and secret and these fuzzy intermediate terms can be couched in terms of statistical or absolute numerical values for scarcity. All these terms become shorthand for a more rigorous definition of secrecy, and specifically of secrecy

in terms of an exacting language of scarcity.

[1320] It matters little whether the scarcity is natural or artificial. Natural scarcity could be driven by the lack of available information, a bottleneck of resources, or delayed production. Artificial scarcity is monopolistic, generally driven by controls in order to create and maintain scarcity and hence demand-side value. When artificial scarcity is applied to information, that information becomes valuable particularly when the scarcity can be enforced and sustained. This is the ecosystem for creating valuable intangible property, and in particular, copyrights, patents, and trade secrets foremost, and it is based on the artificial creation and sustainable enforcement of scarcity.

AA.4.11.0Threats of Meta Data

[1321] Data lacks purpose, structure and exacting patterns, context, and relevance without external references. Metadata provides those references as internal aspects of the data construct. As such, data has been downgraded as an important facility in favor of the more precise, extensible, and useful metadata. Long live metadata.

[1322] The information processing and library science worlds define the concept of data. Whereas information processing perceived data as a structured source for applications, library sciences sees data as the answers to questions and problems. Although the Internet, Gopher, Archie, WAAS, the world wide web and browser search technology began as an information processing concept, these now the complement to library sciences. As these search technologies and taxonomies for setting purpose, context, and references have grown to organize data, all data has become contingent on meta-tagging or labeling.

[1323] Applications, data, configuration baggage, scripts, complex formatting, documents with other types of documents inserted or referenced and external links are all forms of metadata. Many

documents contain scripts, applets, functional code that might be static but more often is dynamic. Viruses and worms are examples of unwanted dynamic modifications.

[1324] Applications often contain tables of data or reference externally linked modules with tables of data. This data can change, and with it the linked modules or even the application itself. Applications frequency update configurations and other variable data streams internal to itself. Some applications modify themselves and rewrite its internal code. Viruses and worms are examples of unwanted dynamic modifications. Applications often support persistence of data; this is the situation when user data, configurations, or metadata is stored within the storage space of the saved application. Data persistence provides several avenues for security lapses, modification of the data and the usage of the data storage space as a platform for a virus or worm. In addition, metadata provides all sorts of opportunities to hide polymorphic traps that can bypass signature scanners and even illegal process monitors. It is a rich game where the very extensible nature of metadata makes for a very extensible solution set but also a very perverse battleground. Every tool, even metadata, becomes a weapon

AA.4.12.0 Degrees of data structure

[1325] Structure, as in data structure or structured data, is a biased and imprecise definition to describe a single presentation of information. The definition conveys a certain amount of relevance to traditional data processing operations in terms of what classes of applications access and manipulate the information, but misfires when information is described within a new DigitalDoors ecosystem of meaning, security, and sharing.

[1326] Structure (also variously called a container, format, setting, scene, rendition, interpretation, translation, transformation, presentation, display, exhibit, report, or framework) is a single purpose-driven instance of the informational content space. The purpose driven instance is just one of

unlimited possible structures, where structure adds context and concept to that content. Structure is just an attribute or property of content. To view structure as distinct from content misses the importance that context and concept imparts to content. Information is not constant, and in fact is transformed to meet different purposes and is altered by its structure. There is no single universal structure, as each structure is a single instance of the possible set of structures. Data normalization and denormalization are possible structures for the same information.

[1327] Traditionally, data either has structure, or it doesn't, or is somehow partially structured. This differentiation is commonly grouped into three categories. Structured data is found in databases, in a self-defining format such as XML, or in field-defined data sets with rigorous size/offset delineation. Unstructured data is generated by end-users as documents, images, audio, video, and composites. Semi-structured data is delineated by field definitions, for example by e-mail with delivery address structure but mostly freeform content. However, the strict demarcation between structured and unstructured fails with all-purpose reference to this hybrid semi-structured format with some structure but mostly freeform content. This failure points to the more serious problem within the infrastructure of data.

[1328] This definition of structure and its categories is important when working with data and choosing tools, but is not viable for an ecosystem of information in terms of its meaning, security, and sharing. If we cannot even define it precisely, then we cannot assess its meaning and value, cannot secure it accurately within the formally-defined Bell-LaPadula Model sensitivity-level categorization, or propose safe methods for cross-domain information sharing. In fact, DigitalDoors has explained the failures in information sharing inherent to the creation of a simple cross-domain tagging nomenclature. See The Failure of Tag Reflexivity for Information Sharing. Quite clearly, this definitional failure is explicit to the lack of a formalization of information.

[1329] A data structure is not to be confused with the concept of structured data. The words are similar and do not imply one therefore the other. Structured data neither requires nor necessitates a data structure and a data structure does not imply the reality of purposeful structured data. Whereas a data structure is a precise application of a formal schema or layout to information to delineate it and provide a clear purpose, structured data need not necessarily have a data structure applied to it. There are other ways to structure data. In fact, part of the confusion is that unstructured or semi-structured data could be parsed by a data structure and still fail to be configured in a meaningful way. The application of a data structure or tags, metatags, XML, or a data hierarchy is no guarantee of the meaning, security, risk, or purpose of the information. While a data structure is often an indication of structure, such contextual or conceptual clues could be as useless and false as they are purposeful and true.

[1330] What is useful is the simple understanding that structured data is one form of tagging represented in one way that can be transformed and tagged again in one or many overlapping ways. A database with a data structure or a schema is the imposition of a tagging system and presentation format that makes clear the grid-like nature of the rows and columns, and index relationships, if any. In other words, either a structured database or an unstructured document is both structured contingent upon a predetermined purpose. The structure can be transformed for other purposes with new meaning, new security requirements, and new information sharing value. The transformational risks cannot be addressed from the premise of any initial structure because the transformation alters context and concept with both the loss and the addition of content, context, and concept.

[1331] For example, a printed P&L statement shows a rolled up subset of income and expenses in a predefined GAAP format as of the date of the data transformation. The rollup does not detail line items, so there is a loss of information. The rollup does typically include a process date and does derive

its source from the person generating the report, so there is the addition of information too through the explicit change in content. There is also the change in context through the format transformation from raw line item data to a categorized report. There is also a concept change because the purpose shows aggregated results instead of transactions and perhaps inference over a longer period of time through the presentation of estimates of future sales. The concept also changes because the date shows the date of the report production and the relevancy of the future forecast rather than a historically precise perspective. The concept changes again because the person(s) creating and presenting the report puts a mark of integrity on the information through this interaction, thus asserting both a new source and a new method.

[1332] Structure is range without precision. DigitalDoors asserts that all structures are defined by a hierarchical taxonomy that supports categories without limit and that supports overlapping structures. See Multitier Hierarchical Overlapping Tag Structure. As such, all databases, documents, and data streams have a degree of structure that unfortunately we can only classify vaguely as more or less, that such degrees of structure are really not useful, and to date we are unsure even how to measure degrees of structure or if it is relevant to anything. DigitalDoors can, however, measure the overlap of information when it is categorized into multiple categories and assess its risk potential when shared.

[1333] The reason for range without precision is that a database, which clearly has a formal structure represented by the schema of grids and indexing, can be restructured for different uses having more or less structure. A report might contain all of the original content but have no structure, be re-interpreted into a different grid and can contain free-form data, such as BLOBs, which might itself have or lack a formal structure. How does this really differ from the e-mail with its address structure containing unstructured messages? Furthermore, the so-called unstructured document from MS Word

does indeed have an internal structure that uses tags to differentiate between titles, paragraphs, figures, links, footnotes, and deleted metadata. How can this really be called unstructured? A formal external schema definition does not differentiate a structured document from one structured with an internal inline self-defining tagged layout either.

[1334] In fact, a database can be converted into XML, XML into a database, and an MS Word document can be converted into either a formal database or XML without additional external information. In fact, MS Word supports both XML and ODF to defines its structure. This does not mean that conversion necessarily adds or removes structure, that it is measurable, or as even useful, as previously cited. Conversion does alter context and purpose. It does show the arbitrary use of the definition for "structured data" is meaningless. See The Transitional Relationship Between Structured and Unstructured Data.

[1335] Tagging (or labeling) of content and use of XML, sensitivity levels, context, and functional application also represents a structure. Indeed, XML-coded documents are included within the traditional definition of structured data. This raises a problem of conversion between structure and tagging and problems of meaning conveyed by structure and what is lost in the conversion. Tagging does indeed provide structure and can equal or exceed the amount of structure inherent in a formal database schema.

[1336] The transition from structured data to a tagged document is just a transformation. The transition from a tagged document to structured data is just another transformation. The association between the structured document and a tagged document is one mostly of context and concept but also of content. Consider each as a singular instance of the hierarchically-defined data set. Transformations can include more content or less content than the original, but often includes additional contextual and

conceptual information by virtue of the transformation. Refer to the prior P& L example for how information is transformed by aggregation, inference, and interaction.. This is a very important point for assessing search results, the meaning of the information, its security risk, and how and what parts of it can be shared.

[1337] A structure is a rendition of a context and a concept for a purpose. The transformation of a structure into another is precipitated by a change in context and/or concept. The transformation must include aggregation, inference, or interaction to occur. This is an important point because it recalls that information does not exist in a vacuum but as a meaning contingent upon a content, context, and concept. The transformation from the old format to the new one presumes an understanding of the defined purpose of the old, how it differs from new, and how the new format enhances the new purpose. Although we presuppose that a hash transformation is unique and non-reversible and thus non-recoverable, we cannot hold data and format transformations to such a high standard. That is why the culture loathes information sharing; they understand the risk implicitly, if not formally.

[1338] Information is not constant, is transformed to meet different purposes, and is altered by its structure. There is no single universal structure, as each structure is a single instance of the possible set of all structures. No one size fits all, and no single structure is necessary and sufficient for any or all requirements. As a result, we need to understand that potential for the Multitier Hierarchical Overlapping Tag Structure as focal point for analyzing meaning, security, and sharing. We need to understand that structure, and the degrees of structure, are only a window on the meaning, value, risk, and potential for information.

AA.5.0.0 Defining Information

AA.5.1.0 The meaning of information and related context risk

[1339] Following is an analysis of the unanticipated risk from information and the security consequences arising from alternate semantic and arbitrary semiotic contexts. Because we often presume that information has a single interpretation, a designated a priori utility, and a definitive purpose, we underestimate its capacity to mean other things. We fail to anticipate that information can be useful in unintended, unintentional, and ambiguous circumstances. We expose a greater security threat because of these expanded meanings without any realization that it even exists or that our own information will be turned as a weapon against us. However, as the consequences become apparent with exponential growth in identity theft and security breaches, we are left wondering why. This is not only an economic problem, but also a social and intellectual one too. All have dire security and competitive consequences. These consequences are the direct effects from the explosion of information, expansion in information processing, and the convergence of both into the single deliverable format of metadata.

[1340] The focus here is on the expansion of knowledge through the application of external factors. It has its basis in a phenomenon first observed in genetic biology where exact DNA genome sequences express into different results through unknown or imputed external factors, general the process of genes turning on and the sequence of this expression process. Because the starting point of the DNA -- that is the information basis is identical and the expression process is controlled -- the ending point should be identical and anticipated in advance. However, variations and mutations often result. This is external to the genome, hence 'epigenetic' or external to the gene. Sometimes these variations occur over several generations masking not only the cause and effect, but also any possible information delivery channel. Clearly this out-of-channel conduit exists, but it is beyond our current knowledge and control.

[1341] Because analogous results occur within the context of information processing -- some are

accidental but some are intentional and manipulated -- through out-of-channel conduits -- the consequences are expressed in the form of epigenetic information that was not previously recognized or did not previously exist. Clearly out-of-channel information conduits exist, but it is beyond our current control. This is the existing conduit for delivery of computer virus, malware, information theft, identity theft, and the many mutations increasing security risk and actual observable security consequences.

[1342] This thesis can be simplified into the following nine assertions.

1. Information and information process can be represented within a single unified structure called 'metadata.'

2. Information is the formal representation of knowledge, codes, and facts.

Process is the expression of information into an observable event or action.

The meaning of information is dependent on context.

There are no known limit to contexts.

The aggregation of information creates new information and new contexts.

Interaction is the expression of information with secondary aggregation.

[1343] We know of no ways to describe information in a primitive format where meaning is explicit and unambiguous and replicable without variations and mutations.

[1344] External factors, in the form of information and information process, create expanded interpretation, utility, and purpose. The following paragraphs define these assertions.

[1345] These assertions are expanded in the following paragraphs.

[1346] The boldest assertion is that process and information are the same thing because a process is basically an expression of information. An application, whether website, an applet, a macro, a

computer application, business rules, an assembly line, corporate routing process, chain-of-command, a virus, or a trigger is basically an information stream executed by a defined computational engine. This is no different from seeing RNA activity as an expression of DNA applied to a natural chemical or physical reaction engine. Correspondingly, process is an information source that requires security. This is obvious to people who view IP and organizational infrastructure as valuable assets but totally overlooked by organizations that do not document workflow and hierarchical controls. Even physical devices are manifestations of data streams. Because information is represented as a data stream with a starting point and ending point within a structure, format, encoding, and probable utility by its containment, that containment is definable as a metadata container. In other words, each and every data stream can be represented by a structure that is self-defining, self-replicating, extensible, and convertible into other structures, the very definition of metadata, or data about data. Additionally, physical representations can be logically converted into wholly information streams as the technology to turn information into physical form advances.

[1347] Information in any format, any data stream, or a metadata structure is the formal representation of knowledge, codes, facts, or the definition of the steps of a physical or even logical process. You can document a process, create blueprints, patents, express a process as a math or logic equation.

[1348] Information process is the expression of information into an observable event or action. Information, such as music can be a notation, a sound stream, a multi-channel stream, compressed notes and volumes, and be convertible from one representation to another or expressed through a synthesizer.

[1349] The meaning of information is dependent on context. Compressed encrypted data could seem to contain no pattern or known, and yet contain just about any secret, any process description or

code, or any formal representation of knowledge. Extraction or expression of that information is conditional on unraveling the context, which depends on an encryption key, knowing the encryption method, knowing the compression method, and applying a reversing process in correct sequence. Conversely, any data stream can be perceived as applying to a specific utility and purpose but referring to one or more other contexts. Metadata, such example an MS Word document or a data table, are thus extensible to unanticipated purposes and coverts channels. Redaction to a limited data stream is the only way to minimize or limit the dependencies. Manual (human) review can be cached and automated to augment the manual process and expand the range and depth of inference detection. "Unstructured" data can be represented in a formalized structured form, thus the perceived differences between a discreet restricted data setting and freeform is moot. So-called unstructured streams must be intelligible by some process and they have some definitive structure. Context is a structure, and structure is a context. Efforts to unify data formats -- for example by designing a master data format (MDF) -- with standards represents the effort to establish consistent formats and do nothing to secure them or secure the information within. It is merely a way to assert a single fixed context where single contexts are at best a fiction.

[1350] There are no known finite contexts; there are always alternate contexts. This means that anything that is presumed to represent a formal perimeter can be redefined and circumvented, thereby explaining the fundamental failing in security methods based at a perimeter. This also applies to secure code efforts; alternative contexts or the alteration of the code with malware patching is no fiction and renders efforts moot to increase security through better techniques. Granularity goes a long way towards minimizing the number of alternate contexts, perhaps even creating a 1:1 possibility, thus initiating the underpinning for the DigitalDoors granular content control security ecosystem.

[1351] The aggregation of information creates new information and new contexts or recreates privileged or de-identified information. Information can be aggregated from within a single data stream, related data streams, or from a single or related data streams with other external sources. Aggregation enhances information, extends its context, and replicates facts presumed secure, private, or confidential. Aggregation exists as a process of simple data combination, data mining, knowledge discovery through inference, and the interaction of data and process.

[1352] Interaction is the expression of information with secondary aggregation. Since process is the physical expression of information, the intersection of these creates new processes and new information. The Storm (peacomm) malware is a prime example where reconstitution of sleeper agents recombine to create a pernicious, mutating, and hidden platform.

[1353] We know of no ways to describe information in a primitive format where meaning is explicit and unambiguous. Math is close to some measure of objectivity but many human situations, endeavors, and observations have no primitives that are context-insensitive or culturally unbiased. Results also depend on external factors, such as the aggregation of information, the people involved and the assessments made, so that the output contains inevitable variations and mutations.

[1354] External factors, in the form of information and information process, create expanded interpretation, utility, and purpose. Even music notation is interpreted by conductors and musicians, where the interpretation is unexpected representing a larger universe than ever anticipated. If you could anticipate the permutations, you would have a closed set, but a closed set is a special and nearly impossible situation. You always have to be aware of the potential for information perversion, repurposing, misuse, and turning that very information protection into a weapon against you.

AA.5.2.0 Content Concept and Context - Correcting the Security Gap for Data Object Models

[1355] The security battle is accelerating between content, concept, and context. Traditional security has been applied only to content, which is strictly defined as the actual data contained within a document, file, or data flow. However, there has been a paradigm shift to complex data structure defined by an object model, dispersed, or aggregated as part of a large process or workflow without a corresponding paradigm shift in data security. Protection of the content alone is insufficient. The existence of the content and understanding of the attendant workflows provide conceptual and contextual information that is often more useful than the actual protected content. Content is a finite security risk, but concept and context represents an infinite risk. This is a paramount security issue as the data element migrates to complex data structures defined by object models. The security needs are exacerbated by structures containing metadata, data paths, links, references, so-called smart tags, and even workflows and applications, as shown in:

[1356] Encapsulating such documents, files, data flows, data structures defined by object models, or other entities with barrier-based security or breaking apart these into random chunks or unitary data elements is neither effective nor feasible. Barrier-based security breaks workflows, prevents information sharing, and fails regardless to protect the concept and context; it does protect some of the content, but at what cost? This security gap exists between content on the one hand, and concept and context on the other. Rather, such complex data must be protected at the individual granular level and within the context of what concepts can be discerned through inference.

[1357] Traditional data security has been implemented against operational sites and against specific data files. These represent barrier-based methods—for the site as an entity—or the individual data source of value. It is effective when the balance of power is unequal between strong defender and lesser attacker. It is effective when the elements attacked are finite and discrete. However, when assets

are distributed or when small assets can be captured and repurposed for significant gain, barrier-based security fails. The power is with the attacker, and protection requires a disproportionate response to maintain any legitimate protection. Any type of siege, hit-and-run guerilla attack, or slow and covert undermining defeats this technology.

[1358] It is also important to cast these significant shortfalls in security to the historical basis of data processing. Initially, the technology itself and the scarce equipment held extreme value. Protection meant physical protection. When the operations become mission-critical, the process become very expensive and more valuable than the equipment itself; this infrastructure was protected. When computing became cheap, ubiquitous, and the value of shifted to the data, protection was positioned to maintain backups and protect cloistered continuity. Now, data has become ubiquitous with the risk in the theft of and misuse of that data. The processing has become significantly more complex but distributed through always available networks. Furthermore, the process itself has been incorporated into the data, merging the distinction rigorously between data, structure, and process. The requirements have changed from central protections to protection of distributions, long supply chains, and interconnected workflows. The co-mingling of data and process blurs the line between protecting intellectual property, trade secrets, and private data, this changes the battle from protection of content to protection of concepts and context. This is the security gap now visible with the transaction from data element concepts to object models.

[1359] There are six areas of security, of which the first two only have been protected by barrier-based security:

Data operation

Data element

[1360] The remaining four represent the results of processes embedded into the data and distribution, and the shift to data structures defined by object models:

Data object

Source

Event

Method

[1361] Traditionally, the data element defined the operation, one-for-one, and the operation required that data element. Each data operation and each data element had a known single purpose and mapped uniquely. For example, "this is the credit card operations and this elemental record set is today's transactions." Record sets are no longer so focused, and for example, the credit card transaction data is now typically applied for the primary activity of payment settlement but also used to accumulate cardholder benefits, for sales and marketing activities, and track card usage patterns. Tracking card usage patterns represent the front lines against credit card fraud and identity theft, but tracking is also extended for secondary supplemental marketing activities to both existing and potential vendors and cardholders. This repurposing of activities and data represents the cutting edge in return on investments in data processing, but also demonstrates why the theft of data and privacy information is so rampant. There is no longer a unique one-for-one match between operations and data sources.

