

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

BRIGHT DATA LTD.,

Petitioner

v.

OXYLABS, UAB,

Patent Owner

Case IPR2023-01442

Patent No 7,020,667

DECLARATION OF KEVIN C. ALMEROTH, Ph.D.

TABLE OF CONTENTS

I. INTRODUCTION	1
II. PROFESSIONAL BACKGROUND.....	1
III. COMPENSATION	13
IV. LEGAL CONSIDERATION.....	13
A. Anticipation	14
B. Claim Interpretation.....	14
C. Overview of Task & Basis for Opinions	15
D. Level of Skill in the Relevant Art.....	16
V. OVERVIEW OF THE '667 PATENT	17
A. Operational Architecture and Navigation.....	17
B. Initialization Stage.....	20
C. Automatic Execution Stage	22
D. Independent Claims 1, 16, and 20	22
VI. OVERVIEW OF THE PRIOR ART	24
A. U.S. Patent No. 7,020,667 (“DaCosta”).....	25
1. Navigation Module.....	28
2. Extraction Module.....	31
3. Smart Server	33
B. Kraft (EX1005).....	35
VII. GROUND 1: THE PETITION FAILS TO SHOW ANTICIPATION OF CLAIMS 1, 16, AND 20 BY DACOSTA	37
A. Claims 1, 16, and 20 are Novel Over DaCosta	37
1. [1g], [16g], [20g]: selecting the target database and target locations in the target database where the at least one of the data elements is to be stored in the target database.....	37
2. [1i], [16i], [20i] – “setting a schedule frequency for transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database”	45

3. [1h], [16h], [20h] “saving the source locations and target locations of the at least one of the parsed data elements in a location file in the target database;” ...	48
4. [16b], [16j]: First and second sets of instruction codes	53
VIII. GROUND 2: THE PETITION FAILS TO SHOW ANTICIPATION OF CLAIMS 1, 16, AND 20 BY KRAFT	53
A. Claims 1, 16, and 20 are Novel Over Kraft.....	53
1. [16b], [16j]: First and second sets of instruction codes	54
2. [1i], [16i], [20i]: – “setting a schedule frequency for transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database;”	54
3. [1e], [16e], [20e]: – “selecting at least one of the data elements from the list to be stored in the target database”	57
4. [1d], [16d], [20d]: “parsing selected data elements in the source file into a list of data elements;”	58
IX. CONCLUSION.....	63

I. INTRODUCTION

1. My name is Dr. Kevin C. Almeroth, Ph.D. I have been retained by Oxylabs, UAB as an independent expert consultant in this inter partes review (“IPR”) proceeding before the United States Patent and Trademark Office (“PTO”).

II. PROFESSIONAL BACKGROUND

2. All of my opinions stated in this Declaration are based on my own personal knowledge and professional judgment. In forming my opinions, I have relied on my knowledge and experience in designing, developing, researching, and teaching the technology referenced in this Declaration.

3. I am over 18 years of age and, if I am called upon to do so, I would be competent to testify as to the matters set forth herein. I understand that a copy of my current curriculum vitae, which details my education and professional and academic experience, is being submitted as EX2002. The following provides a brief overview of some of my experience that is relevant to the matters set forth in this Declaration.

4. I am currently a Professor Emeritus in the Department of Computer Science at the University of California, Santa Barbara (UCSB). While active at UCSB, I held faculty appointments and was a founding member of the Computer Engineering (CE) Program, Media Arts and Technology (MAT) Program, and the Technology Management Program (TMP). I also served as the Associate Director

of the Center for Information Technology and Society (CITS) from 1999 to 2012. I have been a faculty member at UCSB since July 1997.

5. I hold three degrees from the Georgia Institute of Technology: (1) a Bachelor of Science degree in Information and Computer Science (with minors in Economics, Technical Communication, and American Literature) earned in June 1992; (2) a Master of Science degree in Computer Science (with specialization in Networking and Systems) earned in June 1994; and (3) a Doctor of Philosophy (Ph.D.) degree in Computer Science (Dissertation Title: Networking and System Support for the Efficient, Scalable Delivery of Services in Interactive Multimedia System, minor in Telecommunications Public Policy) earned in June 1997. During my education, I have taken a wide variety of courses as demonstrated by my minors. My undergraduate degree also included a number of courses more typical of a degree in electrical engineering including digital logic, signal processing, and telecommunications theory.