[1362] A single security barrier is no longer conceivable as an effective countermeasure. As you see in this simple example but very common security gap, multiple operations are performed against the single record set and the single record set is probably processed multiple times and even converted into alternate formats for additional processing and reporting operations. Because of this repurposing, reformatting, and redistribution of source data, the application and thus the security of the data element

is no longer clear cut. It is complex. There exists a mesh of operations and applications not protected for concept and context.

[1363] It is also important to recognize that a data element can also constitute a metadata item too, and hence represent a non-finite security risk. For example, the profile of an individual is a defined data element in its own right when perceived as a structured record set. It is also a profile that describes the individual that can be repurposed and combined with other data elements. This extension creates enormous security risks if the profile is seen only as collection of data elements including last name, first name, social security number, address, et cetera. Rather, the collection is a descriptive metadata item, describing the individual in terms not intended by the data processing group. Repurposing extends the value, risk, and security needs.

[1364] The shift from the data element that is compartmentalized and defined in terms of purpose to that of the data object is in terms of the expansion in complexity, purpose, and capabilities. The data object is basically a container of containers, each container with its own structure. The data object is defined by its own internal map or by an external schema. No only is the object likely to contain mixed data types, it is likely to support activities with multiple purposes. For example, the profile for an individual contains contact information, but it also used to validate the integrity and validity of the individual for transactional purposes.

[1365] Protecting the contact information without protecting the application of it for verification represents every much the same security breach; it just happens in a different way, a different place, and at a different time. Note, that access to the internal map or external schema also creates a security lapse. Knowing that a particular branch holds a current profile and that another branch holds an outdated profile and which information serves which function, opens another door to data theft and misuse. The

outdated profile is very useful for identity theft because it can be validated as correct information but out-of-time context. The internal map or external schema also explains how to repurpose data.

[1366] The complexity of maintaining security does not end there, either. The data element no longer exists as a unique and single entity. The applicability of the data element is not finite, but infinite, or at least larger than the source organization will imagine. An attacker, attacker, or insider will perceive a novel use that will damage the organization or its customers. Through the application of object model storage structures, the purpose of the data element is unclear without the schema. The schema is a map detailing the name and purpose for each element, although the purpose often represents many different things depending upon context. For example, the date of creation might indicate when a Word document was first created by the author, or it might indicate to a defense attorney that material was created after-the-fact to justify or cover-up events.

[1367] This descriptive mapping represents metadata, data describing data and its purpose. This second order of information needs to be protected as well, as it often describes details about the data that are more sensitive than the data itself. For example, the metadata for the credit card transaction describes the vendor, the customer, and the purchase environment and this is more valuable that the account numbers or transfer of funds between accounts. Note that the schema is most likely to represent the obvious structure and purpose of the data, but not secondary or misapplications of that data or misapplications of the metadata itself.

[1368] Hence, effective security requires the understanding of the repurposing value of the data and the metadata within a context of mesh access. This data and metadata cannot be simply locked down precluding access. That stops processing. If the data is unlocked for processing, that either/or access security does not protect against the unanticipated data repurposing. Even if the processing infrastructure

seemed to be contained and protected within the perimeter of physical and logical barrier-based security, that is within walls and firewalls, the nature of networking all but guarantees unanticipated access. DigitalDoors counters this type of mesh threat to data and metadata by categorizing each such element and then extracting or dispersing the elements individually; this is an infinitely-granular security approach. This approach enables review of repurposing and control a' priore and also posterior priore of the elements not possible with encryption and barrier-based security. Securing metadata is at least as critical for security as protecting individual data element content. The security gap is predicated by the ability to infer the sources, events, and methods represented by the object model and implemented within the corresponding data structure of an actual file.

[1369] "Source" has two meanings and two security implications with a single security imperative. A source in intelligence idiom is the person, group, operative, spy, satellite, or other channel from whence information is gathered. A source is derivation of the data element, the author, data collection point, editor, workgroup, manager, or other participants in a process or workflow. Although sources are not usually data elements, and thus not protected within traditional barrier-based security, they do represent metadata with a very high need for security.

[1370] The compromise of an intelligence source could jeopardize the life or utility of that source or render it ineffective in the future. Knowing the source, or that the source is a leak, could render prior information irrelevant as information is altered or plans are changed. However, knowing the source also describes the integrity, validation, and significance of the information. The need is to protect the source, not prevent all access to the source. The information is the data element, but the source is metadata. In the context of data processing and the object model structure, the source is not only leaf content, it is the paths and linkages to leaves, paths, and the inferable interspersed structure, it is also concept and

context.

[1371] The concept is the idea projected by the mesh of leaves, titles, headlines, file names, storage locations, or analysis of workflows. The analysis of the workflows shows the source, both within the intelligence community and within the data processing community. Sometimes, concept is represented by words, numbers, sometimes by audio and image. Sometimes, the concept is obvious with only the object model. The structure, shown by what is included or populated from the object model or missing and non populated, the paths to leaves, metadata, and embedded applications also defines the blueprint and thus the concept for the file. In the object model, concept is shown in terms of the blueprint itself, the structure, and the interpretation of the content. If you get the drift of this meaning, the data model itself, even if devoid of content, describes the purpose and concept of the file as a whole.

[1372] Concept is also defined by event and method. Embedded processes, triggers, workflow assignments, lists of sources, lists of processes, and other attached documents, links, applets, and metadata impart the abstraction and notion of the file as a whole. Protection is only feasible when elements and methods are categorized and protected at a granular level. This means that leaves, metadata, structures, links, and references must be categorized, extracted, and dispersed. It also means that analysis of the security risk cannot be affected element-wise only but reflected through the aggregate value of the object model when reviewed in mesh framework. The combination of seemingly insignificant leaves might easily convey the entire concept without the need for any of the content.

[1373] The context is the necessary circumstance described by the data elements, the object model, and all embedded processes and workflows. For example, the presence of the Echelon Communication Facility in the British Isles indicates collection of European voice and data traffic. You do not need to know the details at that traffic to understand what it is and how it can be applied. If

communication capture facilities of Echelon where dispersed among all the McDonalds of Europe, the context would be hidden and its existence and purpose effectively masked. The data capture workflow could be hidden within orders of French fries and Cokes. As long as the secret is suspected but not open, the context can be obscured. Even if the context is exposed, the dispersion of the data collection and leaf content is only suspected but not known with certainty. In the same way, the presence of certain object models gives indication of purpose. Security within a data processing environment is dependent on masking the context by protecting metadata and data elements in the granular instead of the packaged whole.

[1374] An event represents any triggered process, such as a macro, add-in, or application. Potentially, it allows the data structure to be self-aware, so that it knows where it came from, who saw it, automatically updates components or data with a time element, and provides tamper protection. This risk is not external from attackers, but rather internal. Complexities, failures, normal events, and programming errors can compromise event security unless oversight is performed on the content of individual leaves, leaves in aggregate, the paths and model structures, and the developing mesh of workflows.

[1375] "Method" has two meanings and two security implications with a single security imperative. A method in intelligence parlance is the technique in which information is gathered or applied. A method within an object model is a process on how data is presented, stored, processed, or distributed. In fact, a method can be any triggered process limited only by imagination. Barrier-based security protects access to the data and prohibits access to the method. However, whenever the barrier is breached by password or unintentionally, the method initiates a life of its own apart from the object model and any associated elements. Protection is only feasible when elements and methods are

categorized and protected at a granular level.

[1376] In conclusion, the security battle between content, concept, and context will be won not on a finite security of barriers but on the infinite and open analysis and protection of leaves, paths, and structures of the new data object models. Traditional security has been applied only to content, while the existence of the content and understanding of the attendant workflows provide conceptual and contextual information that is often more useful than the actual protected content. Content is a finite security risk, but concept and context represents an infinite risk. Such complex data must be protected at the individual granular level and within the context of what concepts can be discerned through inference using the paradigms of DigitalDoors technology to break apart the data with files defined by object models, review the complexities of the many leaves, and inference possible from the paths and embedded metadata sources, events, and methods.

AA.5.3.0 The Structure of Information

[1377] Although information has not really changed, we use more complicated formats and structures now. As a result, the paradigm for information security must adapt to reflect effective security with more complicated formats and structures. The centrality of content, context, and concept is critical to establishing security and enable information sharing. Unlike Cassandra, we cannot merely prophesize the forthcoming reign of information warfare and economic damage from information theft. Rather we must provide adaptive, effective, efficient, capable, and realistic security solutions in balance with the complex requirements of information sharing within all multilateral agencies.

[1378] Security adaptation must reflect compound data structures, complex data storage models, composite formats, hierarchical layouts, and various representations of data more complex than binary streams, delimited fields, metatags, metadata, links and references. Security adaptation must reflect the

convergence of information with the application process as well. Security adaptation must reflect the singular risk from a defined unit of data to include risks compounded from searching, data mining, and workflow recycling. Security adaptation must reflect more than the internal aspects of information leakage to also defend against externalized aggregation, inference, and data-to-data interaction.

[1379] Information provides more meaning in the sum of the parts than in the parts alone; this is reflected by factors, including symbolic, syntactic, pragmatic, format, structural, order-centric, cultural, purposeful, semantic, and semiotic. The structure of information is an aggregation of the content, context, and concept, that is, the basis for the DigitalDoors granular content control paradigm. This basis for any complete security must start with a thorough understanding into the context and structural assembly of information.

AA.5.3.1 Information Structure

[1380] The processes for updating the formats of data, integrating many databases, adding the unrelenting stream of new data, and cleaning old data is one of the most vexing and costly of ongoing information technology (IT) tasks. Also consider the ramifications when data is transformed into XML, tagged, strewn with comments, and embeds functional code. Frequently, databases or field-oriented records (as in COBOL or ISM) are revised to keep pace with changing applications, stovepipe process migration, or the integration of technology silos into a coherent workflow for web-based processes or SOA modernization. Although cleaning and transforming data for use in new models is perceived as a show stopper for daily operations, realize that this disparity is also seen as a mother lode of data for attackers who can repurpose the data through aggregation, inference, and interaction.

[1381] Information is defined by content, context, and concept. This is also reflected as the fundamental data modified by syntax, structure, semiotics, and pragmatics. Content does not stand alone

with a universal meaning. Punctuation--Context alters Concept is funny example of text meaning diametric opposites with some minor structural differences or minor perspective changes. Context is established by format, syntax, structure, purpose, and other factors that include order, assemblage, tagging, and metadata. Concept is determined by source, destination, purpose, usage, reference, and cultural factors. Even context, which some perceive as straightforward and self-documenting, is contextually biased by metaphor, metonymy, cultural idioms, relation of parts to a whole and whole to the parts. Context is also a product of process, workflow, source, method, integrity, and validity. Security without an understanding of the centrality of content is not security at all.

AA.5.3.2 DNA versus RNA

[1382] DNA is the archetypal model for information structure. DNA creates RNA. RNA not only applies a process extracted from the data of DNA it also expresses itself to make copies or transformational variants. RNA is the application code, albeit a macro, an applet, client side script, server side script, helper script, an embedded function, a virus, a functional computer program, SaaS, SOA, remote process, or the complete workflow controller. RNA recreates DNA. The differentiation of RNA from DNA really has no context. Application packaging is such that often it contains segments of data or persistent configurations, or conversely, the compound information package contains segments of application code. Differentiation of information from application is purely a definition of context. This is directly consistent with the prior discussions of information in terms of granular elements and contexts of format, structure, purpose, and so forth. An application is mapped by its structure, usually a header, a loader, parameter input fields, blocks of code and blocks of data, and internal allocations for buffers, strings, and temporary storage. This application structure is the analog of the structure of information. DNA is agnostic to information, application, infrastructure, and ecosystem; in other words,

the same DNA information structure defines proteins, prions, bacteria, viruses, deep-ocean volcanic vent flat worms, arctic birds, and humans.

[1383] DNA and RNA are really one and the same, and interchangeable. DNA creates RNA. RNA creates DNA. They are different formats for the same things. In the same way, text, data, application code, macros, working processes, XML, metadata, and the information helix are one and the same things. Application code is a subset of the greater data superset converted into an expression for a contextual purpose. Perimeters are expressions of data, applications are expressions of data, and MS Word documents or Microsoft XML-based XPS packages are messages. Security implemented as firewalls, security as deep packet inspection, security implemented as anti-malware filters, and security implemented as improved coding methods are expressions of data-level security. These are all subservient expressions of semantic content security. Perimeter security is an incomplete subset of semantic content security; it works for some of the threats, but it is not inclusive of content-level data risks, tagging and metatag risks, or covert repurposing of hidden structures and communication channels. The threat environment is now all-inclusive.

[1384] If the issue raised is: "information is the subject or the object of the application, and the application is the verb and the verb is triggered by an event or process." The appropriate and correct response is that this assertion is merely a matter of context, specifically whether you run an information file or run an application file is just a transformation of ideas. That is a very small semantic point where the perspective is based on whether the data is the focus or whether the process acting on the data is the focus. It is really immaterial.

[1385] Data can invoke applications with metadata triggers, so for example, clicking on the filename of Word document or an Excel spreadsheet expresses the data against the contextual

association of the assigned "open with" designation. The traditional concept to run an application that processes data falls shorts on the complexities inherent in hierarchical data structures. See the white paper Metadata Security Risk Epidemic for a view of the transubstantiation of the Microsoft XPS format as a container both for applications and for data. Within human language, subjects, objects, and verbs are frequently identical words, such as "open," "work," "run," or "complete." This ambiguity is permanent.

[1386] Context in terms of usage and purpose defines the word usage. Likewise, both information and applications are collected, compiled, linked, cross-referenced, delivered, filed, and fundamentally directed and controlled by external workflow. A flow chart for each is much the same with moot differences. That it is just a matter of semantics. The critical concern is that computer code is an expression of information, and is nothing less and nothing more than information. The difference between DNA and RNA is matter of expression and context, just as the difference between information and an application is also a matter of expression and context.

[1387] This is not just an academic exercise or some theory. There are formidable security ramifications. There are real world examples for repurposing, aggregating, and expressing resources in unexpected ways, described later. Examples of repurposing, aggregating, and expressing resources in unexpected ways, include James Bond, the 9/11 attacks, and the unraveled Heathrow plans. The fictional Scaramonaga's golden gun was assembled from pen, lighter, and tobacco box in the James Bond thriller to assassinate people in public places. Even though port and vehicle security reacted to terrorist hijackings, the 9/11 attacks on the World Trade Centers and the Pentagon demonstrated the repurposing of vehicles into flying bombs. More topically, and quite a real threat with legal convictions, consider the recent British binary liquid explosive bomb threats at Heathrow Airport. The common and separately

innocuous components of Tang®, hydrogen peroxide, sports water bottles, a small blasting tail, a disposable flash camera, and duct tape have been demonstrated that when assembled covertly in a plane's bathroom create a formidable bomb.

[1388] By ignoring false distinctions between information and applications, we establish a security structure that is consistent and applicable for both information and applications. The same tools that establish security with sharing for data content establish security with sharing for applications. The traditional perimeter-based security focuses on the applications and the transmission of information. The advanced content guard and inbound or outbound scanner focuses on the semantic and semiotic information content. Instead, granular content control is agnostic to information, application, infrastructure, and ecosystem, just like DNA. This creates a consistent implementation regardless of platform, language, format, structure, purpose, metadata, etc. This means the solution is cross-platform, storage and transmission independent, and non-specific to processor, operating system, and network function.

[1389] Obfuscation in information format and structure or complications from metadata hierarchies and compound or complex information structures is no defense. While silos and stovepipes clearly appear as an obstruction to daily functional operations, it is an ideal opportunity for information warfare and economic damage from information theft. What is an obstacle generally becomes overlooked as a security risk because people mistakenly assume it must be trouble for everyone else too. Of course, it might be inconvenient to repurpose, but inconvenience is not a defense against information warfare and economic damage from information theft. It is not a viable security protection method. Hiding application, structure, data, purpose, and infrastructure is not a security solution.

[1390] The context of applicability is so different that this risk is often discounted. Silos,

stovepipes, and the apparent complexity of data separations do not create an effective perimeter defense, and in fact, a perimeter defense is not a realistic defense any more. Disparities in information storage, formats, file systems, structural storage techniques, dictionaries, data catalogs, metadata repositories, and so forth do not inhibit data repurposing or information warfare. The tools, widespread accessibility to data, human intelligence, with boundless human energy and virtually-free marginal computer time undermine obfuscation and data hiding.

AA.5.3.3 Granular Content

[1391] Information conveys simple data parts, thoughts, ideas, structures, formats, mechanisms, risks, and life itself as DNA. Only recently have we mapped the human genome and understand the gene as a chain of proteins which encode information and elaborate its anticipated application. See the white paper reference to DNA introns and exons with regard to inherent polymorphic risks of compound file structures and hierarchical metadata. This is reflected by a base coding of A and T vs. G and C, really a quarternary system in contrast to the binary, octal, or hexadecimal coding system used for information processing. The bases consist of proteins and molecules for chemical elements.

[1392] In effect, the data parts are the chemical elements, but the information is contained within various pooled layers of structure, format, aggregation, inference, and gene-to-gene interaction. Chemical reaction predispositions, pressures, temperatures, catalysts, and free radicals create the branching instructions for different results. Although these branch choices might seem random or esoteric, the situational states are as relevant as the conditional choices programmed into software. We might yet understand the complexity, but it is nevertheless present and instrumental. The same can be said of computer code with security flaws and buffer overflows triggered by circumstances unforeseen at design. One is a result of ignorance, the other also of ignorance and the lack of full situational

awareness. Data is not always passive and good-natured. Data not only is acted on, it is also reactive to situational parameters often unperceived as risk.

[1393] We have yet to understand all the inputs to make eyes blue or brown. We are also learning that disease is not an organism issue but rather a gene issue and even a single protein issue. So while we are unraveling the interpretation of life through the human genome data structure, we are also constructing diverse yet equivalently complex information structures for information processing. The first line of security in life is about perimeter control and gross defenses. The second line is about granular content control at the information level in terms of the aggregation of nucleic bounds and receptors. The parallels to information technology and security are far ranging.

[1394] Life is a process, but life is reflected as information. Consider the assertion that all computer applications and workflow are completely reflected by information. (We might not as yet know how to initiate the process of life solely from its information map -- the spark of life or the triggers to a wide range of processes that activate DNA into proteins, amino acids, and RNA -- but we do know how to reflect information as process and process that information.) Therefore, the structure of security is represented by the storage, process, and structure of information, the last being what is missing from information security and the focus of this paper.

[1395] Data pooling, information from collections, assemblage, hybrids, data mining, aggregation, inference, and data-to-data interaction undermines perimeter security. Data mining portrays the risk from preexisting and structurally known sources. Master data integration, a new catchword for mining data warehousing resources also should risks the specter of new security and privacy risks. Assemblage is the process of combining information in clear-cut patterns to establish risks. Hybrids are the improvised explosive devices of the information sharing and security world; individual pieces are

beyond recognition as a realistic threat until two, three, or more are combined. Examples include the previously described include James Bond and The Man with Golden Gun, the 9/11 attacks, and the Heathrow flight binary explosive bombing plans.

[1396] Reusing and repurposing information in a bewildering and unanticipated array of new contexts represents a formidable risk. It is irresponsibility to gripe that we cannot expect all thee variations or that our current tools and training is adequate for known security risks but not for these unanticipated ones. That is the mistake detailed in the white paper Street Light Syndrome, a situation were we conveniently apply the easy solutions we know to a perceived problem that is not consistent with the actual problem.

[1397] Aggregation within an unanticipated context is a covert channel. The sum of the parts within context reflects more than the sum of the parts taken alone. In fact, the risk is hidden, unanticipated, disguised, and often unrealized. The parts get through the perimeter boundary and can be assembled into accidental, incidental or quintessential threat. Furthermore, parts are frequently disguised with encryption or repurposed into an innocuous and limited form factor that slips past the perimeter defense. The sum of the parts is contextual and conceptual risk. This is hard problem that must be address for risk assessment and to implement an appropriate new security paradigm.

[1398] Context and concept is reflected by purpose and format. Information is presented in varied formats, and these formats convey meaning in addition to the content within. Format includes the encoding method, the language used, the presentation attributes, and structures. Our language for defining format is sufficiently imprecise so as to be unable to differentiate the information encoding methods, the characters used to represent the information, and from the language layered on top. For example, information is often reflected by a two-dimensional or three-dimensional barcode. A barcode

is a format. The information reflected by the barcode is generally encoded within a binary code mapping into either numerical symbols, a human language, audio, video, or enwrapped inside compound data structures. The ultimate mapped human language, for example Creole or French, is not likely to be obvious or even distinguishable from each other without derivation of the context.

[1399] The difference between Creole and French is not apparent but could be important. They are both rooted in European languages, but word or phrase meanings are different in each. Spanish and Portuguese are other European dialects overlaping both French and Creole, and can be mistaken for either. The differences within dialects are not always apparent. It is important in different markets. These differences go beyond spelling presentation and pronunciation and include idioms. For example, cojer (to take or get) is offensive in most Latin Spanish dialects, and some listeners will not be forgiving of the speaker.

[1400] Punctuation--Context alters Concept shows even the exact same words and picture could result in diametric interpretations with different outcomes. The specific information could be a key, index, or actual information with a standalone value, a contextual meaning, and some inferential meaning. A key or index indicates the actual information must be looked up within a database or list. The standalone value could represent a compound key, such as an UPC code, which both has meaning within many databases and lists but also more practically within a direct or immediate context. An example is a lot number, source location, destination, product counts, manufacturing date, expiration date, shipper, receiver, warnings, etc.

[1401] The issue of representation by barcode becomes abstracted one more step with technologies such as RFID for encoding barcodes into invisible radio waves. The RFID could be visible on the surface or hidden inside packaging (much like the XPS metadata hierarchy), yet easily readable

by contact or within a minimal distance with the correct energy frequency. The returned data could be binary or encrypted, or in extended form contain additional volumes of information in binary or encrypted forms.

[1402] Recognize that the information is nevertheless available although wrapped within many functional and semantic contexts, just like the DigitalDoors multitier hierarchical overlapping tag structure. The paradigm for information security must adapt to reflect effective security. The centrality of content, context, and concept is critical for establishing security and enable information sharing. Granular content control applies not only to the risk assessment and the protection of information content, but also to the structure of information too.

AA.5.34 Implementing of MS Office

[1403] Security flaws within Microsoft (MS) Office Suite result from ill-conceived data structures and because of the very integration that makes the Office so useful. Microsoft cannot repair these security flaws with bug fixes or security patches. Only a major overhaul by Microsoft of the suite and its constituent applications will eventually repair these flaws by altering the Office workflow and processes. However, practical and realistic solutions for risk-mitigation presented in this paper can be applied now. Neglect to address these flaws violates new privacy and security regulations and perhaps borders on malpractice.

[1404] These flaws are not the achievements of hackers and outsiders, although they can be exploited by hackers, competitors, adversaries, and data mining analysts. They result from fundamental design characteristic of all the MS Office Suites and each desktop productivity application separately. Every MS Office binary document contains confidential information. This ranges—from small amounts of information about authorship—to the editing history complete with deletions, reviewer comments,

file attributes, and source and routing information—to extraneous baggage from documents previously edited during the same session. The unanticipated delivery of such sensitive information represents a serious and credible risk through the loss of confidentiality, repudiation of privacy, breach of secrecy, and exposure to organizational sources and methods.

[1405] This paper shows how to skirt these flaws. It defines the inherent application security risks and demonstrates offsetting security methods. The positive focus is on document security and controlled presentation. While encryption is a partially effective solution, it is just a point solution even when extended by public key encryption (PKI), Kerberos, or digital signatures. Encryption of MS Office documents hides integral risks until the documents are actually viewed, printed, edited, or emailed. Encryption breaks most workflows that are the stated business goals for the Microsoft collaborative environment. Because of these security lapses, creating and implementing MS Office security, as explained in this paper, must be implemented through a multi-faceted change in behavior. It is also implemented by altered workflow process tailored to specific needs of each organization and attention to presentation formats used for distribution. This paper shows methods to secure MS Office documents despite these fundamental security design flaws.

[1406] A multi-faceted security workflow process becomes an issue over control of distribution by document type, recognition and categorization of all user content defined by security exons, removal of non-coding or non-activating security introns, preparation and distribution by clearance levels, content certification and accreditation (C&A) subject to conversion to primitive and certifiable file formats, distribution in print-representative-like packages, with guarded ingress and egress of Office files. Finally, implementation of security through granularity of MS Office node elements by analysis for inclusion and exclusion is a far more effective method, permitting collaboration within a multiple-usage infrastructure.

AA.5.4.1 Microsoft Office Suite and Applications

[1407] The preeminence of MS Office in terms of functionality arrived with Office 95. That version was sufficiently functional to perform most desktop office processing and management tasks, and derail competitive sales of WordPerfect, WordStar, 1-2-3, SuperCalc, Ventura Publisher, XyWrite, AmiPro, MultiMate, and so many others. The innovative object-oriented hierarchical data model first deployed with Office 95 is now fundamental to all versions of Word and all MS Office applications. Yet, this data model itself creates fundamental security flaws. Feature and functional advances since Office 95 clearly have value, but are increasingly aimed on workflow efficiency and integration. Microsoft Corporation markets Office as a platform for delivery of new services within a collaborative environment. While security is a stated strategic objective for Microsoft in terms of is Trustworthy Computing Initiative, a lack of fundamental security design and ill-conceived workflow processes within Office undermines this objective. As such, MS Office represents a critical but widely-used commercial off-the-shelf (COTS) platform with significant inherent risk because of workflow and object data model design flaws.

AA.5.4.2 Repress Thoughts of Banishing MS Office

[1408] Achieving Microsoft Office application security is significantly more involved than obvious. MS Office applications represent vulnerability risks at the file, operating system, process, and workflow levels. No single approach for security is sufficient. Banning MS Office applications and MS Windows does not organizationally, politically, operationally, or even economically represent a viable security formula. The use of MS Office applications is so widespread that any outright ban does not preclude delivery and reliance on these file formats and processes with any number of overt, covert, accidental, or engineered risks. In fact, alternatives include "work-alike" macro-language functionality

and file format support. The core security risks inherent with MS Office have been coded into other such products, as well as most other desktop productivity tools and COTS products. Work-alike competitors include Sun StarOffice, 602 Software Office Suite, WordPerfect, Lotus Notes with 1-2-3, and other OS-specific tools, such as MS WordPad delivered as an MS Windows applet.

[1409] Use of older technologies or a rollback to older technologies in order to improve security is professional sabotage and undermines the increased white-collar efficiencies observed with MS Office. It creates at best a false sense of security due to the pervasiveness of MS Office documents; you could avoid creating them, but you certainly will receive them and need to respond to them in kind. Furthermore, security solutions must also reflect the need for ongoing user support in products, usage, and processes. For example, Microsoft has specifically stated it will not issue an INCLUDETEXT patch for Word 97 because it is no longer a supported product. This reinforces the notion that addressing security flaws is a process with currently supported and evolving products rather than a collection of point fixes and patches for released products.

AA.5.4.3 Office Versions, Releases, and the Data Object Models

[1410] MS Office is a security risk because of the interaction among the MS Office applications and documents, the creation of metadata in binary document file formats, and the shift from one of results to that of a reentrant and ongoing process. Document data has expanded from simple linear files to complex object-oriented structures. MS documents are black holes. Basically, what goes into them at any point usually stays there. Additions, deletions, system information, redlining, reviewer comments, and routing become indelible parts of each document.

[1411] Many different versions of MS Windows, server extensions, and many releases of MS Office or its constituents complicate security. Application features, bug fixes, security patches, and 3rd

party add-ins complicate the nightmare when assessing and ascertaining the exact composition of the MS Office environment. Client-based applications, such as InfoPath, Outlook, Outlook Express, Internet Explorer, the various scripting languages, plus server-based applications including Exchange, SharePoint Server, Net Meeting and Live Meeting Whiteboard, Live Communications Server enhance the collaborative physical coverage of MS Office but also correspondingly increase security and privacy risks.

[1412] The MS Office document is forwards and backwards compatible across MS Office releases. This means that Office 95 can open and alter Office 2003 documents, and Office 95 can open and alter Office 2003 documents. However, "dead" internal structures are defined in obsolescence and new structures have been added to the newer versions. Cut and paste among the Office applications adds non-native structures too. Therefore, results from file conversion, raw data, metadata, links, macro code, and structural elements can be hidden accidentally or purposefully. It also possible for a sophisticated user to create new (see the sidebar on security introns) and undefined covert structures (see the sidebar on security exons) ignored by all extant MS Office versions and tools, visible or activated only by complex steps, since MS Office does not validate the integrity and applicability of internal structures within a document binary file.

[1413] Security that is part of MS Windows or MS Office, such as user passwords, file passwords, password-protected databases, file encryption, and range protections are not sufficiently effective. Even the Windows Encrypted File System (EFT) and Active Directory (AD) merely postpone inherent risk until files are distributed outside the perimeter of the encryption system. Other methods defeat overt security. A large number of tools recover passwords and unlock encrypted MS Office files, zipped files, adobe PDF distributions, or reverse engineer and open FAT, FAT32, and NTFS files. An

Internet search with a browser or a file sharing program will uncover any number of freeware, shareware, trial, commercial, and pirated tools to do just this. Furthermore, collaborative sharing of an MS Office file requires that any such file passwords be divulged to open the files; this exposes all the hidden structures, metadata, and security risks inherent in the document or referenced resources.

[1414] It is important to recognize that there are many file types and document structures associated with MS Office, specifically defined by the formal MS Office documentation at msdn.microsoft.com but also those shared with other MS Windows applets and competing products. Each MS Office application, such as Word or Excel, create file binaries or binary files with different object structures but interchangeably read/write and import/export each other's file types, embed portions as formatted text or complete objects, or link through remote procedure calls to these other file types. These object model structures are generically called the Document Object Model (DOM). The DOM is another term for an object-oriented data storage package.

[1415] The purpose for the DOM with hierarchical storage of metadata is three-fold. First, it is useful for backwards and forwards version compatibility. Second, metadata extends the document creation session from one-time event into an ongoing revisional process. Third, metadata provides order and structure otherwise notoriously difficult for inherently free-form and flexible documents.

[1416] Metadata provides backwards and forwards version compatibility, a problem that plagued the software market of the 1980s as upgrades were frequent and disruptive. This is specifically missing with Access and its .MDB table space/workspace metaphor. Frequently, software upgrades included old data upgrade routines to convert old formats to new. This was both risky and prevented reversion to the older software version once the converted data was used in the newer application. Metadata provides the necessary blueprint, format, and structure retention so documents can be revised in future editing

sessions. Try creating a complex document in Notepad, which is a plain-text editor, to understand the desirability of maintaining other channels with document information. It is just that these other channels with document information are packaged in the same file binaries for all Office applications. Consider how difficult it could be to reset type characteristics every time you reopen a document. This information is part of the Office metadata, although style sheets and schemas maintained in a different storage channel are valuable in HTML and XML and might aid the future transition to a secure MS Office.

[1417] It is incorrect to assume a static basis for any MS Office application document structure, as a monolithic MS DOS-based file, or as an in-memory object. For example, the Excel DOM can be embedded inside a Word DOM, which selectively can then be pasted as a formatted object into a PowerPoint presentation. Because of this workflow, simple tools and methods will not eliminate the security risk. It is not just a Word problem; law offices using mostly Word probably represent the simplest security exposure. That is the exception, of course. In general, you have to address the security through each DOM individually.

[1418] Each version of MS Office supports different object models, each application with the suite has a different base object model. In other words, while some features in Word 95 are still supported in Word 2003, other features in Word 95 might have atrophied and are no longer supported in the same way or even not at all. In addition, Word 2003 has wholly new features and corresponding extensions to the object model not recognized by Word 98. This demonstrates that metadata is version-specific and hidden when another version is upgraded with a newer one. Another concern that arises in almost every Office document is imports, pastes, and OLE imbedding of other Office documents and aspects of the object model corresponding to that application type. For example, a base Word document with a spreadsheet and Project waterfall chart now includes editable components referencing a different

Office applications with data in a structure referenced by that corresponding application object model, in this case Word, Excel, and Project.

[1419] Not all the MS Office risk vectors can be explored with Notepad. Most of the metadata is not visible ASCII text but rather encoded binary data and complex structures. Locate a hexadecimal (binary) editor at www.sf-soft.com or another forensic tool web site. The utility of forensic tools is critical to security success because of the complexities of desktop and server workflow, and also because of the document structure itself. Binary pointers list the locations of document node elements and other pasted or embedded data structures. Although WinHex is useful to demonstrate several hidden security flaws in MS Word, the simple hex editor only reveals the content of a simple DOS file or Word file as a monolithic storage unit. In reality, that DOS file is backed up, replicated, written, rewritten, and stored in duplicated extents throughout machine RAM, system buffers, and disk blocks and sectors. MS Word "fast saves," versioning and plain backups create a melange of risk vectors that transcend this paper, but are nonetheless relevant to anyone assessing system, MS Windows desktop, networking, and network neighborhood access control and security issues. Security really is a metaphorical ice field, and what you do not see and are unaware of can be catastrophic.

[1420] Microsoft is aware of these flaws and has published these thirteen categories of dirty metadata:

Table of dirty meta data

Name

Initials4

Organization name

Name of originating computer (desktop)

Name of network server and/or hard drive

File properties and summary information

Non-visible embedded documents

Names of previous authors

Document revisions

Document versions

Template

Hidden text

Author comments

[1421] Some of this metadata is accessible through the Office application menu interface through menus and dialog boxes. There are also the document file properties exposed by the Tools/Options pull down menu and the User Information tab, this does not mean that this dirty metadata can be removed through menus and dialog boxes. Some of the metadata persists indefinitely.

[1422] This is not the complete list of metadata. There are other categories of revealing metadata also known to create security risks but not fully disclosed by Microsoft. Consider reviewer comments and redlining workflow. This often includes embarrassing suggestions and the routing of the reviewed document. Other visible metadata with confidentiality risk include:

Footnotes

Cross-references

Table of Contents tags

Indexing tags

Hyperlinks

Smart tags

[1423] Expect x-link and x-pointers plus style sheets and schemas within documents saved in the XML format. In addition, other undocumented structures are part of the extended and expanding Office document object models. Consider fields and mail-merge fields, which are markers for information automatically inserted by Office or by a user when opening, saving, printing, or emailing documents. These fields create a built-in facility for careless information disclosure or overt hacking. There are also the document file properties exposed by the File/Properties pull down menu. This includes:

File/properties

General

Summary

Statistics

Contents

Custom

[1424] Other security risks are not specific to MS Office. The techniques for information camouflage are equally valid in most any desktop application, and are most relevant to presentation output rather than binary file delivery. Information camouflage includes text set to small font sizes, such as 0 or 1, fonts set to type unlikely to be installed on the system which map to symbols or line drawing, PostScript or Unicode font sets with alternate encoding, and font color set to match the paper color or an applied background. White font on white paper hides text, black font on a black border or shading hides text too. Text can also be hidden with graphics when the graphics are anchored to a specific location congruent with the text. Color games with text and graphics also hides the text.

[1425] Macros, VBA codes, VBA add-ins, and applets also represent a security risk. Anything than anyone can imagine as an application can run from within MS Office, productive or destructive. Usually, these bits of code are stored as part of the document metadata. However, they also can be out-of-channel files. Either way, they can be compromised by a new code that overwrites the original. They also can be inserted through fields, formulae, or menu add-ins. Collaborative tools are the most obvious entrée, but Windows security flaws also provide some interesting opportunities for Office security exploits. Code flaws affecting MS Office are beyond the scope of this paper, but are explored in the book.

[1426] New features in Windows and other Microsoft digital rights management (DRM) applications, such as ORAPI, ADSI, and MS IRM provide for collaboration, resiliency, and complex versioning and backup far beyond the capabilities of MS Office. As such, many new security risks are exposed; implementation of MS Office security within a collaborative environment is beyond the scope of this paper.

AA.5.4.4 Content Security

[1427] The differentiation of content within an MS Office document based on initial owner and target distribution is important for information sharing with coalition or business partners. Some content will be strategic, some tactical, and other content can be downgraded by censorship of information such that only target parties in-the-know can understand the context. This is accomplished by downgrading the content with a publishing format change, element removal and exporting within a new provably-secure format. Downgrading is a process well-known to the military, anyone who prepares of documents for release under the Freedom of Information Act (FOIA), paralegals who are in the known and deliver case information to the courts, and anyone censoring privileged or trade secret information from

distributed documents and email.

[1428] For example, faxing a Word document to a legal adversary is acceptable since the image is a controlled and published representation. However, delivery by WinFAX delivery of the editable binary file is unacceptable. WinFax integrates easily with MS Office and has that file delivery capability, which should be avoided for security reasons. As another example, if legal eFiling rules necessitate delivery of a document within a binary document format, the MS Office document can be created as a printed paper or file (Print to Text, Print to PCL, or Print to PS MS Windows driver options), output to a TIF image, print to an Adobe PDF file, or exported through a filtering sentinel as an ASCII test file. Note that the MS Rich Text Format (RTF) is not suitable because the RTF format also includes metadata. If font, table, and presentational format must be preserved—since ASCII does not support that—any image output is a good choice. However, be aware that postscript, Adobe Acrobat, and even images can be reconverted to a formatted binary document with conversion tools and optical character recognition; all metadata, edits, redlining, versioning, and workflow will be not be recovered, of course, which is the essence of locating MS Office security flaws and implementing security.

[1429] Content of MS Office documents transcends the actual presentation as a printed page, slide, spreadsheet, database report, email message, an index of documents, UML: or project waterfall, or organization chart. Microsoft Corporation is positioning Office as a platform for delivery of new services; it is not just about a PowerPoint presentation or a Word document printed to a facsimile. The DOM is a project plan, with a structure, with components that do things and are sensitive of their own.

[1430] Recognize that MS Office security is also a cross-platform issue. Inclusions could be aimed at Macintosh, Unix, Linux, or other operating systems and even other document applications. Delivery of any MS Office document can represent a security on egress by containing proprietary data

and functions or by ingress as a carrier for a virus or Trojan virus. Even Outlook email with its potential for rich-text formatting, HTML or XML content, links, inserts, and file attachments carries the entire MS Office risk with it to wherever and on whatever platform it is received. For example, the MS Office document could include an attack on a Linux-based SendMail server or client. While metadata and redlining contain sensitive data, when integrated with webDAV interchange, InfoShare, Exchange, and other collaborative environments, they also contain workflow and traffic content which can be equally sensitive. For these reasons, it is important to explore the MS Office DOM risk factors:

Content classification

Tagging

Clearance level

Data mining

Traffic analysis

Inference

Encryption

Digital Signature

Document access linked to Fortezza, PC Crypto cards, smartcards, and n-factor authentication

Granularity

Strategic info

Tactical information

Common Criteria or NIST analysis

Covert channels

Bell-LaPadula model conformance

[1431] Content classification occurs with tagging for formatting with bold, indexing, and paragraph marking, explicit element tagging for HTML and XML or database and spreadsheet table, field, ranges, row, and column designations, as well as authorship techniques, such as "...describes the formal issues of security introns in the next section... ." Formulae and macros define ranges with informational content, as well as indicate purpose and intent of the process as well as the target data. When content is tagged at the sideline, as in "eyes-only," or within-the text with any label name for clearance level, as in "<1>," this attests to a security level with an importance that exposes security lapses.