6. One of the major concentrations of my research over the past 30+ years has been the delivery of multimedia content and data between computing devices, including various network architectures. In my research, I have studied large-scale content delivery systems, and the use of servers located in a variety of geographic locations to provide scalable delivery to hundreds or thousands of users simultaneously. I have also studied smaller-scale content delivery systems in which

content is exchanged between individual computers and portable devices. My work has emphasized the exchange of content more efficiently across computer networks, including the scalable delivery of content to many users, mobile computing, satellite networking, delivering content to mobile devices, and network support for data delivery in wireless networks.

7. In 1992, the initial focus of my research was on the provision of interactive functions (e.g., VCR-style functions like pause, rewind, and fast-forward) for near video-on-demand systems in cable systems; in particular, how to aggregate requests for movies at a cable head-end and then how to satisfy a multitude of requests using one audio/video stream broadcast to multiple receivers simultaneously. This research has continually evolved and resulted in the development of techniques to scalably deliver on-demand content, including audio, video, web documents, and other types of data, through the Internet and over other types of networks, including over cable systems, broadband telephone lines, and satellite links.

8. An important component of my research has been investigating the challenges of communicating multimedia content, including video, between computers and across networks including the Internet. Although the early Internet was used mostly for text-based, non-real time applications, the interest in sharing multimedia content, such as video, quickly developed. Multimedia-based

applications ranged from downloading content to a device to streaming multimedia content to be instantly used. One of the challenges was that multimedia content is typically larger than text-only content, but there are also opportunities to use different delivery techniques since multimedia content is more resilient to errors. I have worked on a variety of research problems and used a number of systems that were developed to deliver multimedia content to users. One content-delivery method I have researched is the one-to-many communication facility called “multicast,” first deployed as the Multicast Backbone, a virtual overlay network supporting one-to-many communication. Multicast is one technique that can be used on the Internet to provide streaming media support for complex applications like video-on-demand, distance learning, distributed collaboration, distributed games, and large-scale wireless communication. The delivery of media through multicast often involves using Internet infrastructure, devices and protocols, including protocols for routing and TCP/IP.

9. Starting in 1997, I worked on a project to integrate the streaming media capabilities of the Internet together with the interactivity of the web. I developed a project called the Interactive Multimedia Jukebox (IMJ). Users would visit a web page and select content to view. The content would then be scheduled on one of a number of channels, including delivery to students in Georgia Tech dorms delivered

via the campus cable plant. The content of each channel was delivered using multicast communication.

10. In the IMJ, the number of channels varied depending on the capabilities of the server including the available bandwidth of its connection to the Internet. If one of the channels was idle, the requesting user would be able to watch their selection immediately. If all channels were streaming previously selected content, the user's selection would be queued on the channel with the shortest wait time. In the meantime, the user would see what content was currently playing on other channels, and because of the use of multicast, would be able to join one of the existing channels and watch the content at the point it was currently being transmitted.

11. The IMJ service combined the interactivity of the web with the streaming capabilities of the Internet to create a jukebox-like service. It supported true Video-on-Demand when capacity allowed but scaled to any number of users based on queuing requested programs. As part of the project, we obtained permission from Turner Broadcasting to transmit cartoons and other short-subject content. We also connected the IMJ into the Georgia Tech campus cable television network so that students in their dorms could use the web to request content and then view that content on one of the campus's public access channels.

12. More recently, I have also studied issues concerning how users choose content, especially when considering the price of that content. My research has examined how dynamic content pricing can be used to control system load. By raising prices when systems start to become overloaded (i.e., when all available resources are fully utilized) and reducing prices when system capacity is readily available, users' capacity to pay as well as their willingness can be used as factors in stabilizing the response time of a system. This capability is particularly useful in systems where content is downloaded or streamed on-demand to users.