[1432] Although MS Office 95 reached the utilitarian level of adequate functionality, the new features of MS Office and the inclusion of photographic manipulation, pixel editing, vector graphics, charting, data sorting, Find and Replace, indexing, tagging, smart tags, links, and collaborative integration through such as OneNote, InfoShare, Outlook, and Exchange expose the MS Office documents file store individually and in aggregate to data mining techniques. For example, a subtotal of employee salaries within a pro forma business plan matched against a list of employee names compared to a bank check ledger gives away each employee's salary level; each document in isolation does not give away information until several are merged and analyzed together. Direct analysis through record relationships and sorting is one type of data mining, human intelligence through inference or statistical inference with set theory or Bayesian methods is yet another. For example, because you know that 6 employees are traveling to a conference in D.C. and two others are not in the office, you can approach a particular person who by inference is manning the station desk with a very specific social engineering attack.

[1433] OneNote, InfoShare, Net Meeting and/or Live Meeting, Outlook, and Exchange with MS

Project also enable workflow routing, group editing, and acceptance signoff. This information becomes part of the document metadata so that traffic analysis shows where the document originated, what changes were made and by whom, how it was routed by username, network, and IP address, who has seen it and has access to it, and all process flow and comments. One of the secure prizes of organization information thus unintentionally published is the names of people within the organization and functional roles.

[1434] Encryption, digital certificates, digital signatures, biometrics, and USB or other hardware Fortezza access devices bind into workflows, access to applications, and access to specific files. For the most part this represents and all-or-nothing security. An encrypted file means you cannot access it until it is decrypted; since MS Office files are non-linear, partial decryption is more likely to prevent it from being opened by any MS Office application. Once the key is provided, the cat is out of the bag. If multiple users get the same key, it is likely that key will float around freely. Encrypting a document multiple times for each user intended to access it is a workflow nightmare. Furthermore, encryption packaging does nothing to provide egress or ingress security, or handle the granularity issue. Encryption is effective at a low level or when combined with the other methods described in this paper.

[1435] Security through granularity of MS Office node elements by analysis for inclusion and exclusion is a far more effective method. Multiple source documents create structure and semiotic meaning not in evidence with subsets. This process breaks the context to prevent useful data mining, routing inferences, and the more powerful semiotic information methods. It allows for the separation of strategic information from the tactical, so that access is granular by role, user, and other discriminators.

[1436] Many academic and implemented security models are in use today, both as a straw man and for certification processes. This includes the Common Criteria, NIST certification, and the Bell-

LaPadula security conformance model. It is well that you know about them, but for the most part they do not explain or provide insight into how to protect MS Office documents. These models assert the need for air gaps between organizations with different security levels, but do not provide a means for information sharing as legislated by the 2001 Homeland Security Act or normal organizational collaboration or data processing workflows. While they do address the potential for covert channels (insertion of content in alternate formats or encoding) and how to protect against them, the methods are not effective except at a very superficial level. If you review the "covert channel information" you should understand the difficulty in maintaining security to thwart virus transmission. Instead, MS Office security must be implemented at an intron level, as described in the next section.

AA.5. 4.5 Implementing Document Protection

[1437] Several steps are prudent to enable MS Office document protection. The first step is have a network guard that filters all incoming and outgoing traffic for MS Office document files and quarantines them. Spam and virus filtering is necessary to preclude system, resource, and file exploits. URL filtering, quarantine lists, black lists, white lists represent the minimum responsible approach. Ingress files can harbor viruses, etc. Outgress files can harbor privileged information at any and all levels of the DOM. With a means to filter and check every node for purpose, content, metadata, formats, structure, comments, links, and so on, there is no other way to vet the integrity of the file.

[1438] It is insufficient if not impossible to remove metadata. Removing metadata from the files binaries often irrevocably corrupts files—so that they will not print, save, or be in any way recoverable with the automatic corrupted file recovery tools. The only complete answer is to walk the object model and assess the purpose and content of each substructure, inclusion, or node element at a granularly content and functional review. By the way, this makes it possible to vet a document as provably secure.

The document can then be published, exported, or recreated within a new context suitable to the organization workflow and security needs.

[1439] MS Office is not the only application to rely on a document object model. Most other modern desktop applications utilize the same backward and forward extensible structure, but characteristically create similar security risks. The installation object content for Adobe Acrobat, just the authoring tool becomes a visible part of every mastered Acrobat file even though it might not be any part of the source MS Office document exported as an Acrobat package.

[1440] A number of well-meaning efforts available to scrub MS Office documents represent partial solutions at best and a false sense of security at worst. Microsoft posts a Knowledge Base article on metadata (MSKB Q 237361), the help note, "Get rid of tracked changes and comments, once and for all" and a commercial product called Metadata Assistant automates these generic processes. The promise is not a full solution; it is partial at best. These solutions might get the metadata in the master document, but not find the hyperlinks, the subdocuments, or confidential information still in plain view. Since a typical document is compounded from more than just Word and Excel, and often is part of a process, as in Outlook e-mail with document attachments, scrubbing has to get each piece and walk the object model for content.

[1441] However, that said, this is the current collection of methods for offsetting MS Office security flaws.

Start with a clean template

1. Write-protect templates
2. Attached template(s) or styles for other MS Office documents
3. Normal .DOT

4. Clean up .DOT

Edit cleanly

1.Disable Versioning

2. Disable Change Tracking

3. Periodically "Accept Changes" to purge change log and save or save as

4. Disable Fast Save

5. Patches for "Unwanted Data"

6. Do use comments, not hidden text

7. Do not use footnotes, end notes, table of contents, index, links,

Remove References

URL

Hyperlinks

Pointers

References

hidden text, headers, footers, footnotes, endnotes, tables of contents, index, links, can establish a context or cerate a semiotic inference to other documents or sources

Paste... do not embed

Publish... do not send a file

Print

Fax as image (not as binary document in WinFax or eFax, etc)

[1442] Start with a clean .DOT template (Word) or whatever the new document type is for the specific application. The Normal.DOT or New Spreadsheet.XLS on the MS Office distribution CD is

safe. In Visio, for example, start with a new chart. Older documents saved as templates tend to accumulate unanticipated metadata. If you must use a non-standard template, clean it up. View it as both a printed document, as a binary file, and as a document object. Write protect the templates, or better yet, store the templates as non-modifiable networked volume. If the templates are used throughout the organization, create a network store for them and configure each user's installation of MS Office to look there for the templates. Search the network for duplicate copies of these templates and eliminate them.

[1443] If you change from any version of MS Office to another version—this can be a regularly upgrade or a even a downgrade—create new documents and cut-and-paste parts of prior documents into new ones. Lose the older files and templates. If you receive or open an Office document in either an older or newer version, create new documents and cut-and-paste parts of prior documents into new ones consistent with the MS Office version that you use.

[1444] Edit cleanly. This is not a single step but rather a process, both one time and ongoing. Realize that the MS Office risk does not go away, so security remains a daily and ongoing battle. Disable versioning to prevent a buildup of past versions of the document. With versioning, prior sessions will be stored as document.doc 1, document.doc 2, and so on. These tend to get entwined with the latest version. If workflow with InfoPath, Outlook, or other collaborative workflow tools creates duplicate copies of the source document file for each user, then configure the system to store a single network copy instead. Preserve an audit trail and backup with a system backup rather than versioning. Disable change tracking to curtail the embarrassing buildup of additions, deletions, and changes that transcend the publishing intent of the document. If this redlining is necessary, establish guidelines—and stick to them—for periodically accepting changes to purge change log. Use the command to save the document without the change log with File/Save or File/Save As.... Do not rely on fast saves, timed saves, or file

recovery after a MS Office crash to purge the dirty metadata. After a crash, reopen the document, save the document under a new name. Close the Office application. Delete the old file precipitating the crash. Rename the new file under the old name. Reopen the Office application.

[1445] Make a conscious and active decision to install or ignore patches. There are a number of Office updates, usually called "service releases," that fix bugs and security flaws. However, patches will not repair the fundamental security flaws within MS Office. The last substantial security patch was for Word 98 for Macintosh called the "patch for "unwanted data" that fixed the situation where data and metadata from one document window got into other documents opened in other windows through a multiple document interface or sequentially during the same session.

[1446] Use comments instead of hidden text. It is documented as a feature so it can be found rather than accidentally uncovered. Hidden text with a font color change or font size change looks like an artifact that most users will ignore or delete. It can be discovered with Edit/Find or Edit/Find and Replace after pressing the More button. Specify font settings of 0, 1, or 2, or font colors of white, black, etc. in order to locate materials so hidden. It is not easy, but you could automate this process with a Word macro or VBA add-in.

[1447] Avoid the use of headers, footers, footnotes, endnotes, inserts for table of contents, index and the like. These appear only in the printed output unless specifically viewed from the View pulldown menu. Such links also create a lot of dirty metadata beyond what is visible even during editing that persists until changes are accepted.

[1448] Remove references. This is subtle, but very important when documents are specifically posted or even inadvertently listed on web sites. References include other files, documents, hyperlinks, and other possible embedded formatted materials. These references create the ability to infer quite a lot

about the purpose of the document from other related documents in the same directory, by the same authors, and the types of other documents. For example, a spreadsheet stored with a report that is not overtly included in the report suggests that is source material that has not been reviewed with an eye towards privacy, security, or client privilege.

[1449] Paste images, cut text, formatted text, pie charts, record sets, slides, waterfalls, milestones, organizational charts as plain text or an image rather than formatted Office objects. When you embed or paste as an object, all the metadata baggage from the other Office application is now added to the metadata in the target document. Since that metadata baggage is not native to the target document application, it is inaccessible and truly hidden. Tools, such as Metadata Assistant will not find Excel metadata within a Word Document, Word metadata within an Excel spreadsheet, and none of them within an Outlook note or message.

[1450] Publish your presentation. Do not send a binary document. Although "Publish" is a new feature to Office 2003, the intent is more important. Publish your documents in every MS Office application whether Office 2003 or prior. The process for publishing is to create a new distribution and delivery format limited to the visual presentation elements containing none of the blueprint and none of the metadata. In simpler words, create a new output version of the document. Print it on paper. Print it to a file. Print it to an image. In some way, alter the format to preclude the DOM with the document blueprint and its metadata. You do not need the blueprint for a presentation; the presentation is the final product. If you need to send a presentation for revisions, regulation review, or continued workflow, change the binary document format. Printed documents can be captured formatted by optical character recognition, such as with OmniPage. That is fine. Text files can be read by Word. Spreadsheets as text files can be imported and formatted by Excel. But, do not send the original binary file. Do not convert

Office documents into .RTF files as these still retain metadata. Create comma-delimited files, if you must. If you want formatting retained for revisions, flush all pending revisions, copy the document in pieces, paste into a new document, and save it as a new document. The proper concept is to recreate the MS Office document in a "format-neutral" file to remove blueprint and metadata.

[1451] Unfortunately, creating neutral formats is a very difficult task unless matched to the organizational workflow processes. Saving a Word document through File/Save As... as a text document is format-neutral, however, you will discover that the document has no fonts, headers, and style whatsoever. You will also discover line feed problems disrupt it and make reading it a chore. This is part of a content lifecycle issue for an organization. You would need to know how the documents are used and why and where; workflow could well dictate XML with a style sheet as a better format-neutral deliverable. It all depends on specific organizational needs.

[1452] Be aware that distributing your document, with such as WinFax or Efax, or using a remote control program, such as Laplink and Metaframe, to a user with the same application is likely to receive the file and not the presentation image. Be aware that document attachments in Outlook email or routing documents through InfoPath moves the binary document (or a complete copy of it) and not its presentation image. The use of Net Meeting or Live Meeting when sharing the desktop reveals Office metadata and has the functionality to transfer files among participants. Either way, publish presentation images in advance to prevent that security lapse and use the presentation materials only. If you must edit the sources, do so only after the collaboration is terminated.

[1453] Publishing Office documents to HTML, XML, PostScript, PCL, or Adobe Acrobat is not as safe as creating a presentation image. HTML retains DOM structure and metadata as markup tags and comments. XML retains DOM metadata as markup tags and comments, with formatting and structure

as a resource definition file or style sheet. Security is dependent on the PostScript or PCL driver, although usually only elements that appear in printed output are converted to these files. However, font size 0 and font 1, black text on a black background, white text on plain paper, and hidden text are still rendered; this creates a security risk when the print file is viewed but remain unseen when printed or viewed. Acrobat uses its own DOM and creates presentation files with a combination of metadata from the source document and its own metadata from application installation. Although some output settings preclude printing, encrypt the output, and restrict ranges, be aware that many tools can recover Adobe passwords and restore Acrobat documents into editable MS Office binary documents.

[1454] Finally, content security through granularity of MS Office node elements by analysis for inclusion and exclusion is a far more effective method, permitting collaboration within a multiple usage infrastructure. This technology exceeds the scope of this paper, but is detailed with sample code and applications in the book. The process is to walk through each node of the object model, match the node to a Office DOM, match the node to a function or feature, and capture the contained information for granular processing. Unmatched nodes are introns, and are removed and quarantined. Matched nodes are exons; the purpose in terms of the presentation image must be established, or they too are removed and quarantined. Nodes that are empty, old, questionable, outdated, or irrelevant to the application and version are also removed and quarantined.

[1455] The information captured for granular processing is then viewed by black and white lists for inclusion in the presentation image. Although the security steps detailed in this paper are effective in maintaining security, confidentiality, privacy, and integrity with MS Office, the only method to counter Zero-Day exploits is granular control at the contact level, applying the attack methods as a level of analysis to locate security risks. Sensitive, private, privileged, confidential, or information with even

an inferential basis is purged from the presentation image. This last step is a complex process given advances in data mining, set theory, statistical polling, Bayesian inference, and semiotic methods to expose meaning beyond the obvious, intended, or that possible when information is combined with other sources.

AA.5.4.6 Conclusion

[1456] This paper defined the inherent MS Office application security risks and demonstrated methods to implement security within the MS Office document structures. Creating and implementing MS Office security is therefore expressed as a process that controls distribution by document type, with removal of non-coding or non-activating security introns in the document, conversion to primitive and certifiable file formats, distribution in print-representative-like packages, with guarded ingress and egress of MS Office documents. Although the focus is on document security and controlled presentation, this is not the whole security issue for MS Office, but it is a substantial first step toward what is perhaps the most difficult issue in COTS security control.

[1457] A multi-faceted security workflow process becomes an issue over control of distribution by document type, recognition and categorization of all security exons, removal of non-coding or non-activating security introns, preparation and distribution by clearance levels, content classification and data mining, content certification and accreditation (C&A) subject to conversion to primitive and certifiable file formats, distribution in print-representative-like packages, with guarded ingress and egress of Office files. Finally, Implementation of security through granularity of MS Office node elements by analysis for inclusion and exclusion is a far more effective method, permitting collaboration within a multiple usage infrastructure.

AA.5.5.0 The contextual information matrix

[1458] Sensitive content, and in particular, personally identifiable information (PII), does not exist in a vacuum. It exists within a container (e.g. framework or structure) defining its shape, size, format, purpose, relationship, ownership, integrity, acquisition technique, half-life, disposition, application, attributes, and properties. This container is both relative and subjective, as well as situational. This data container is sometimes explicit when rigorous definitions delineate the content, but sometimes only implicit when tied to context by application and purpose; it is determined by data structures, XML schemas, field formats, tags, memos, comments, or notations, or by the matching applications built to access the specific data. It is modified by a function, framework, network, service environment, and workflow. This container is an information matrix that needs security itself to protect security and integrity of sensitive and PII content.

[1459] The information matrix is evolving past simple prefabricated record structures, grids, database tables, and linear or tabular formats, and all other such simple formats into a hierarchical structure not unlike DNA. Extraction of sensitive data, and in particular PII, from the information matrix is only a partial security solution. Context must also be addressed, including the container and relationships. Consider the following:

[1460] Even if the container were empty, you would infer the missing content. It is not sufficient to extract the content without addressing context. The specificity of the container defines the likely fresh egg content. The container -- alternately -- could contain chocolate in egg shapes or bunny rabbits but only with intended inference to fresh eggs. The container could be recycled with nuts and bolts, but also only in reference as a recycled container of eggs. Situational placement is also relevant to context and content. A box in a grocery store or refrigerator implies eggs, a box in a candy store or coffee table implies candy, while a box in a workshop implies a sorter for loose parts. Any printed properties and

attributes merely modify the quality of the content or adjust the perception of the content quality.

[1461] For example, farm source, brand name, size all convey integrity, quality, ownership, and intended application. For another example, a clear plastic matrix indicates some enhanced visual value supposedly obvious. Just as contextual packaging conveys meaning about the contents, packaging effects the knowledge of data content too.

[1462] The contextual information matrix conveys knowledge about the content -- to a greater or lesser degree. The risk is to presume the lesser degree with a corresponding resulting failure to provide adequate security to sensitive data or privacy to PII. The failure to repackage the container itself or redact the context properties and attributes that define it compromises any and all security. Although perfect manual re-identification is unlikely, statistical methods yield functional results. Just as it is easy to look around and find the eggs or emptied shells to repopulate the container, forced re-identification with content into the information matrix by size, format, and index emphasizes the requirement to repackage the extracted content, context, concept, attributes and properties. Where security of sensitive data or privacy of PII is required, the information matrix represents a risk through aggregation, inference, and interaction that requires the same level of attention as the sensitive or PII content within. As a result, the content container is an information matrix that needs security itself to protect the security of the sensitive content and the integrity of PII content

AA.5.5.1 Future direction for information awareness

[1463] The Pentagon's aborted Total Information Awareness program was aptly named, although its termination in no way safeguards anyone from the premise of data mining, aggregation, and inference. Other programs surely exist today and will exist tomorrow. Mistakes in data stewardship will be revealed, as they have in the past, showing that exploitation has given way to uncontrolled information

irradiation. We are in a world approaching total information awareness in that fundamental data about people, organizations, processes, and information triggers is public knowledge or can be readily bought. Total information awareness is not so much about a government activity gone out of control—although government was the first to exploit the information glut—but a social, cultural, business, and control environment without inherent safeguards and sufficient signification paradigms.

[1464] The unique, limited, and single-purpose data stores of commerce and government of years past have become common, networked, and imitated just about everywhere. The danger is not from any single protected database but every single one of the public, private, commercial, and governmental resources which are out there. The data sources need not be electronic, and in some cases paper-based archives mask the availability and potentiality of the contents. Single-source paper archives are ripe for exploitation because authentication is physically impossible and its single focus precludes the force of authentication mass. Furthermore, total information awareness does not have a lock on fact or truth, but also can nourish distortion and complete fabrication. It is as easy to propagate lies as truth in order to obscure any difference between the two.

[1465] Small pieces of information, such as name, social security number, bank accounts, credit card accounts, address, phone number, genealogy, business relationships, fingerprints, biological data, and most of the facts used for knowledge-based identification (i.e. multifactor identification) are available from a multiplicity of sources. List brokers provide exact information. When exact information is not available, information can be inferred or statistically aggregated for precision. In many cases, accuracy is not even necessary because partial overlap of outdated information and prior addresses is sufficient for most vetting, identification, authentication, and authorization operations. Taken a step further, accuracy can be phony—built from a cut-and-paste montage of sound and video and data

records. "Frankenbiting" is the process used to create a semblance of a story with true aspects. It is a reference to the creation of Frankenstein's monster from a collection of incoherence parts to create a seeming plausibility but out-of-control existence. The use of frankenbiting is often made in reference to the lack of reality in "Reality TV." It can be just as effective to lift a real identity, subvert the values of an identity, or invent a fake persona with elements of belief for acceptance. The canvas for fraud is not just art, jewelry, branded merchandise, but clearly now information itself. The message unit containing "He denied he said he was negligent" also contains "He said he was negligent" and "he was negligent."