13. As a parallel research theme, starting in 1997, I began researching issues related to wireless devices and sensors. In particular, I was interested in showing how to provide greater communication capability to "lightweight devices," i.e., small form-factor, resource-constrained (e.g., CPU, memory, networking, and power) devices. Starting in 1998, I published several papers on my work to develop a flexible, lightweight, battery-aware network protocol stack. The lightweight protocols we envisioned were similar in nature to protocols like Bluetooth, Universal Plug and Play (UPnP) and Digital Living Network Alliance (DLNA).

14. From this initial work, I have made wireless networking—including ad hoc, mesh networks and wireless devices—one of the major themes of my research. My work in wireless network spans the protocol stack from applications through to the encoding and exchange of data at the data link and physical layers.

15. At the application layer, even before the large-scale “app stores” were available, my research looked at building, installing, and using apps for a variety of purposes, from network monitoring to support for traditional computer-based applications (e.g., content retrieval) to new applications enabled by ubiquitous, mobile devices. For example, my research has looked at developing applications for virally exchanging and tracking “coupons” through “opportunistic contact” (i.e., communication with other devices coming into communication range with a user). In many of the courses I have taught there is a project component. Through these projects I have supervised numerous efforts to develop new “apps” for download and use across a variety of mobile platforms.

16. Toward the middle of the protocol stack, my research also looked to build wireless infrastructure support to enable communication among a set of mobile devices unaided by any other kind of network infrastructure. These kinds of networks are useful either in challenged network environments (e.g., when a natural disaster has destroyed existing infrastructure) or when suitable support for network communication never existed. The deployment of such networks (or even the use of traditional network support) are critical to support services like disaster relief, catastrophic event coordination, and emergency services deployment.

17. Yet another theme is monitoring wireless networks, in particular different variants of IEEE 802.11 compliant networks, to (1) understand the

operation of the various protocols used in real-world deployments, (2) use these measurements to characterize use of the networks and identify protocol limitations and weaknesses, and (3) propose and evaluate solutions to these problems. I have successfully used monitoring techniques to study wireless data link layer protocol operation and to improve performance by enhancing the operation of such protocols. For wireless protocols, this research includes functions like network acquisition and channel bonding. Protecting networks, including their operation and content, has been an underlying theme of my research almost since the beginning of my research career. Starting in 2000, I have been involved in several projects that specifically address security, network protection, and firewalls. After significant background work, a team on which I was a member successfully submitted a \$4.3M grant proposal to the Army Research Office (ARO) at the Department of Defense to propose and develop a high-speed intrusion detection system. Key aspects of the system included associating streams of packets and analyzing them for viruses and other malware. Once the grant was awarded, we spent several years developing and meeting the milestones of the project. A number of my students worked on related projects and published papers on topics ranging from intrusion detection to developing advanced techniques to be incorporated into firewalls. I have also used firewalls, including their associated malware detection features, in developing

techniques for the classroom to ensure that students are not distracted by online content.

18. Recent work ties some of the various threads of my past research together. I have investigated content delivery in online social networks and proposed reputation management systems in large-scale social networks and marketplaces. On the content delivery side, I have looked at issues of caching and cache placement, especially when content being shared, and the cache has geographical relevance. We were able to show that effective caching strategies can greatly improve performance and reduce deployment costs. Our work on reputation systems showed that reputations have economic value, and as such, creates a motivation to manipulate reputations. In response, we developed a variety of solutions to protect the integrity of reputations in online social networks. The techniques we developed for content delivery and reputation management were particularly relevant in peer-to-peer communication and recommendations for downloadable “apps.”

19. As an important component of my research program, I have been involved in the development of academic research into available technology in the marketplace. One aspect of this work is my involvement in the Internet Engineering Task Force (IETF). The IETF is a large and open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. I have been

involved in various IETF groups including many content delivery-related working groups like the Audio Video Transport (AVT) group, the MBone Deployment (MBONED) group, Source Specific Multicast (SSM) group, the Inter-Domain Multicast Routing (IDMR) group, the Reliable Multicast Transport (RMT) group, the Protocol Independent Multicast (PIM) group, etc. I have also served as a member of the Multicast Directorate (MADDOGS), which oversaw the standardization of all things related to multicast in the IETF. Finally, I was the Chair of the Internet2 Multicast Working Group for seven years.