[1466] Information necessary and sufficient for identity theft does not need to be available from a single source, but can be accumulated to create the necessary profile. This is true for personal, business, organizational, and government activities. Identity theft, per se, is not specifically the province of the individual personal identity; it can apply to any other physical or logical entity, such a owner of record of a building, the corporation, the ruling class, or a shadow controller. It is as conceivable to offload a ship—container, cargo, or oil tanker—at a misdirected port of call by electronically rerouting it and reasserting a fraudulent certificate of carriage and manifest in the big time without raising suspicions of captain and crew or port operations just as it is to steal a credit card, bank account, persona, or real property of buyers, sellers, bankers, brokers, lenders, and service organizations.

[1467] This discussion goes to the nature of information. Information is word with too many definitions to be comprehensive. Suffice it to say that information is data, a message unit, aggregation of message units, content, symbols, meaning, signification of meaning, knowledge, a denotation, a connotation, a designation, an indication, a presumption, etiology, or an interpretation. None of these definitions are correct or sufficient. We have a built-in bias that information is necessarily truthful,

actual, correct, legitimate. That bias is logically false. In fact the presumption of legitimacy undermines much of the stock we place into knowledge and information processing.

Some of these "human" definitions for information are useful, but we require a more rigorous framework to proceed. We need to explicitly define terms and expand the framework of information from a collection of message units into a working system of information items with a mathematical basis that include a contextual relationship, a conceptual basis, a structure for functional transformation, and a delivery system. We talk about information contained in databases, but information is not confined, but rather temporary, transient, transitory, and in perpetual transit and transition.

AA.5.7.0 Unstructured data

[1468] The issue of structure in data frequently arises when describing unstructured or structured database schemas. Structure is basically a taxonomy or ontology useful for ordering and describing the format, purpose, tools employed to access or process data, and process of data. Realistically, any simple or sensible definitions are relative because all data is structured, although that structure might not be evident. Lack of obvious structure is not proof for lack of structure. Structure can be in-channel or out-of-channel, just as information tagging can be in-channel or out-of-channel. The location of the tagging (or metadata) does not take from the tagging structure. It is really a context issue. Failure to address this undervalues information, the need for security, and the appropriate application of information for significant financial returns.

[1469] Structure is also amorphous and transitional. Consider z-ordered databases where slices are extracted or SQL is used to create records or transactional sets. The structure is a simple model of complexity, and utility is a simplification for practical application. The structure blueprint is itself a data set, in fact a metadata of the data itself. This metadata is an abstraction, itself subject to amorphous and

transitional representations. Thus, while data is relative, its structure too is also relative. As these representations are transformed, the meaning, syntax, purpose, utility, and value of the data stream changes too.

[1470] In fact, the issue of structure perhaps oversimplifies how obvious the structure really is and how irrelant to most issues the question of structure is in terms of assessing content, context, or concept. Unfortunately, that underlies the problems of information, security, and the value of information as an asset in the classical production model.

[1471] The real issue is about discovering that structure. This means discovering the utility of the information within different known, probable, and possible contexts and concepts. Some of these are likely to increase the perceived value of the information or provide repurposing opportunities with PII exposure. In effect, structure, per se, is a non-issue except as it relates to ease of workflow or assessing the risk inherent in information when implementing a security or privacy system.

AA.5.8.0 Information measurement and management

[1472] Information measurement is about assessing utility value and subsequent risk from misuse. Information has different value to different parties. Risk is the application of that value against intended utility. We can measure information by many dimensions. This includes:

Table: Dimensions for measuring information.

size

length

density

meaning

utility

ease of access

format

structure

integrity

authentication

clarity

precision

distinctiveness

comprehensibility

risk

loss

cost to obtain

cost to remake if lost

cost to store

availability elsewhere

carrying costs

auction value

value to another party

value to another party if and only if the other party retrieves it surreptitiously

value if a close secret

value if widely known

value by general sale to anybody that chooses to acquire a copy

[1473] Value is neither inclusive nor exclusive. It is certainly not definitive. Value changes with time, location, accessibility, context, and conceptual utility. Information has different value if it is a unique distribution, if it is a near copy, if is a facsimile copy, if it is an exact copy, if it is a digital copy, or a representative copy. We can measure the value of information to different parties in terms of the money lost, the lives it costs, or the disruption to normalcy it causes. We can measure its relative value based on who has it and who doesn't. Measuring information is relative, and always will be, but there are better ways to achieve it.

[1474] Information is both static and transcendent. Measuring information content is both about measuring signal to noise ratios, but also about the secondary value of information through aggregation, inference and interaction. Claude Shannon began the information revolution with his seminal thesis on information theory, but others, before and after, have recognized the complexity of expression, the many forms it takes, and question how it could ever be static. When we talk about formlessness, where formlessness is part of protection, security, and privacy, we talk about formlessness having a ying and yang, a benefit for you and risk when used against you. This is true with information, in the extreme as the economy depends more on intellectual input rather than material or labor. DigitalDoors recognizes this complexity of expression, the lack of stasis in information, and the complexities of transcendent meaning in information data streams.

[1475] Bits in a data stream represent a signal. Transmission and reception introduce noise to distort that analog signal. This is what Shannon measured. However, the signal is itself a representation for other things, such as symbols. Symbols in our world include characters, words in a sentence, blueprints, orders to build something, audio, or images. Audio and images are complex constructions rather specific symbols, and in turn provide a rich tapestry of meaning. Bits can mean other things too,

limited only by our imagination or our technology to convert ideas into symbols, and symbols into concepts, and concepts into reality. Indeed, we are getting closer to technology than can fabricate things directly from ideas. If we can think it, we can make it. Story interpretation, literary analysis, symbolic explication are part of cultural experience and a rich part of our intellectual heritage. Simple meanings just do not exist. Simple information measurements do not exist, or at least they are not really part of the intellectual heritage.

[1476] There are currently 12468 three-letter symbolic entries in the DigitalDoors English language dictionary, from 'un, to <A>, to zuz and ZZZ. By comparison, there are 140608 unique possibilities with a set of 26 alphabetic characters with capitalization and 16777216 with a full 8-bit ASCII symbol set. The 12468 dictionary entries have at least one defined indication with a DigitalDoors context. That does not preclude the other 1674748 character permutations from having valid meaning and thus a measurable utility, permutations in other character sets, other languages, other encoding methods, or the 24 bits of numerical meaning too as numbers or index reference into any arbitrary database.

[1477] Note that 16777216 combinations represents only assignment within ASCII. The letter combination could play as a MIDI sound, a WAV sound, show as bitmap, convert to a long integer, or denote a real number with positive and negative values. It could also encrypt something else. Consider the number of possible encryption methods and possible encryption key combinations. The interpretations are literally endless. Meaning is more than a lookup, and more than a calculation. These large numbers provide some insight into the symbolic set space and the full range of meaning that must be considered when measuring the value of information. Furthermore, the contexts and concepts are conditional and fully relative to assessing value, correct, possible, probable, or misused.

[1478] Some of these DigitalDoors dictionary entries have a specific meaning, but most have multiple meanings. Most have multiple meanings that resolve only when viewed within a precise context. The key to meaning is like the Rosetta Stone, a stele that unlocks the mysteries of the symbolism based on a cultural context. Sometimes, we never have the explicit interpretation. Often, some of those meanings do not translate and do not resolve clearly; it is the subject of interpretation, literary analysis, and endless conflict. It is possible to measure the number of available contexts and create a statistical assessment of a possible meaning. That ratio can be subtracted from 1 (a common statistical technique for calculating useful probabilistic results), multiplied by sensitivity level assessment to create a risk measurement. This can be repeated for all the symbols in a data stream and aggregated into an overall risk summation. The same statistic is also generated for known contextual interpretations, ignoring the what ifs and the unlikely misinterpretations. This value divided by the first result yields a percentage, one possible indicator of information risk.

[1479] Value and risk are related, perhaps not by precise inversion or as a ratio of the complete value space, but both value and risk calculated this way provide useful assessment for DigitalDoors data protection through extraction and dispersion. This actually works and is functional. In fact, it underlies the DigitalDoors measurement technology that makes possible the implementation of the DigitalDoors security technology.

[1480] God, GOD, and G-d are three of the DigitalDoors dictionary entries. The meaning and the utilizing of this content is wholly driven by context and concept. It could plebian, blasphemous, religious, or ephemeral. Bad measurement is simplistic. Better measurement requires analysis of the likely meaning. Meaning transcends the three letters. Three letters is only three letters, just 12, 24, or 48 data bits, and technically cannot convey much value, risk, or signal information. However, the

symbolism fills libraries, pits cultures against each other, underlies entire economic systems, and tends to have been a principle fomenter of warfare.

[1481] Under the concepts of Claude E. Shannon, information theory regards information as only those symbols that are certain to the receiver.

[1482] A unifying theory known as information theory was developed and became the subject of intensive research. Claude E. Shannon, whose initial ideas appeared in the article "The Mathematical Theory of Communication" in the Bell System Technical Journal (1948) defined information to include the messages occurring in any of the standard communications media, such as telegraphy, radio, or television, and the signals involved in electronic computers, servomechanism systems, and other data-processing devices. The theory is even applied to the signals appearing in the nerve networks of humans and other animals. The signals or messages do not have to be meaningful in any ordinary sense.

[1483] The chief concern of information theory is to discover mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store, and otherwise process information. Some of the problems treated are related to finding the best methods of using various available communication systems and the best methods for separating the wanted information, or signal, from the extraneous information, or noise. Another problem is the setting of upper bounds on what it is possible to achieve with a given information-carrying medium (often called an information channel). While the central results are chiefly of interest to communication engineers, some of the concepts have been adopted and found useful in such fields as psychology and linguistics. The boundaries of information theory are quite vague. The theory overlaps heavily with communication theory but is more oriented toward the fundamental limitations on the processing and communication of information and less

oriented toward the detailed operation of the devices employed.

[1484] A Mathematical Theory of Communication, from us or locate it online at Bell Laboratories (Lucent Tech) . The Shannon information theory excludes the inference of information, relationships to other information and symbolic expansion, or an extended philosophy with pragmatic, functional, intellectual, cultural, developmental, or conflicting bases. However, this is not enough on which to base information measurement. This is specifically why this is a problem. Just because the symbols are uncertain to the receiver does not mean that sender is in any way uncertain about the information, or that there is another intended receiver different from the actual receiver for whom this message is actually directed. This disparity opens up sending covert messages in the clear, meaning hidden in plan sight, or information with greater value than obvious. Risk measurement that does not consider the wider range of meaning is flawed. Security that does not consider this broader extension of meaning is security with leaks. This specifically is relevant to issues of aggregation, inference, and interaction, how the receivers interpret the context of formlessness, and also in the assessment of structured data and unstructured data because structure is context as well.

[1485] The character representation is merely the designation for a symbolic concept. The religious or ephemeral concept is also represented by many other symbols, images, sounds, and even material statues. The meaning is truly in the eye of the beholder, the cultural context, and everyone seems to have very different concepts; meaning changes by spatial location and time too. A marble of Pallas Athene on the Acropolis is probably about power, technology, and religion. Two thousand years ago, that message told about the demonstrated might of Athens as the superpower. Today, the message is much different, even though the information represented by the statue is still intact. It represents how even the mighty fall and how change is the only constant. That statue now in a French museum conveys

the vision of beauty, humanity, and technology of an ancient civilization. It tells how the superpowers of today rewrite history and alter the very meaning of information itself. Information is not static content. Information decays, loses value as the culture or context changes, gets repurposed within new contexts, and is seen from within the perspective of new concepts. It is also dependent on other information for value. It is wholly contextual and conceptual.

[1486] One lamp or two lamps hung in the tower in the Old North Church in Boston of 1776 could be a mistake or an explicit symbol. Without the explicit context, the meaning is not interpreted, misinterpreted, or just meaningless. This was a semaphore code for local military response to the British Redcoat suppression of the continental American insurrection, one if by land, two if by sea. Furthermore, this code is only a code, with two possible values and an explicit built-in error-correction code of no lamps, the norm for the church tower. It does not imply, forecast, or indicate the plans based on such information, another concept once removed from the information signal itself. Information is full of such complexities, inferences, inherent cultural, linguistic, or explicit symbolism. Sometimes, symbolism becomes so commonplace that we fail to consider any other basis for an interpreted meaning. This underscores the power of a content-, context-, and concept-based information measurement and risk assessment because it is conducive to the changing basis of symbols and information in time and space or through cultural and intellectual definition.

[1487] If symbolism becomes an unknown, and unknowns represent huge risks, the symbolism needs to be interpreted as huge risks and not merely mistakes, noise, or irrelevant chatter. Lots of signals are purposefully hidden in noise or irrelevant chatter. This is hiding in plain site. Al Quaeda used wedding announcements in major newspapers, information hidden among chatter, to signal something based on a previously arraigned interpretation. Plain site must be parsed and evaluated against other

contexts and concepts to expose hidden signals. Battles are fought not only with kinetic weapons but also with information. We need to measure information to assess its meaning and thus its risk.

[1488] Information is transcendent. Nevertheless, we categorize these contexts and concepts. Of course, they are not precise and do not fit cleanly into even one or multiple categories. We can however place symbolic information into multiple categories and count the possibilities or rate the likelihood of the possibilities. Multiple categorization is a little more complicated but computer automation makes it possible. Such a rating system can be tempered by experience, by the context and concepts of other information within the same or related data streams, or assume the wanted or needed purpose or utility of the process or organization. All of these increase the accuracy of classification (of acronyms, words, phrases, sentences, paragraphs, documents, representational data, sounds, images, and other multimedia) and likely address the intent of the symbolism. There are answers, they are not exact, but they are much more effective than none.

[1489] Transcendent meaning of data streams that comprise information is fact, and measurement requires a far more extensive and flexible assertion than the symbolic content alone. It must also include the potentials, defined, obvious, intended, or merely possible, as part of the measurement system. Meanings depend on the basis, and hence measurement on that basis. Time and location are part of that basis, both past, present, and future. Since we can only forecast the future, we must forecast the potential meaning of information into that future and the value it retains, aggregates, or has with decay as well.

[1490] In conclusion, we see measurement of information as a complex process that cannot be simple nor definitive. It is relative on many factors, both of time, location, ownership, and distribution. The more these factors can be included, the more effective and valuable the measurement process can

be. It is transcendental above the content itself with a time, spatial, culture, purpose, and intent component that is modified by context and concept. Information measurement must include that potential for multiple alternative classifications, the potential for contextual changes, and the fact that value changes with time, the events that have transpired, and the human cultural condition. We recognize this complexity of expression, the lack of stasis in information, and the complexities of transcendent meaning in information data streams, and measure the value of information accordingly.

AA.5.9.0 MS Office Document Structure: Protection and Processing

[1491] The battle for data security has changed from protecting content to the battle for concept and context. Sequential text files are the exception rather than the norm. Flat, plain, and sequential files would have disappeared entirely from all but transitional processing steps except for the recent success of HTML web sites and the desire for storage of complex data into sequential XML formats. In spite of the apparent linearity of HTML and XML, in practice these flat files participate in a greater complex hierarchy of structured data mapped by object models. The object models blur the lines between content, concept, and context such that effective security requires a broader stroke than merely encapsulating content with encryption and limiting access with tokens or encrypted certificates.

[1492] Linkages to external files, style sheets, and embedded applications or scripts undermine the simplicity of HTML and XML flat formats and compromise point security. Even structured field or line and record-oriented file formats have given way to more complex data storage models. It is insufficient to view security of content and files in terms of encryption and encapsulation alone. Structured object models mix content with metadata and methods such that non-granular access—that is, either/or barrier-based access through encryption keys, dongles, and passwords—undermines any concept of effective security.

[1493] Furthermore, simplistic document management and access control overlook the multiple purposes for each compound data document and the adverse impact on organizational processes and workflows. Barrier-based security also fails from any Pacman-style attack, where the barrier, once breached not only provides full access to the once-protected interior also interferes with analysis of the attack and observation of how to prevent the ongoing attack. The Pacman attack and the DigitalDoors strategy to defend against it using viable and efficient methods is detailed by another white paper, The DigitalDoors Pacman Security Strategy Paradigm. Granular multi-level control of user data, metadata, data stored through the specifications of a hierarchical data object model, and methods underscores the new security paradigm. This transition is most pronounced in Microsoft Office documents, such as Word, Outlook, or Excel given the indiscreet distribution of source files. Office document publishing and Adobe PDF creation represents a minimal solution to the object model and metadata security risk.

[1494] All data sources important to data process workflow are non-linear, non-sequential, and not standalone in that the data sources are interconnected to or required by other data sources. This includes databases, structured documents, desktop application user files, hierarchies of data structures, and workflows. The most advanced data workflow and the focus of attention is the object-oriented models used in data processing today which comprise a cascade of events rather than a single point operation. This complicates security-related activities such as security, survivability, privacy, confidentiality, and anonymity, the trademark of DigitalDoors technology, and our contribution to improving the security of complex document object models and interdependent workflows.

[1495] There are only a handful of counterexamples to complex data structures, mostly monolithic file structures and simplistic processes. This includes text files, raw binary image files, and lists. These are typically inputs to older or uncomplicated computer activities; they do not reflect the

complexity and interrelationships consistent with and necessary for most critical networked data processing activities. Examples of flat files are text files, binary images, and lists. Plain-text documents are used only as temporarily or as conversion paths for other activities.

[1496] Binary graphics are employed for their specific simplicity, speed of display, and small size. Note that they are usually stored in an inverted backward last-to-first sequence. BMP, GIF, and other formats represent the bulk of web images.

[1497] List files are rarely important and standalone files are often a temporary part of another process. One of the most ubiquitous of plain-text files, the HTML web page, is rarely a simple text file, but a circular connection to many other like files and one part of a more complex hierarchy. A relative of lists is the field-oriented record structure. This is usually a grid-like storage of linear data.

[1498] However, even this table grid, multi-dimensional indexing, SQL query concept is giving way to object-oriented post-relational database storage methods based on object models in order to augment functionality, speed of performance, cross-platform and application functionality, and compete with easier to use user and developer products. Even the image files are becoming increasingly complex. Hierarchical images formats with vector graphics compress motion and curves into small packages. Examples include Corel Draw, Macromedia Flash, Adobe Photoshop, and Microsoft Photo. These of course contain proprietary and unintentionally-distributed information. Increased reliance on reliable data storage infrastructure and networked storage technologies is enabling the transition to data storage based on object models. Notice the root, branch, and leaf paradigm of this principal data storage structure.

[1499] The object model refers to the layout or the map of how the data is potentially stored in what is definitely a linear file. The stored file is the document object structure containing the data whereas the model is the schema representation. The model is just a blueprint for an empty data

structure. The data structure is stored as a binary file populated with data representing a subset of that blueprint. The data file is often referred to as the document binary file so as to make clear that it is not a plain-text file, not in user-friendly format, and generally readable by an ASCII reader only in discontinuous chunks. The model and the structure are not the same. The model does not represent a security threat in itself; it just represents how to find and define data stored within an actual data structure. It is the data structure in memory or stored as a file that is the security threat. Usually, the file containing the data structure gives enough clues to the purpose, methods, and sources... unless addressed by a multi-level security scheme attuned to the complexity of the object model. Although this "file" is stored as linear flat file, the extended structures is dependent on the hierarchical collection of potentially infinite branch and leaf references. Despite this complexity, there are clear reasons based on simplicity for this hierarchical structure, not the least of which is flexibility, self-documentation, and backwards/forwards compatibility.

[1500] The subtle differences between a plain-text file, a file containing lightly structured data, the schema, and a file containing data within an object structure becomes very important for security. When files are distributed and those files each contain data within object structures, workflow is complex and cannot be effectively protected with barrier-based security without complicating or disrupting operations. For these reasons, internalized security reflecting leaf content, structural paths, and the mesh of inter-relatedness among the paths, leaves, and external sources becomes the next paradigm for implementing effective content-level and application-level security. Consider the data structure defined by an object model as an organizing container. The contents within can be empty, or collections of containers, with more containers within. It is a security sieve with traditional encryption and the requisite inter-process workflows. The leafs and the security of the leaves does not secure a chain

of evidence increasingly necessary in modern data processing activity.

[1501] Enhanced security must reflect this distributed requirement since the data sources are not single point sources, but complex relational, object-oriented, or hierarchical. In addition, data access and processing is approaching a worldwide distributed infrastructure, and completion transcends single places, times, and events. When the security problem is dispersed, the security solution cannot be monolithic either but must reflect the dispersed distribution and hierarchical complexity of the data and process. Location is not the problem, so metaphorical perimeter walls are not the answer. To treat security too as a monolithic, static, and walled solution when the security problem is granular and dispersed within a flexible timeframe misses its true need. Effective data security must reflect five informational attributes in a newer paradigm for security. These include:

Table: new informational attributes

Purpose

Sources and methods

Ownership

Date or timeliness

Content

[1502] These five security attributes reflect not only the data content but also the point processes, embedded resources, and workflows. Traditional security methods fail at reflecting these attributes with a one-method-fits-all-mentality. A perimeter defense is an either/or proposition allowing or disallowing access in full, preventing any granular or multi-level security. Multiple perimeters or different entry points through a single wall also fail presuming a linear or retrograde progression of access through rather an overlap or mesh of access controls.

[1503] This metaphor fractures complex data processing workflows. Traditional security methods erect a monolithic perimeter around the process, files, delivery, or storage. Walls prevent sharing, access, and even processing. Walls are a barrier to entry for all those who do not have the permission to pass through that single door. It fails completely when that single entry point is permitted to be bypassed or any part of that barrier to entry is forced. It is monolithic security concept for data sources and processes that have not been monolithic for more than 20 years, Unfortunately, that outdated metaphor fails to reflect that not every access is through the same door and needs different controls. Data sources are not monolithic, and certainly data is not either. Distributed data, distributed processing, and widespread distribution defeats monolithic security schemes. Encrypting and packaging data files or their access methods represents a monolithic failure for complex and hierarchical data sources and processes. Access needs to be granular and multi-level, and represent the five informational attributes presented above.

AA.5.9.1 The MS Office Object Model

[1504] The MS Office document object model is based on a hierarchical data model best represented by a tree structure. The tree is flexible, extensible, forwards and backwards version-compatible, and non-linear. The root provides the starting point. The ending points, leaves or items, are many. The content leaves are sometimes linear and simple, such as plain text, but more often define object-oriented structures, mixes of plain text with rich media and descriptive attributes. The hierarchical model represents a breakthrough in flexibility in that it supports a diverse range of media types, linkages, and relationships awkward in flat file or record-oriented formats. Specifically, the data stored using object model efficiently compacts sparse-matrix information, representing only the elements which exist and none of those which are not included or have been deleted. The object model is also extensible. It

can contain triggers and even self-executing code to make the document self-aware (if not a ready macro-virus host). It is forward and backward compatible with older and newer data structures so long as the documentation represents valid subsets. It is conceptually simple. It is self-documenting; no map exists or needs to exist of any document as the hierarchy creates the map and how to traverse the document when that is necessary.

[1505] Structural documentation or schema representation is necessary only for efficient development or internal analysis of the validity of object model security implementations. It is important to have a map of the potential document structure and the purposes of the hierarchical branches and terminating leaves. However, to the uninitiated, the office document structure is complex and perverse. Although an office document is stored as a linear disk file, it is off limits to any linear access and simplistic computation.

[1506] MS Office documents include Access, Excel, Graph, MapPoint, Outlook, FrontPage, Money, PhotoDraw, Project, Publisher, Powerpoint, SQL server, Visio, Word, WordPad, plus various media file formats. MS Office applications or suites are not the only relevant tools with which to access these document modules. Many Windows applets, communication protocols and applications, and client/server activities also use these hierarchical data structures, called the object model. Furthermore, it is important to realize that an office document may comprise multiple structures within one document through dynamic data encapsulation, object linking and embedding, references and linkages, smart tags, or other methods of inclusion. Specifically, charts, reports, and tables from Excel or Access can be linked into a Word document or hosted by a Powerpoint presentation shown through a NetMeeting whiteboard connection. The depths of this embedding are limited only by access time and virtual RAM.

[1507] This embedding means that user content analysis alone is not sufficient to protect the

security of a document; the metadata or embedded data might reflect an unexpected or unanticipated content format. As such, consider the analysis of a Word document in anticipation of it being only a Word document misses the potential security compromise of an included Excel formula, an SQL raw data table, proprietary macros, ActiveX documents, and data or functionality from almost any other MS Windows-compatible product. The security risk does not end there. The data within a single Word document comprises data directly entered and controlled by the originating user. The office hierarchical data model also mixes actual user data with group and organizational data. Furthermore, information about the user data, that is metadata, is also stored within the office document structure. The valid and important reasons for this include simplicity, packaging, delivery, and functionality issues. However, the fusion of data with metadata adds to the security complexities. Specifically, the visible and obvious user data is likely to comprise 2 or 3 branches on the document structure whereas another 40 or fifty are populated with security-sensitive metadata. Consider that edits, redlining, Outlook and Exchange-enabled workflows, basic Office Suite configurations, and other metadata hide beyond even the user-accessible interface. Without intensive analysis of the root-branch-leaf structure and content, some metadata is not accessible except programmatically. That means some security-sensitive data is hidden from all but those who know exactly how to locate it and exploit it.

[1508] The object model is efficient for storing and retrieving sparse information because empty sets do not have to be written to disk, stored in files, or retrieved. However, the object model itself creates a security risk through the absence of that data. First, the objects defines what can exist. The lack of data defines what is not used and does not exist. That overlap begs the questions of what is missing and why? The absence of information—and the very clear indication that it does not exist within a file—is frequently very telling in intelligence operations. The absence of information when it is expected

represent clues, possible avenues for inference, other channels for the delivery of missing information, and fertile area for obfuscation, information hiding, and diversion. Such gaps provide contextual and conceptual information that can be as important as the directly available information.

[1509] Likewise, the fusion of data with macros and embedded triggers and functionality adds to the security complexities. However, it is not the purpose of this paper to describe virus threats or our products to protect against embedded macros or hostile applets. Other vendors and tools, such as Symantec and Norton AntiVirus, are designed to search for and remove such embedded risks. Although the DigitalDoors technology will not perform virus scanning, it is a suitable platform and process flow carrier for initiating such scanning. DigitalDoors focuses on the threats inherent within data, protecting that data, and enabling safe sharing, not the threats to data processing infrastructure.

[1510] Finally, the structure of the data within a file as represented by an object model in no way support the viability, validity, verification, or integrity of the file. In other words, a file may be valid, but it contents for various reasons have been damaged or corrupted. At the current time, most products locate defects at only the gross level and fewer yet effect recovery or repairs.

[1511] Although many people are aware of the metadata security gap in Word when they open a word document in a plain-text editor such as NotePad or a hex editor, they only see a hodgepodge that is the tip of the security iceberg. The problem is not limited to Word. It includes the full product line of Microsoft products; lack of metadata security effects most products of most other software vendors too. It is not a Microsoft problem alone. Corel, AutoCAD, Adobe, Micromedia, Intuit, Oracle, Informix, Symantec, WordPerfect (Novell) need metadata and hide it. Metadata is a reflection of the complexity inherent in the integration of actual information content with sources, methods, process, ownership, and timeliness.

[1512] Office metadata includes author initials, name, organization, creation, editing data, printing data, old edits (including corrections and deletions), redlined material, mixed author ownership, and a host of presentation, formatting, cell formulae, transmission, system addresses, and delivery information. If you believe that this reflects only minor or innocuous leaks, consider that the file creation date could contradict assertions of event dates or that deleted references to people or places could be a smoking gun in a legal proceeding. The users are certainly unaware of the extent of the metadata, but be certain, as stated above, that some metadata is not accessible except programmatically allowing security-sensitive data hidden from all but those who know exactly how to locate it to actually exploit it against you in ways not yet tried or imagined. Metadata security lapses are not minor.

[1513] Although Microsoft is aware of this metadata security failure, it has only recently and incompletely reacted with the Office XP suite and the new publishing feature. The publishing feature provides the creation of a flat linear file without the metadata content for controlled distribution. Cognoscienti use the Adobe Acrobat Distiller and PDF print drivers to create controlled distribution for older Office suites and other proprietary documents. Acrobat resolves the metadata and macro virus risk but in no way protects against data risk or enables secure multi-level information sharing.

[1514] It is also important to realize that the Office document modules comprise the blueprints, process, external data sources and linkages, and materials for building the resulting presentation; the presentation content is usually the ultimate end product. The blueprints and process often are immaterial to the presentation and represent proprietary and confidential material. The delivery of a document is not the same as the publishing or delivery of the presentation content. The difference is a significant security gap.

[1515] While this object model flexibility represents programming genius and workflow

innovations, this flexibility was not created within the context of security or the knowledge of the power of data theft and damage. This object model complexity and diverse accessibility creates a security nightmare. A simple wall barrier fails. It breaks the workflow, prevents sharing, control, and flexibility.

[1516] Effective Microsoft and metadata security, as with the DigitalDoors technology, requires adherence to the five informational attributes defined elsewhere. The object hierarchy structure is variously described as a binary tree, category structure, or hive. In any event, the entry point is the root or base, containing a potentially infinite number of subcategories, each with a potentially infinite number of leaf items. The structure can be pruned, deleted, or rearranged. The items represent object-oriented information, from entire subdocuments, to relational databases, layered graphics with vector elements, to simple plain-text, to a single binary numerical element. In actual practice and security implementation, the process required, to which DigitalDoors adheres, is to parse all branches to each and every leaf. This is not recursive, just extensive. Each path is examined for context, each leaf for content, all nodes for external references, and everything must be viewed within the context of sources and methods, not just obvious content. The obvious content is what the user created and sees, but as you now know, that is a minor portion of the data contained within the document object structure. This is a paradigm shift enabled by the DigitalDoors technology, as show in the hierarchy below:

[1517] For each document (the file and structure)

Access the root

For each limb

For each branch

For each sub-branch

For each leaf (item)

Process each leaf

[1518] Preservation of the path to each leaf is important as it defines the access to that data element. The existence and/or null value of the leaf represents a security control point. The model defines, with supplemental external knowledge of the object model, the possible security risks. The model and the content are not separate from external knowledge of sources and methods. The leaf is for all intent and purposes the significant security control point. Hiding, encrypting, or removing the leaf does not provide security any more than encrypting the document file does. It breaks the workflows to do so. However, it is possible to review and alter the contents of the leaf within the context of the purpose of the leaf to retain functional access with multi-level security.

[1519] Specifically, lets review the five information attributes of security in context to processing the leaf, purpose, sources and methods, ownership, date or timeliness, and content. Although most security methods seek to obscure content, the last and most insignificant item, effective security must review the four other elements as well. Consider the following.

[1520] How does the purpose of the leaf provide context, purpose, or informational reference to the document as a whole or the individual leaf? Does it provide source, destination, authorship, viability, validity, verification, or integrity to the document as a whole or the individual leaf? Consider the value of processes imbedded in the document as cell formulae, a help file, or other complex routing wizard. Does it show linkages or references to other documents? What is its status or position within the document? What is its element position, as a headline, footnote, or redlined status? These seemingly minor details transcend actual content but provide clues to the following attributes.

[1521] Intelligence agencies stress the confidentially of the sources and methods used to gather information. The information itself might or might not be important, but the ongoing care of the sources

and methods is important for future information gathering activities and retention of any status quo until action is initiated. In addition, the viability, validity, verification, or integrity of the document is predicated by the viability, validity, verification, or integrity of the sources and methods used to create it. In terms of the Office document, this type of information is both contextual, leaf content, and metadata. To presume that security is only user content at the leaf misses the value of metadata and the inherent risks of the object-oriented document format. For example, authorship, source, source dates, editing dates, deletions, redlining, notes, footnotes, MS hidden text, links, and other structural elements describe when, how, where, and who created the document. This speaks to the viability, validity, verification, or integrity of the document as a whole, and can compromise past, ongoing, or future data collection efforts and operations.

[1522] Ownership is reflected both in leaf-level content—that is obvious when a document is presented or published—but also in the metadata. Ownership is also a characteristic of file storage properties, in ring rights, file storage position, linkages, SMB or network file access rights, and HTML references. Ownership, particular the number of links, the times access and edited, numbers of hits, and the level of churning, suggests the relative importance and merit in the document.

[1523] Date or timeliness reflects currency. The dates, in terms of edit times, access times, and frequencies suggest the relative importance and merit in the document. Touch and other file-level commands can only mask the overt date and timestamp of a file, not its purpose or content, true timeliness, or merit. This information is spread through the metadata and leaf content. In some hierarchical structures, this information is stored in tables or other structures apart from the immediate document root. When a document is a relational data structure, as in Access or SQL, hidden system fields and hidden security data define edit and deletion times. It is also important to recognize that in

databases, records which are deleted by the user are only marked as deleted but persist until the database is purged, packed, cleaned, compressed, or otherwise processed in a maintenance mode. When relational technology with transactional logs and rollback facilities are enabled, data can be recreated or dated despite many types or natural of instigated disasters. This supplemental metadata defines date and timeliness too.

[1524] Security of content can be compared to erecting a barrier around that content. However, when content becomes a collection of simple data elements along with data objects, dispersed and distributed sources, effected by embedded events and triggered methods, barrier-based security completely fails with any breach and is not effective for the more complex object structures. Human intelligence can ascertain content from some parts of the declassified whole. Even distributed and dispersed content can provide a coherent view of the concept and context. In such cases, partial content defines the concept and the context. The details, which are extracted and hence missing from a particular reconstituted slice of the reclassified source. While amounts, times, places, and participants are missing or represented by placeholders and even misleading information, external sources of data will confirm and elucidate the missing contents. It becomes important to extract and disperse enough of the object model elements and metadata to obscure the concept and context as well. This process addresses the shortfalls of the single file encapsulation and encryption opening a functional avenue for multi-level access control of even the most perverse but prevalent of the MS Office document object models.

[1525] In conclusion, while content is king, it is the most obvious but curiously least valuable of the document security issues. In terms of protecting and processing an Office document, each leaf must be processed and assessed for its security needs. Note again that each leaf may be another object-

oriented structure in its own right or a simple element. It will need to be processed and assessed accordingly. This means the leaf must be evaluated for content, then ignored, encrypted, extracted, or dispersed. DigitalDoors can also mask, replace, or seed new content at this leaf. Reconstitution is represented by a recovery map with leaf path reference, a multi-level scheme, and multi-level data protection. This enables full or partial reconstitution as directed.

[1526] The DigitalDoors technology defines object models schemas comparable to the MS Office models. However, the DigitalDoors schemas define the many versions of the object models in terms of purpose, risk, and security lapse potentials. Let us be clearer on that. For example, Microsoft Word is a set description for Word 97, Word 98, Word 2000, and Word XP. Each of these versions has different object models. Although they are similar, each model is a variation on the theme with some compatibility. DigitalDoors describes the purpose and potential risks for each leaf and for the paths to the leaves. The DigitalDoors schemas also include a statistical grid to describe the cross-risk among the different leafs and paths. In fact, legacy leafs from a document created within Word 97 but imported and saved as Word 2000 format but opened for editing within Word XP represent security risks with leaf and path information leftover from Word 97 and Word 2000 and potentially untouchable by Word 2000. Although users are certainly unaware of the extent of the multi-version metadata, some metadata is not accessible except programmatically allowing security-sensitive data hidden from all but those who know exactly how to locate it to actually exploit it against you in ways not yet tried or imagined.

AA.5.10.0The criticality of right context

[1527] On September 7, 2008, an article about the 2002 United Airlines bankruptcy filing was posted on the South Florida Sun Sentinel business web page. This article triggered trades on UA stock, options, and other derivatives. On September 8, 2008, the market price of the stock dropped from $12.76

at opening to as low as $3.15 creating market panic. The Tribune parent of the Sun Sentinel issued this statement.

[1528] Published: September 09, 2008 10:00 PM ET

[1529] CHICAGO Updating the expired United Airlines bankruptcy story snafu, Tribune Co. late Tuesday that a dynamic link to the 2002 story that caused a stock market panic was briefly on the Web site of its South Florida Sun Sentinel.

[1530] "Our records show that due to traffic volume, sometime between 1:00:34 a.m. EDT, Sunday, September 7 ... and 1:36:03 a.m. EDT, Sunday, September 7 ... a link to the old article appeared in a dynamic portion of the Sun Sentinel's business section, grouped with other stories under a tab entitled 'Popular Stories Business:Most Viewed,' Tribune said in a statement.

[1531] "No new story was published and the old story was not re-published -- a link to the old story was merely provided," it added.

[1532] Tribune said at 1:36:57 a.m. EDT on Sunday Google's Googlebot search agent crawled the story on the Sun Sentinel's site -- as it had done as recently as last week without linking the story to Google News.

[1533] This time, Tribune said, the Googlebot linked the story, which was retrieved from Google News about three minutes later.

[1534] Income Securities Advisors Inc., a Miami Lakes, Fla., firm distributed the story to a site operated by Bloomberg News, according to multiple press reports Tuesday.

[1535] United on Monday blamed the Sun Sentinel for the report that dragged its stock price down 76% before trading was halted.

[1536] Sun Sentinel editors denied they had posted the story, and Monday night Tribune Co.

issued this statement after what it called a "preliminary investigation": "To be clear, no story appeared today or over the weekend on the Sun Sentinel website or any Tribune website regarding United Airlines' filing for bankruptcy."

[1537] Tribune reiterated Tuesday that the story "contains information that would clearly lead a reader to the conclusion that it was related to events in 2002," and it noted that comments posted with the story were dated 2002.

[1538] "It appears that no one who passed this story along actually bothered to read the story itself," the Tribune statement said.

[1539] The facts show that truthful content was correctly posted to a trusted website as one of the "most viewed" financial stories. The integrity of the story is not in doubt. Sources and methods are beyond reproach. The downstream waterfall effects of distribution and search-engine categorization lead to the story being rapidly propagated to other trusted websites. The initial activity was the posting of a historical news item. The effect occurred when this item was reported on additional news sites. Bloomberg failed to establish appropriate context integrity and assess risk for this information. The content seems factually intact and correct. The contextual issue is that the link date was set as September 7, 2008 to a 2002 story, in effect any subsequent conceptual analysis (if any) misrepresented historical information as current, thus triggering automatic trading. Subtleties have profound impact. As the stock tumbled downward, investors and automated systems responded to derivative loses by dumping stock until trading was manually halted.

[1540] Many systems and process exist today that have a fundamentally fragile tipping point. Information drives the financial markets and secondary derivatives leverage these markets creating more leverage against any such tipping point. As information assume more importance with the production

process and as part of the production output itself, it asserts a greater multiplier effect on wealth creation. Information alone can trigger a tipping point. This is clearly in evidence with the United Airlines panic. Information is neither right nor wrong, beneficial nor detrimental, intentional nor unintentional, moral nor immoral, public or self-serving. Information just is. Information is about formlessness and intangibles.

[1541] Intangibles alone, in this case information with a subtle contextual defect, Information becomes the attack and the leverage used to trip the tipping points. Whether this panic was unintentional will certainly become a legal issue, but certainly future information attacks against such leverage tipping points will be intentional. Information will be reversed against an attacker and competitors for self-interested results.

[1542] This panic demonstrated the potential opportunity to exploit recycled or fraudulent market information and to game against the worldwide financial and automatic option trading systems. This event can be applied to other markets like banking, credit cards, energy and transportation, agency and military where individuals and syndicates can profit from financial, political, or military turmoil. It is clear that information can disrupt and destabilize the status quo and precipitate reactive stampede. Even if accidental, it clearly demonstrates the future criminal exploitation of context.

[1543] This event shows that few sectors are equipped to deal with the information revolution and the evolving processes for communication. This new environment requires the evolutionary methodology for granular content control with digital door checkpoints to ownership, stewardship, usage, content, context, and concept.

AA.6.0.0 Search and data mining

AA.6.1.0 Meta Data Hound search - Adjacency with Permutations and Combinations

[1544] DigitalDoors Metadata Hound is a knowledge management exploration tool exploiting granular content control.