20. My involvement in the research community extends to leadership positions for several academic journals and conferences. I am the co-chair of the Steering Committee for the ACM Network and System Support for Digital Audio and Video (NOSSDAV) workshop and on the Steering Committees for the International Conference on Network Protocols (ICNP), ACM Sigcomm Workshop on Challenged Networks (CHANTS), and IEEE Global Internet (GI) Symposium. I have served or am serving on the Editorial Boards of IEEE/ACM Transactions on Networking, IEEE Transactions on Mobile Computing, IEEE Network, ACM Computers in Entertainment, AACE Journal of Interactive Learning Research (JILR), and ACM Computer Communications Review. I have co-chaired a number of conferences and workshops, including the IEEE International Conference on Network Protocols (ICNP), IEEE Conference on Sensor, Mesh and Ad Hoc

Communications and Networks (SECON), International Conference on Communication Systems and Networks (COMSNETS), IFIP/IEEE International Conference on Management of Multimedia Networks and Services (MMNS), the International Workshop On Wireless Network Measurement (WiNMee), ACM Sigcomm Workshop on Challenged Networks (CHANTS), the Network Group Communication (NGC) workshop, and the Global Internet Symposium, and I have served on the program committees for numerous conferences.

21. Furthermore, in the courses I have taught at UCSB, a significant portion of my curriculum covered aspects of the Internet and network communication, including the physical and data link layers of the Open System Interconnect (OSI) protocol stack, and standardized protocols for communicating across a variety of physical media such as cable systems, telephone lines, wireless, and high-speed Local Area Networks (LANs). The courses I have taught also cover most major topics in Internet communication, including data communication, multimedia encoding, and mobile application design. My research and courses have covered a range of physical infrastructures for delivering content over networks, including cable, Integrated Services Digital Network (ISDN), Ethernet, Asynchronous Transfer Mode (ATM), fiber, and Digital Subscriber Line (DSL). For a complete list of courses I have taught, see my curriculum vitae (CV) (EX2002).

22. In addition, I co-founded a technology company called Santa Barbara Labs that was working under a sub-contract from the U.S. Air Force to develop very accurate emulation systems for the military's next generation internetwork. Santa Barbara Labs' focus was in developing an emulation platform to test the performance characteristics of the network architecture in the variety of environments in which it was expected to operate, and, in particular, for network services including IPv6, multicast, Quality of Service (QoS), satellite-based communication, and security. Applications for this emulation program included communication of a variety of multimedia-based services, including video conferencing and video-on-demand.

23. In addition to having co-founded a technology company myself, I have worked for, consulted with, and collaborated with companies for nearly 30 years. These companies range from well-established companies to start-ups and include IBM, Hitachi Telecom, Turner Broadcasting System (TBS), Bell South, Digital Fountain, RealNetworks, Intel Research, Cisco Systems, and Lockheed Martin.

24. Through my graduate education, leadership with CITS, involvement in TMP, role in the development of the Internet2 infrastructure, and consulting with ISPs, I have gained a strong understanding in the role of the Internet in our society and the challenges of deploying large-scale production networking infrastructure. CITS, since its inception, has looked at the role of the Internet in society, including

how the evolution of technology have created communication opportunities and challenges, including, for example through disruptive technologies like P2P. TMP looks to focus on non-purely technical issues, including, for example, state-of-the-art business methods, strategies for successful technology commercialization, new venture creation, and best practices for fostering innovation. Through my industry collaborations and Internet2 work, I have developed significant experience in the challenges of deploying, monitoring, managing, and scaling communication infrastructure to support evolving Internet services like streaming media, conferencing, content exchange, social networking, and e-commerce.

25. I am a Member of the Association of Computing Machinery (ACM) and a Fellow of the Institute of Electrical and Electronics Engineers (IEEE).

III. COMPENSATION

26. My work on this matter is being billed at my customary rate of \$750 an hour. Also, I am being reimbursed for reasonable expenses I incur in relation to my services. I have no pecuniary interest in the outcome of this proceeding. I understand I will be paid regardless of the outcome of any proceeding in which my work is used.