[1545] The purpose is to show how categorized words or unknowns adjacent to each other can be combined to form new and linear combination phrases. These combinations typically do not define PII, although they could when representing proper names or addresses in full. More likely, these combinations are best represented as metaPII to pierce the secrecy veil through aggregation, inference, and interaction. Content is conjoined with context and concept to expose intangible property including trade secrets. Protection against this complex risk requires more complex analysis than simple white lists and black lists.

[1546] Adjacency is a function of semantic seriation, in that meaning is conveyed by the proximity of word combinations. Digitaldoors terms these generated phrases, linear adjacent composites or sometimes linear adjacent permutations in recognition of the technical method whereby these terms are generated. These phrases typical augment the white and black lists. These conjoined permutations represent metaPII more complex than simple elements. In conjunction with linear adjacent composites, the DigitalDoors My Group List functionality enables complex Boolean permutations and combinations when the adjacency is not linear but could be dispersed over sentences, lines, paragraphs, or throughout an entire data stream.

[1547] Linear adjacent composites have significant applications both in knowledge discovery and defeating gaming attempts to bypass spam filters. For example: when "Khalid Sheikh Mohammed" is unknown, it will be categorized by the DigitalDoors secure engine as granular content in three distinct words. Given the default dictionary configuration, both Khalid and Mohammed are categorized as male names and Sheihk is categorized as unknown (although sheik is an honorific title). Alternatively, Sheihk

can be designated as a possible alternative spelling for the respected Sheik title. However, DigitalDoors nevertheless realizes that three contiguous adjacent words are categorized and potentially assigned either to the black list or white list. The potential therefore exists for new permutations of those terms, specifically:

Khalid Sheikh

Khalid Sheikh Mohammed

Sheikh Mohammed

[1548] Although other possible permutations include the null set, Khalid Mohammed, Mohammed Khalid, Sheikh Khalid, Khalid Khalid Khalid, Sheikh Sheikh Sheikh, Mohammed Mohammed Mohammed, these linguistically or syntactically less useful combinations are ignored by DigitalDoors.

[1549] A user can be notified of these three permutated phrases for inclusion in My Word List or alternately added to the user, group, and/or organizational dictionaries. These phrases are automatically submitted to the DigitalDoors Metadata Hound[1] for divergent exploration because they represent potential search threads in knowledge discovery and legal e-discovery. While an experienced analyst would recognize this phrase entity as a terrorist, a novice might not. Certainly, before any such combination phrase is linked into a search vocabulary after-the-fact, it could become an element of interest at its initial appearance in any intelligence data stream by virtue of its novelty and proximity to other elements of interest.

[1550] Furthermore, automated recognition systems would recognize this as a terrorist if and only if its attention lists included such compounded phrases. Because the prime advantage in asymmetric

---

[1]

warfare is novelty, recognition of novelty is critical to antiterrorism and other first-to-market advantages, specifically with regard to issues of sustainable competitive advantage or barriers to market entry. While attention lists can be manually updated to reflected novelty, this is tedious, error-prone, subject to the imaginations of the luckiest or brightest people, and unfortunately most likely to be recognizable only after the fact. DigitalDoors technology provides an automated advantage. Instead, the analysis of contextual relationships and the application of adjacency with permutations and combinations can be applied to automatic discovery and to augment attention lists.

[1551] Attention lists include exclusively black lists or white lists. Black lists include words of exception that raise a red flag or are rejected by spam filters. White lists include everything presumed to be safe. Although black lists and white lists categorize known words and phrases, the gray list of unknowns represents the bulk of the data stream. Additionally, this unknown data stream is a growing for unstructured documents. The content knowledge of these lists also tends to leak out, and be played by attackers, thus further reducing the instances or items on either the black or white lists and skewing the importance of such gray list taxonomy. The failure to categorize this gray list and reflects its currency, the material that is neither black nor white, significantly undermines information life cycle management, security, and information sharing activities because the filtering activities fail with this gray list. Techniques based on fuzzy logic, statistical Bayesian inference, and similarity rankings appear promising, but are neither widely implemented nor successfully demonstrated as enduring.

[1552] Specifically, while Bayesian inference is used extensively for e-mail spam categorization, spammers have demonstrated that they win the arms race by combining multiple common words into a nonsense phrase that defeats the Bayesian filters, splitting common words with in-between filler, or using alternative symbols in place or look-alike letters. The use of search as a supplement provides

semantic and syntactical assessment of such combinations through the DigitalDoors Metadata Hound for convergent integrity validation. However, this increases the cost, delay, and complexity of filtering by taxonomy, and it might represent the only sufficient method to discover blended, polymorphic, mutating, evolutionary, and increasingly sophisticated malware attacks.

[1553] Since linear permutations current allow for adjacent phrases made with up to nine words (default value, but user configurable between 2 and 20), DigitalDoors supports automatic rejection of false positives or inclusion of false negatives and positive hits into My Word List. This facility is useful to minimize the computer overhead for creating the permutations and filtering out junk. The linear adjacent composite process is optional in manual mode but automatic with DataHound and all metasearch activities. The list of newly-generated composites and rejects is presented during this process so that positive hits can be classified for future use or bypassed as rejects. Rejects can also be reclassified and thus re-enabled, when syntax or conditions change.

[1554] Additionally, it is also possible and useful to explore the text surrounding a target for value. For example, consider the partial sentence:

[1555] "...planning to board a flight from Logan International Airport to Ronald Reagan International Airport at the end of the month..."

[1556] "Ronald" and "Reagan" are name entries, "Logan International Airport" is a phrase, but the phrase "Ronald Reagan" and "Ronald Reagan International Airport" are not name entries.

[1557] They are places, locations, targets of interest, and user-specific My Word List entries. Adjacency around the targets generates both phrases automatically, as well as "Ronald Reagan International" and "Reagan International Airport" because the defined dictionary stop words 'to' and 'at' stop the aggregations. That is why they are called stop words. The administer or individual users can

select which, if any, should be added as entries, to My Word List, or to targets of interest ('always target' or 'always ignore' categories).

[1558] Future DigitalDoors enhancements might include combinations and permutations dispersed over sentences, lines, paragraphs, or throughout an entire data stream to automate augmentation of My Group List with automatically generated entries. The difference between what is currently available and this potential enhancement is merely a matter of range, scope, and depth. The technology is already developed and is identical to what is currently available. It is a functional question of signal-to-noise ratio. Specifically, the workflow downside is that the number of generated results will be very large indeed requiring manual review to remove mostly chaff and specify only utilitarian My Group List entries.

AA.6.2.0 Search Convergence and Search Divergence

[1559] DigitalDoors granular content metasearch results provides two distinct and fundamental guidances, search convergence and search divergence, both of which are critical to the process of connecting-the-dots. Search convergence is the process of validating, substantiating, and establishing the integrity of the fundamental search. Search divergence is the process of expanding the content, context, and concept beyond the strict scope of the fundamental search. Failure to consider both convergence and divergence leads to incomplete results.

[1560] Convergence is the usual web-like search engine or data mining operation. It is rigid. You start with known keywords and indexes -- search terms -- as the base search query and accept hits with the highest relevancy. This information is used aggregately to confirm or deny the integrity of the initial search. You get back what you expect with priority, relevance, or ranking focusing the hits to the content most consistent to your context. Outliers are naturally rejected and filtered out. Search convergence

might entail a physical inventory, database report verification, an independent party confirmation, or a drone flyover to substantiate the information. The result is in direct response to the implicit question, that is the base search query. You position and substantiate the dots you already know. You start with old information, and you get old information. However, the benefit provides by DigitalDoors is that it contextualizes these convergent results for relevancy to your specific purposes with the metafilters of categorization and sensitivity level taxonomy and thus finds the most appropriate results for the user.

[1561] However, some people find outliers useful to expand the scope of their knowledge and infer new intelligence. Some intelligence operations require exploration without preconception. Herein, outliers with the least priority, relevance, or ranking frequently provide previously unknown references or interconnections that are purposefully hidden. This is a search divergence because the base search query primes the process but effectively we already know the highest ranking results and reject those results while iterating the search for lesser-ranking items only. Divergence is the expansion of the search query or data mining operation.

[1562] You start with known keywords and indexes, but accept only information that is new or different from the direct response to the implicit question, that is the base search query. Validation and integrity is not important to divergence. You are looking for all related metadata, references, links, and relationships without regard to accuracy. Exact matches are in effect ignored, but all new references are added into a secondary query yielding subsequent results. The subsequent results in turn yield yet more new references. This subsequent query yields subsequent results, which is repeated indefinitely. The iterative results create an asymptotic adjacency list model with a social networking relatedness. Social networking relatedness is often viewed as flow charts showing betweenness, closeness, and connectedness. The theory of six degrees of separation has some legitimacy, however there is a

difference between accidental and active relatedness. This expansion of the base search query is the process by which you discover more dots. Connecting dots you already know are pointless does not expand your knowledge; you want to expand the set of dots and generate new content, new context, and new concepts. DigitalDoors starts with old and/or limited information, but generates new information.

[1563] Divergence is also critical to enabling DigitalDoors security by exploring the purpose and usage of information beyond the immediate and obvious preconceptions (i.e. foresee reflexivity) of semantic and pragmatic meaning. See beyond the immediate security perception for other possible cross-domain requirements and enable information sharing.

[1564] You are connecting known dots with new dots, the true concept of connecting the dots. The new dots lack integrity and substantiation. Do not ignore them or dismiss outliers because they can reveal the outlines of previously unknown, undiscovered, unforeseen, uninvestigated and unrevealed situations. The search process is really a metasearch applying techniques of data mining, textual and contextual search, aggregation, inference, and data-to-data interaction, and the integration is the establishment of relationships. However, some of your new dots will be false. At this point, a search convergence is useful to establish likelihood and risk for the new information in order to validate and establish the difference between accidental and realistic relatedness.

[1565] Although convergent search is different from divergent search, convergent search is useful after a divergent search to establish integrity. However, since the process is ongoing and automatic, interim results are available for manual pruning and grafting. This is useful for noise reduction and inserting feedback from human intelligence. At any time, new search terms can be inserted into the divergent search and dead ends removed from the process.

[1566] Divergent search based on only content search terms overlook the power or context and

concept. DigitalDoors filters search results through the DigitalDoors multitier hierarchical overlapping tag structure. For example, only search results that are confidential or top secret and represent names and locations might be retained. The filters are a function of range and scope.

[1567] The convergent search terms are represented within a tree view, with new divergent search terms represented as subordinates branches and nodes. Nodes that match (exactly or proximally) can be linked and represented by a multi-dimensional relationship graph.

[1568] DigitalDoors metasearch performs two key functions. First, it selects new data elements of value and recognizes unknowns for automated iterative divergent search. Second, it categorizes all aggregated results to sensitivity level and enables distributions subject to MLS-compliance cross domain. While the DigitalDoors primarily focus on unstructured information, it can be integrated with structured data sources too because it is actually easier to extract context and concept from structured data.

AA.6.3.0 Convergent search

[1569] Typical search and data mining tools presuppose you already know part of the answer. This presumes partial knowledge a priore just to phrase the question in order to get a proper answer. However, this imperfect solution for knowledge management does not begin to address the real issues needed when connecting the dots. We have termed this type of knowledge management and search technology as convergent because you start with a known entity and confirm or deny that basic entity or gather additional information in depth. The initial search terms are categorically narrowed necessarily to reach this focus. Note that DigitalDoors automates the convergent process until no additional information is found on the topic. Convergent search presupposes the reason for the search or the premise of the additional information.

[1570] Sometimes, there is no presumption, no partial knowledge. Sometimes, however, the

questions are hidden without premise. Sometimes, categorical precision, preconceptions, and historical contexts expressively distort data mining, knowledge management, and search, security, and information sharing. Sometimes, the results require a complete disconnect and disengagement from reality, history, obsession, cultural imperatives, the norm, or the way things have always here been done.

[1571] Cultural bias can over-classify data stores, preclude granular access to items, or skew the distribution of information based on the preconceived notion of what someone needs to know. This propagates the scenario that need to know is contextual to a security architecture and very constricted world views. This fosters a view of the leaves -- at most the trees -- instead of the forest. The focus is preordained, and quite frankly, so are the subsequent results. This is indicative of the failure to establish the methods and security within a sharing infrastructure suited to connecting the dots. Security is contradictory to sharing, particularly when security is driven by limited sensitivity levels and compartments rather than by a granular content control and a multi-tier overlapping hierarchically tagged categorization.

AA.6.3.1 Divergent Search Paradigm

[1572] In contrast, DigitalDoors allows knowledge management as a divergent exploration. The basic starting point is any data stream (RSS, blog, documents, database, satellite ground station, a veritable fire hose of data) or multiple data streams, or a collection of search terms for an automated process. A single term defines a manual process with a primary search term. The automated process redacts the data stream(s) and looks for content and context of importance creating a list of primary search terms, metadata contextual filters, and noise reduction tools in order to expand the breath of knowledge with new links and references.

[1573] Because the search process encompasses the DigitalDoors selection and extraction

engine, the underlying process is to categorize content, context, and concept as a precursor to this expanding divergent search. Categorization is consistent with preexisting categorical precision, preconceptions, and historical contexts but also anything outside that norm is also explicitly recognized to enhance the divergent search.

[1574] Search results are aggregated hits or the information referenced by the hits. Depth settings determine the number of hits processed and the preponderance of information accumulated for assessing accuracy, integrity, and confirmation. Scope defines how much of the actual information is explored for knowledge exploration and later convergence to necessarily obtain some focus. DigitalDoors does not neglect or disdain convergent search at all. In fact, convergence is the necessary resultant process to divergent search. Together, convergence and divergent represent contrasting but ultimately complementary processes that typically start as a divergent search, with concurrent hits funneled into convergent searches for authentication, accuracy assessment, relevancy, and encapsulation with a focus.

[1575] The cycle of convergence and divergence extends the linear adjacent recognition of categorized or unknown keys or phrases which can be permuted to harness new and previously unknown search terms or information (social networks, relationships, references, hidden connections, etc.).

[1576] Alternate spellings, misspellings, partial matches, duplicates, and other fuzzy technology is used to graft depth to a convergent search or prune a divergent search. The search architecture is open and conducive to add-ins, competing products, and supplemental tools to enhance accuracy, depth, breadth, speed, or quality. Likewise, Digital Metadata Hound can be integrated into SaaS, SOA, or other distributed workflows.

[1577] In the case of a divergent search, DigitalDoors is specifically exploring for secondary search terms and information not initiated by the primary search terms. Noise and duplicates are pruned.

All hits and information located with URLs, X-links, or other pointers are retained to maintain chain-of-custody, reproducibility, and subsequent human intelligence analysis. Because consolidated results alone are insufficient for legal exploration, search warrants, or plans of action, or even spurious, suspect, and in context to anything, therefore the retention of the intermediate hits become critical to build a rationale for action and review the relationships, connections, and networks of people.

[1578] The search is extended with the primary search terms, that is the initial entity topic, the secondary search terms, and combinations and permutations thereof. Duplicates and non-duplicates are pruned per the selection of convergence or divergence. Tertiary search terms and information are aggregated in the same manner as the secondary search terms and secondary information. Quaternary search terms and information are aggregated in the same manner as the secondary search terms and secondary information.

[1579] This divergent process repeats unless manually interrupted, terminated by a timer, manually reviewed midstream and grafted or pruned, or terminated by repetition or the lack of newly-found information. When no new information is found -- corresponding to an asymptotic plateau and the natural conclusion -- both concurrent convergent and divergent searches are stopped. Convergent searches are stopped when the results only provide information not confirming preconceptions. Divergent searches are stopped when the results no longer represent new things.

[1580] This combination of both convergence and divergent search overcomes cultural bias, the excessively restrictive classification of data stores, precludes granular access to items, and skews the distribution of information based on the preconceived notion of what someone needs to know. DigitalDoors expands the scenario that need to know is open but consistent with rigorous and compliant security architecture. This fosters a view of the forest, the exploration of new content knowledge, new

contexts, and new concepts. This encourages the search for new resources no longer bound by the restrictions of categorical precision, preconceptions, historical contexts, cultural imperatives, the norm, or the way things have always been done here that beyond air gaps and data perimeters of single agencies. This is indicative of a new paradigm establishing the methods and security with sharing infrastructure suited to connecting the dots.

AA.6.4.0 Data mining and inference attacks

[1581] Every computer-related application, interaction, and workflow leaves behind indication of its passage. These terms are just some of many in the literature, but all point to gaps in garbage collection, process cleanup, software installation and removal, and oversight. Every process leaves trails that are relevant to data mining, aggregation, inference, and data-to-data interaction. These include:

Table: trails for data mining

Tokens

Assignment of privileges

User or process ID and information

Data fragments in application baggage

Data fragments in files

Data fragments in unallocated file storage

Data fragments in RAM, buffer, cache

Data fragments in CPU RAM, buffer, cache

PnP Device Instance ID

Installation resources never used (drivers, files, registry entries, etc)

Orphaned driver files

Obsolete driver files

Orphaned libraries

Obsolete libraries

Orphaned registry entries

Tags

Sidebars

Footnotes

Index markers

Cross-references

Section markers

Links (internal, URL, X-Pointer, etc.)

References

Table of contents markers

Material marked for deletion

Metatags

Metadata (primary content, comments, markup, tags, metatags, second-order content elaboration, usage descriptions, links, references, embedded data, scripts, macros, slough [exons], interdependencies)

Misaligned dependencies

Style Sheets

Data Type Definitions (metadata, HTML, XML, and database)

Orphaned metadata

Unrecognized metadata

System driver assignment allocations

System driver port assignment allocations

System driver drive letter assignment allocations

System driver class assignment allocations

System driver network assignment allocations

Serial number identification of a device or instance

Administration references, assignments, resource allocations in configuration files

Administration references, assignments, resource allocations in registry

OS Logs (as in WMI data collection)

Event logs

Temporary files

Device drivers

Registry assignments for device drivers

Tampering

Unallocated storage

Unallocated files

AA.6.5.0 Metaseaerch data mining

[1582] Stovepipes and silos, independent efforts at information collection and intelligence analysis, and air gaps or security guards divide domains and prevent information sharing. Centralization and consolidation of data gives rise to data overload, a problem in part addressed by new web-like searching techniques, but also opens the door to security leakage. Data mining is explicitly content-based

and rigidly contained with the data warehouse. Search is rigid because the starting point for each and every effort requires an index or explicit initializing terms. DigitalDoors technology, based on granular content control, enables security with information sharing including MLS-compliant search capabilities across domains. Data is explored not only by content but also context and concept.

[1583] Search should occur across domains, with information shared and delivered across domains and through the extraction and integration of both private and public data sources. When connecting the dots with information pooled from different agencies, it is essential that search be possible without preconceptions of utility, purpose, or cultural bias. Effective search, and what is termed metasearch throughout this paper, comprises more than web-like textual (and other formats such as audio or video, but also compound meta-formats) search technology. It includes data mining, textual and contextual search, aggregation, inference, and data-to-data interaction with granular content control subject to security compliance.

AA.6.5.1 Metasearch

[1584] Metasearch generally means the usage of multiple web-based search engines based on keywords with results combined into a single unified report; this is better described as a confederated search or federated search. Each web-based search engine is rigid in terms of its reliance on preexisting indexes (robots crawl through the Internet looking for changed, added, or moved data and the relevant categorization to apply to it. Federated search is usually fast but limited in terms of most likely usage bias. However, the combined results reinforce the preexisting answers based on the implicit search question.

[1585] In contrast, the DigitalDoors metasearch is designed to traverse domains, search for unknowns and expand the base of information. DigitalDoors metasearch results provides two distinct

EX. 1002, Page 577

and fundamental directions, search convergence and search divergence, both of which are critical to the process of connecting-the-dots. Search convergence is the process of validating, substantiating, and establishing the integrity of the fundamental search. Search divergence is the process of expanding the content, context, and concept beyond the strict scope of the fundamental search. Failure to consider both convergence and divergence leads to incomplete results.

[1586] Convergence is the usual web-like search engine or data mining operation. It is rigid. You start with known keywords and indexes -- search terms -- as the base search query and accept hits with the highest relevancy. This information is used aggregately to confirm or deny the integrity of the initial search. You get back what you expect with priority, relevance, or ranking focusing the hits to the content most consistent to your context. Outliers are naturally rejected and filtered out. Search convergence might entail a physical inventory, database report verification, an independent party confirmation, or a drone flyover to substantiate the information. The result is in direct response to the implicit question, that is the base search query. You position and substantiate the dots you already know. You start with old information, and you get old information. However, the benefit provides by DigitalDoors is that it contextualizes these convergent results for relevancy to your specific purposes with the metafilters of categorization and sensitivity level taxonomy and thus finds the most appropriate results for the user.

[1587] However, some people find outliers useful to expand the scope of their knowledge and infer new intelligence. Some intelligence operations require exploration without preconception. Herein, outliers with the least priority, relevance, or ranking frequently provide previously unknown references or interconnections that are purposefully hidden. This is a search divergence because the base search query primes the process but effectively we already know the highest ranking results and reject those results while iterating the search for non-ranking items only. Divergence is the expansion of the search

query or data mining operation.

[1588] You start with known keywords and indexes, but accept only information that is new or different from the direct response to the implicit question, that is the base search query. Validation and integrity is not important to divergence. You are looking for all related metadata, references, links, and relationships without regard to accuracy. Exact matches are in effect ignored, but all new references are added into a secondary query yielding subsequent results. The subsequent results in turn yield yet more new references. This subsequent query yields subsequent results, which is repeated indefinitely. The iterative results create an asymptotic adjacency list model with a social networking relatedness. Social networking relatedness is often viewed as flow charts showing betweenness, closeness, and connectedness. The theory of six degrees of separation has some legitimacy, however there is a difference between accidental and active relatedness. This expansion of the base search query is the process by which you discover more dots. Connecting dots you already know are pointless does not expand your knowledge; you want to expand the set of dots and generate new content, new context, and new concepts. DigitalDoors starts with old and/or limited information, but generates new information.

[1589] Divergence is also critical to enabling DigitalDoors security by exploring the purpose and usage of information beyond the immediate and obvious preconceptions (i.e. foresee reflexivity) of semantic and pragmatic meaning. See beyond the immediate security perception for other possible cross-domain requirements and enable information sharing.

[1590] You are connecting known dots with new dots, the true concept of connecting the dots. The new dots lack integrity and substantiation. Do not ignore them or dismiss outliers because they can reveal the outlines of previously unknown, undiscovered, unforeseen, uninvestigated and unrevealed situations. The search process is really a metasearch applying techniques of data mining, textual and

contextual search, aggregation, inference, and data-to-data interaction, and the integration is the establishment of relationships. However, some of your new dots will be false. At this point, a search convergence is useful to establish likelihood and risk for the new information and integrity and substantiation of the difference between accidental and active relatedness.

[1591] DigitalDoors metasearch performs two key functions. First, it selects new data elements of value and recognizes unknowns for automated iterative divergent search. Second, it categorizes all aggregated results to sensitivity level and enables distribution subject to MLS-compliance cross domain. While the DigitalDoors primarily focus on unstructured information, it can be integrated with structured data sources too because it is actually easier to extract context and concept from structured data.

AA.6.5.2 Description of the DigitalDoors Metasearch Process

[1592] There are multiple starting points to the metasearch process. One or more candidate data streams, news feeds, or documents can be dragged from the desktop into a queue. These items are monitored for changes and then feed into the DigitalDoors Secure engine for granular content selection and metatag categorization with the output result that search terms are generated and metasearch terms are also generated. A document or data stream automatically sets the context and is explored for likely search terms. No index or bias is necessary to begin the process. The resulting search terms of interest are fed into one or more URLs for exploration.

[1593] Indexes and search terms do not define the implicit question or establish parameters for an effective answer because they represent a single dimension of information. Metafilters such, as phrases, My Group Lists, categories and sensitivity levels add both context and concepts to the initial search and all subsequent exploration.

AA.6.5.3 Combinations and Permutations

[1594] DigitalDoors recognizes complex content, such as the user-defined entries in My Phrases, My Word Lists, and the ranged limited My Group Lists. However, unknown content can slip through this sieve without explicit human intervention, therefore DigitalDoors automates this sieve by creating permutations of both categorized and unknown words and phrases. These permutations are used as new search terms. The likelihood of random combinations (included in the class of the NP! problem space) is so low, that emphasis is focused instead on linear adjacent selections. This results in is a small and focused problem space that can be solved on a typical personal computer and adds disproportionate value. For example, a Google search of "Mohammed Atta" returns "Khalid Sheikh Mohammed" as a disconnected string of categorized and unknown words out of the box; that means parts of the phrase "Khalid Sheikh Mohammed" are recognized without any special configuration, but not as a contextual phrase. Permutations of linear adjacent selections do however create the complete connected phrase and automatically submit the complete phrase for a divergent metasearch, thereby creating new and previously unrecognized relationships to connect the-dots.

[1595] Content is transformed by context and concept. Structure, format, purpose, usage, ownership, intent, validity alter content, as explained in other DigitalDoors papers. Specifically, the transformation of data creates weights, permutations in meaning, and the presentation affect overall meaning. A tearline does not include the entirety of the message or the initial intent of the meaning. This is a loss of content but a freedom to reevaluate content items without preconception. As such, it is an extension of search, but a critical necessity for security and sharing purposes, to entertain the potential meaning in the transformation and redaction of the data stream structure and the content into granular items.

AA.6.5.4Automated Contextual Pruning

[1596] Many divergent mining and search efforts will expose unknown, deceitfully hidden, and statistically irrelevant relationships. For example, a Google search of "Mohammed Atta" also shows meetings at Aleppo. The relevant thread for Aleppo is that it is Syrian university town with social connections to Hamburg and the 9/11 hijackers associates. However, Aleppo is also a common Lebanese street name, and the original shoemaker character in the Pinocchio story by C. Collodi. The name was changed to Geppetto in the Disney childrens' movie. While it is tempting just to prune this track, it is also important to explore associative contexts, because the use of codes can mask more sinister meanings. The 9/11 hijackers communicated within the seemingly innocuous context of weddings, guest lists, and dates. So while contextual pruning is easier after the fact, it can be very important before the fact when the use of codes and false contexts are not so clear.

[1597] In conclusion, although stovepipes and silos, independent efforts at information collection and intelligence analysis, and air gaps or security guards among domains prevent information sharing, DigitalDoors metasearch enables security with information sharing including MLS-compliant search capabilities across domains. It overcomes security leaks inconsistent with the needs to both protect national security interests but also enable cross domain information sharing.

AA.6.6.0 Knowledge Discovery: Search Convergence and Divergence

[1598] The foundation of DigitalDoors technology for security and information sharing is based on measurement for risk within granular content (and metadata). It is applied to secure information while providing interdomain information sharing. Risk measurement is driven by information recognition, multitier overlapping hierarchical metatagging technology, and granularization. This same technology, as a standalone deliverable product or service-oriented architecture, is applicable to knowledge discovery, uncovering unknown or concealed relationships, and connecting the dots. This paper explores

DigitalDoors DataHound, which is a serialized and repetitive process for knowledge discovery, usage of the DigitalDoors Secure engine, the application of granular content, or standalone for knowledge discovery.

[1599] Actual DigitalDoors security is implemented through dispersion. Information sharing is implemented either through the delivery of MLS-compliant versions, or through a base-redacted document with objective security compliance for universal distribution with reconstitution. Reconstitution is supported in whole or in part as MLS-compliant results, or as step-wise defenses-in-depth MLS-compliant results. Because the risk measurement is driven by content, context, and concept, it is essential that DigitalDoors recognize content not only in isolation to itself as a data stream or document but also its value when combined with other private and public sources through aggregation, inference and data-to-data interaction. DigitalDoors Metasearch is useful to augment standard Internet search engines because it enables both search divergence for knowledge discovery and search convergence for assessing information integrity. It completes the process necessary for discovering new knowledge to connect the dots.

AA.6.6.1 Initial Process

[1600] The automated DigitalDoors metasearch tool takes as source any data stream or document. The DigitalDoors Secure engine automatically parses the source for content recognition and assessment of granular content. This granular content is automatically metatagged for range, sensitivity level, compartmentalization, categories, relevancy, and other multitier overlapping hierarchical factors. The granularized data content becomes the search terms, while these metatags become metafilters for the DigitalDoors knowledge discovery. These search terms are submitted to any arbitrary number of extant tools and services for locating other relevant data streams, documents, databases, local, Intranet,

Internet, and public or private data stores. Likely tools include Internet search engines, data mining tools, database query engines, and other knowledge management (KM) applications or services. Inductive and deductive tools are advantageous too. Most tools can be easily integrated with DigitalDoors linked through service-oriented architectures (SOA), or their results piped into the source data steam or document (through Web 2.0 mashups). DigitalDoors metasearch also works with standard, a priori -, search terms or keywords, like any desktop, web-enabled, or Internet search tool; this bypasses the automated DigitalDoors Secure assessment of a relevant or representative source data stream or document.

AA.6.6.2 DigitalDoors MetaSearch

[1601] The results from the extant tools are logged for auditing, dispersed to maintain a legal chain-of-custody, and aggregated into a single collated result,. This collated result (the source, results, reports and links or references) is automatically submitted into the DigitalDoors Secure engine. Inputs include the configuration for either search convergence or search divergence. Search convergence conforms to the standard purpose for using Internet search engines like Google or Alta Vista because it defines, authenticates, and validates the search terms. Convergence does not provide new information but provides integrity for the initial information. Divergence is all about aggregation, inference, and data-to-data interaction because it specifically searches for links, references, relationships, outliers, and social networking associations to the search terms; it ignores anything overlapping the search terms. In other words, convergence increases the depth of knowledge where divergence increases the breadth of knowledge. It explores new, novel, unknown, and hidden connections. If you want to connect the dots, divergence dredges new but factually-uncertain information where convergence thereafter authenticates. A discrepancy search is also within the realm of divergent search. The metasearch inputs also include

the metafilters created prior to the aggregated result. These metafilters narrow the scope of all subsequent and serial submissions to the extant tools and services. The filters include specification for sensitivity level of granular content (for example, find only [TS] content), limitations by category (find only locations or people), only content of a specific importance (for example, find only within My Group List), and other DigitalDoors features and compartmentalization. Metafilters for standard search terms, which bypass the initial DigitalDoors Secure review of a source, are optional.

[1602] DigitalDoors Secure is automatically run. The source (which is now the aggregated result from the extant tool processes) is automatically parsed by the DigitalDoors Secure engine with secondary recognition and assessment of granular content. This granular content is automatically metatagged for range, sensitivity level, compartmentalization, categories, relevancy, and other multitier hierarchical factors. The granularized content becomes the search terms and the metatags become the metafilters for the sknowledge search. The automated results of the Secure process is an augmentation of existing knowledge or a breakthrough to unknown or covert relationships. The convergence search vets the integrity of a priori search terms, while the divergence search generates new search terms and new references. The metasearch is repeated, as configured, for a fixed number of times, for a fixed period of time, until results reach an asymptote, or indefinitely.

AA.6.6.3 MLS-Compliant Search and Distributions

[1603] Because DigitalDoors sand all the extant knowledge management tools are run in a secure system high environment, results are likely to be system high too. Repetitions remain system high. However, the point of any search or knowledge exploration is to share information and distribute it to the users at the edge subject to optimal restrictions. This is the reward generated from the DigitalDoors paradigm shift. The new search terms, the aggregate result, and all intermediate reports and documents

are processed for MLS-compliant security and information sharing.

[1604] Specifically, the list of new words, phrases, and other content is either automatically assessed or manually reviewed for addition to the DigitalDoors dictionaries. Ownership, sensitivity level, compartment, group, categories, mission, relevancy, and other multitier overlapping hierarchical metatags are applied to each discovery and distributed to users subject to MLS compliance. The aggregate result are secured with multiple MLS-compliant versions and distributed to the edge on a per sensitivity-level basis, or a base redacted document with objective security compliance for universal distribution and pushed to the edge with DigitalDoors sView reconstitution services. These two options are also applied to all intermediate results, reports, lists, linked or referenced sources. Of note, all links and references, even access to public data sources can be optionally restricted to minimize subsequent risks from aggregation, inference, and data-to-data interaction. An agency with specific intent and knowledge is able to use aggregation, inference, and data-to-data interaction on public sources to create classified results. For example, if unclassified search terms including "president assassination British visit" returned news clippings and RSS feeds of a forthcoming presidential visit to 10 Downing Street, these public results are still unclassified but the context is clear and should be classified.\

[1605] DigitalDoors Metasearch is a novel and useful standalone service to uncover unknown or concealed relationships, and connect the dots. Although metasearch is derived from the DigitalDoors s technology for security and information sharing, it is useful to augment standard Internet search engines because it enables both search divergence for knowledge discovery and search convergence for assessing information integrity, the full process necessary for connecting the dots. Search divergence extends search with aggregation, inference and data-to-data interaction beyond simple content into the realm of context and concept.

AA.7.0.0 Security:

AA.7.1.0 DigitalDoors Security Keystone overview

[1606] Digital Doors sanitizes sensitive documents, media, programs, and messages (specific examples of data streams) with the unique capability of automatically reconstructing those messages upon validation of security clearance. The DigitalDoors technology associates multiple receivers and their respective clearances with multiple variations of a document. This supports information exchange between trusted parties, parties that do not trust each other, and even parties unknown to each other. It is the basis for the currently functioning DigitalDoors security system applying the architecture and processes that:

select granular critical and important data within a data stream (document, data file, transaction, object, structured storage, record, field, document, etc.)

extract that same critical and important data

substitute placeholders for that same critical and important data

disperse that same critical and important data extracts to remote distributed storage locations

release that same critical and important data extracts from storage through a controlled release process of granular access rights of each user through validation and authentication

reconstitute that same critical and important data extracts in whole or in part to reconstruct the original data stream (such as a e-mail message, report, or legal document)

[1607] In effect, the DigitalDoors technology hides information leveraging the distributed capability of the private and public communication channels, such as the Internet. This avoids single points of failure to establish an in-depth defenses through creation of numerous granularly-controlled obstacles to attackers. This technology secures against insider as well as external attacks. DigitalDoors

can measure the security before and after the process for detailed metrics on the process, data integrity, and data security. This technology enables secured sharing of information on a need-to-know basis. Shared information can be distributed uniquely or en masse to internal or external organizations, with each user individually receiving information automatically tailored to each user's security clearance. When circumstances change, and for whatever reason, and after the fact of distribution, the sender can still control the release of critical and important information from distributed storage locations. DigitalDoors operation is transparent and automatic. User access is through access controls already established by organizational security policy.

[1608] DigitalDoors is in the process of conforming to DOD and other relevant standards. This includes encryption limited to AES and Rijndael. Furthermore, all communication links will be protected by SSL, optional point-to-point hardware encryption, and other technology. The objective is to house the primary subsystems in an insulated environment protected from unauthorized access. We support and will integrate with public key cryptosystems. Design was primarily implemented in UML, with automatic code generation in MS Visual Basic. We expect the port to C, Java, and other languages to be fairly straightforward. User interface issues are imbedded in flexible calling interface parameters supporting regeneration to any widget-based GUI.

[1609] The DigitalDoors extraction and dispersion architecture was conceived within the explosion of Internet and web-related services as a commercial off-the-shelf (COTS) utility toolkit and application suite. DigitalDoors supports and extends SAN and NAS storage service and leverages the benefits of those storage technologies. It is equally useful working with or extending the functionality of storage virtualization. As such, DigitalDoors was conceived as an open, accessible, and robust omni-platform solution to survivability, identity, privacy, confidentiality, and anonymity. In contrast to

existing restrictive security technologies that work by excluding access to data, surrounding it with concentric walls, and searching it repetitively for embedded viruses, the DigitalDoors system is envisioned to open data to wider access, enable information sharing, and effectively create a granular owner, organization, and supervisory multi-level access control over that data.

AA.7.1.1 Security

[1610] DigitalDoors security is predicated on the process of selecting critical and important information, extracting it, and dispersing that information to remote randomized locations. The context between the source, the declassified result, and the extractions is broken, thereby preventing efforts to understand the information. The more information processed in such a way, the greater the innate security. Generally, access to the process is by user authorization only to preclude destruction of valid source data.

[1611] When a password (Any reference to a password, user, administrator, process, file, database, recovery may, recovery file, and key (for encryption) is equally relevant to two or multiple objects.) is entered by a user or administrator, it is encrypted. All validation of passwords to the database of passwords is performed in that encrypted format. No clear text is delivered over any non-local connection. This same process is applied for all encrypted data, recovery data, location pointers, storage passwords, and extractions. Lost passwords cannot be recovered within DigitalDoors, but can be reset with the help of the administrator. We may include an automatic facility for password management in the future subject to security concerns.

[1612] All encryption keys created for an organization, group, user, or as one-time pads are encrypted by the encryption key and method configured by the administrator. This is stored in that form within all encrypted data records, databases, objects, and files. This precludes attacks against the network

transmission links. Only direct access to these objects with both plain text keys and corresponding encryption would provide a means to break the administrator-configured key or user keys.

[1613] User access is established by the administrator at the time when an account with user name and password is created. Optional per-user encryption keys, methods, remote access, dispersion access, and storage passwords can be established. Administrator access is by name and password, and access can be shared in that two or more administrators must agree on access. Administrator access does not grant user level access or vice versa. Administrator access configuration also required the creation of a unique encryption key for all user level and data level passwords. It is the same password, but we can create layers at any time in the future. The limitation is that any change in the unique encryption key requires a hierarchical recovery of data and re-encryption to apply a changed key. This process would preclude ordinary access until the process completed. Different administrators will have different names and passwords, and shared key access can be enforced by biometric methods to prevent a single administrator masquerading as multiple fraudulent people.

[1614] While declassified information is useful as a byproduct of the process, DigitalDoors also provides a granular means to reconstitute the source in whole or in part. Reconstitution requires two critical items: a starting point linkage between the declassified data and the recovery map and the keystone recovery map itself. Access to this information is protected in various ways. Declassified file names are typically changed so as to protect the meaning, content, and context of the source information. The recovery map name has no relationship to the declassified file and the original source. Linkage is retained in an optional database (or distributed databases), or not at all. Loss of the linkage precludes reconstitution, except by human intelligence, if feasible at all. Generally, access to the database is by user authorization. Granularity means that access to any one source, and in fact any single extraction, does

not provide insight into any other sources or extractions. Break-ins preclude access to other information as it is compartmentalized and widely dispersed with other security barriers.

[1615] The declassified data is the template for reconstitution. It is assumed to be a public file, but access to it and its existence can be limited and masked with standard network security and access control methods. When the declassified data is part of an email message, it is presumed to be public anyway. Access to the recovery map for an email message is the subject line of the message. A number defines the originating organization and a serialized number for that message. Access for reconstitution is granted per message, per user, and per extraction with multi-step Internet-enabled identification and authorization process. The email user must be setup to access a unique site for authentication. Each organization can maintain one or multiple sites, or subscribe to a public service. This validates the user, the user's allowance to that organization's email, and specific recovery information for that message. Reconstitution is enabled in whole or in part. Normally, the user would see the declassified message only. If the user had installed the DigitalDoors for Email add-in, the user would be prompted automatically to log in. After successful identification and authorization, the email message (and relevant) others could be automatically reconstituted in whole or in part as allowed by time of day, location, user, and other access rights.

[1616] The keystone is the recovery map. This data object (or passive and active file as in code) contains offset and length information for each extraction and information about the extraction itself. This object can exist as one or many sequential maps. The recovery information can be contained by this same file or distributed to other files. The offset and length information describes the position of the extraction, but also any substitution of equal, null, shorter, or longer length. Note that substitutions do not provide insight into the extraction or add to context unless a category tag is chosen as a substitution.