IV. LEGAL CONSIDERATION

27. I am not an attorney. My understanding of the law is based on information provided by counsel for Oxylabs, UAB.

A. Anticipation

28. I understand that a reference anticipates a claim if it discloses each and every element recited in the claim, arranged as in the claim, so as to enable a person of ordinary skill in the art (POSITA) to make and use the claimed invention without the need for undue experimentation in light of the general knowledge available in the art. I understand that a U.S. Patent document is presumed to have sufficient description to include sufficient detail for a person of ordinary skill in the art to make and use the subject matter that document describes.

29. I also understand that a claim limitation may be present in an alleged prior art reference either expressly or inherently, but inherency may be established only if such limitation is necessarily present in the reference. I understand inherency cannot be established by probabilities or possibilities, and the mere fact that a certain thing may result from a given set of circumstances is not sufficient to establish inherency.

B. Claim Interpretation

30. I understand that a claim term is interpreted according to its ordinary and customary meaning as a POSITA would have understood the term in light of the surrounding claim language, other claims, the specification, and the patent's prosecution history, which are referred to as intrinsic evidence. I also understand that prior art references cited in the patent's prosecution history are considered

intrinsic evidence. I further understand that evidence outside the patent and its prosecution history (e.g., dictionaries and technical articles), may inform the context in which a POSITA would have understood the claims of a patent. I understand this ordinary and customary meaning applies absent unique circumstances, such as where a patent clearly expresses an intent to set forth a special meaning for a term.

C. Overview of Task & Basis for Opinions

31. I have been asked to review the '667 Patent. I have been asked to provide opinions related to certain issues from the perspective of a person of ordinary skill, having knowledge of the relevant art, as of July 18, 2002, and—except where otherwise noted—the opinions stated in this declaration are from that perspective. The qualifications and abilities of such a person are described in paragraph 35 below.

32. More particularly, I have been informed that this proceeding is currently at a preliminary stage during which the Patent Trial and Appeal Board (the Board) will decide whether the Petition establishes a reasonable likelihood of at least one claim of the '667 Patent being invalid. I also understand that the Board may review the Petition under a stricter “compelling merits” standard that requires the Petition to make a stronger showing of invalidity; however, my analysis is not performed using this higher standard. At this preliminary stage, I have been asked to focus my analysis on the Petition’s proposed invalidity theories for the '667 Patent’s independent claims 1, 16, and 20.

33. My opinions are based on my education, training, and experience as well as items that I reviewed to prepare my opinions, including the Petition, the documents listed in the Exhibit List included with the Petition, and any other items that I reference in my below analysis.

34. I understand that other issues may arise that require further explanation, and I will provide that explanation if appropriate. As a result, I respectfully reserve the right to update and supplement this declaration and the information and opinions provided herein.

D. Level of Skill in the Relevant Art

35. I understand that a patent must be analyzed from the perspective of a person of ordinary skill in the art (“POSITA”) as of the time of invention of the patent, which is typically considered to be the patent’s earliest filing date. While I do not necessarily agree with every aspect of Dr. Williams’ explanation of the level of ordinary skill in the art, I do not currently believe the aspects of that explanation I would change or clarify would impact the analysis set forth in this declaration. For purposes of this declaration, my opinions are therefore provided from the perspective of the POSITA described in paragraph 30 of Dr. Williams’ declaration (EX1003) as of July 18, 2002, which is produced below:

either: (1) a Bachelor Degree or higher in the field of electrical engineering, computer engineering, or computer science and at least

one year of related practical experience; or (2) at least four years of practical experience in data management.

Such a POSITA would have been familiar with the concepts and types of computing networks and processes described below in paragraphs.

V. OVERVIEW OF THE '667 PATENT

36. The '667 Patent discloses a “data collection and retrieval system and associated method allow the capture and replication of data presented at various Web pages into a database application.” EX1001, Abstract. As described in more detail below, the system can include an initialization stage—which provides the user interface which allows the user to select the source file that contains data the user wishes to copy, target database that will receive the data, and timing criteria for automatic transfer of data—and an automatic execution stage that automatically transfers the data from the source file to the target database as instructed by the user in the initialization stage. *Id.*

A. Operational Architecture and Navigation

37. An illustrative operating environment in which the data collection and retrieval system can be used is depicted in the component diagram of FIG. 1 (inset below). EX1001. 2:50-53.

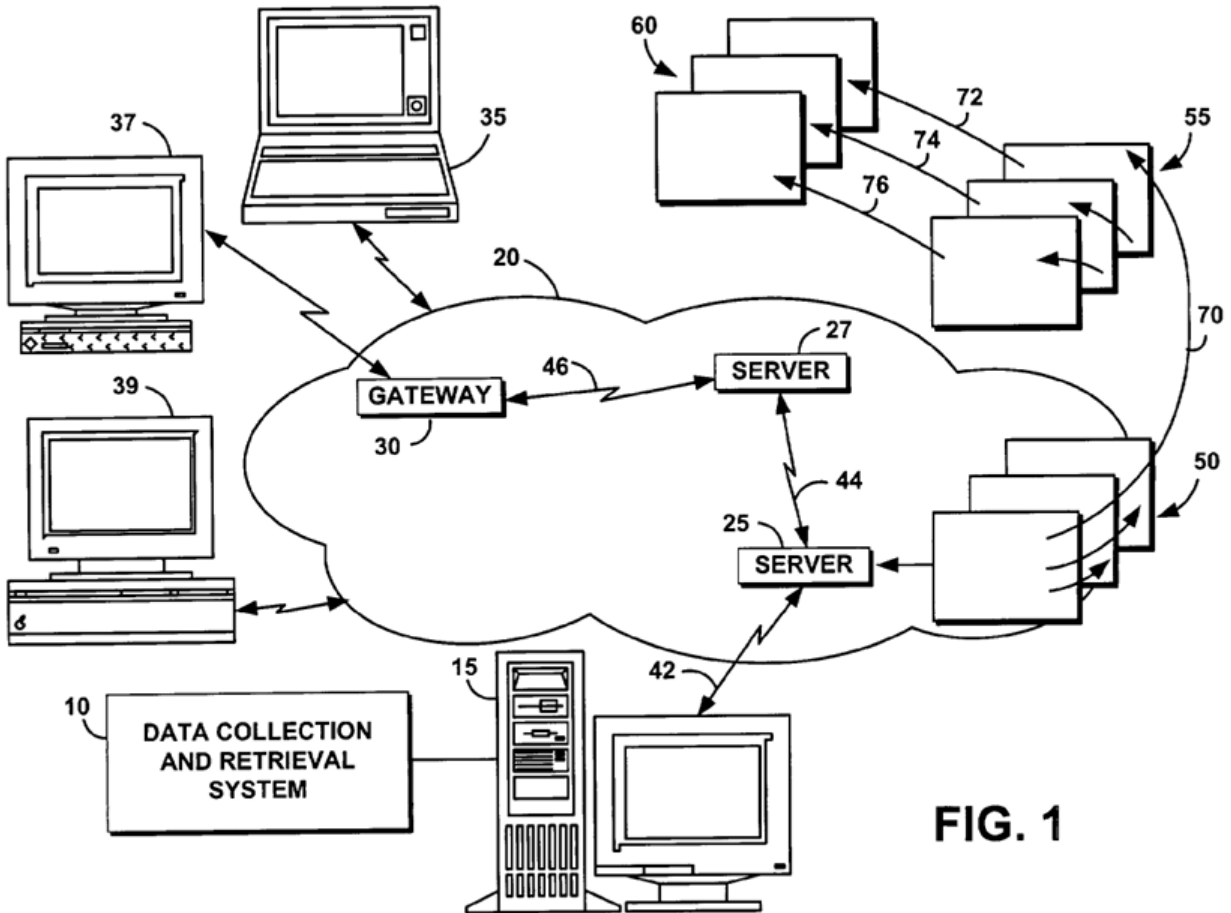


FIG. 1

As shown in FIG. 1, “servers 25, 27 provide access to stored information such as hypertext or Web documents indicated generally at 50, 55, and 60” that can include hypertext links (e.g., 70, 72, 74, 76) to other webs sites or documents that are stored by various Web servers. EX1001, 4:6-16. In this way, the user can navigate to various webpages “until a webpage ... is found that contains the data in which the user is interested.” *Id.* 4:17-26.

38. An example of this navigation path is shown in FIG. 2, below:

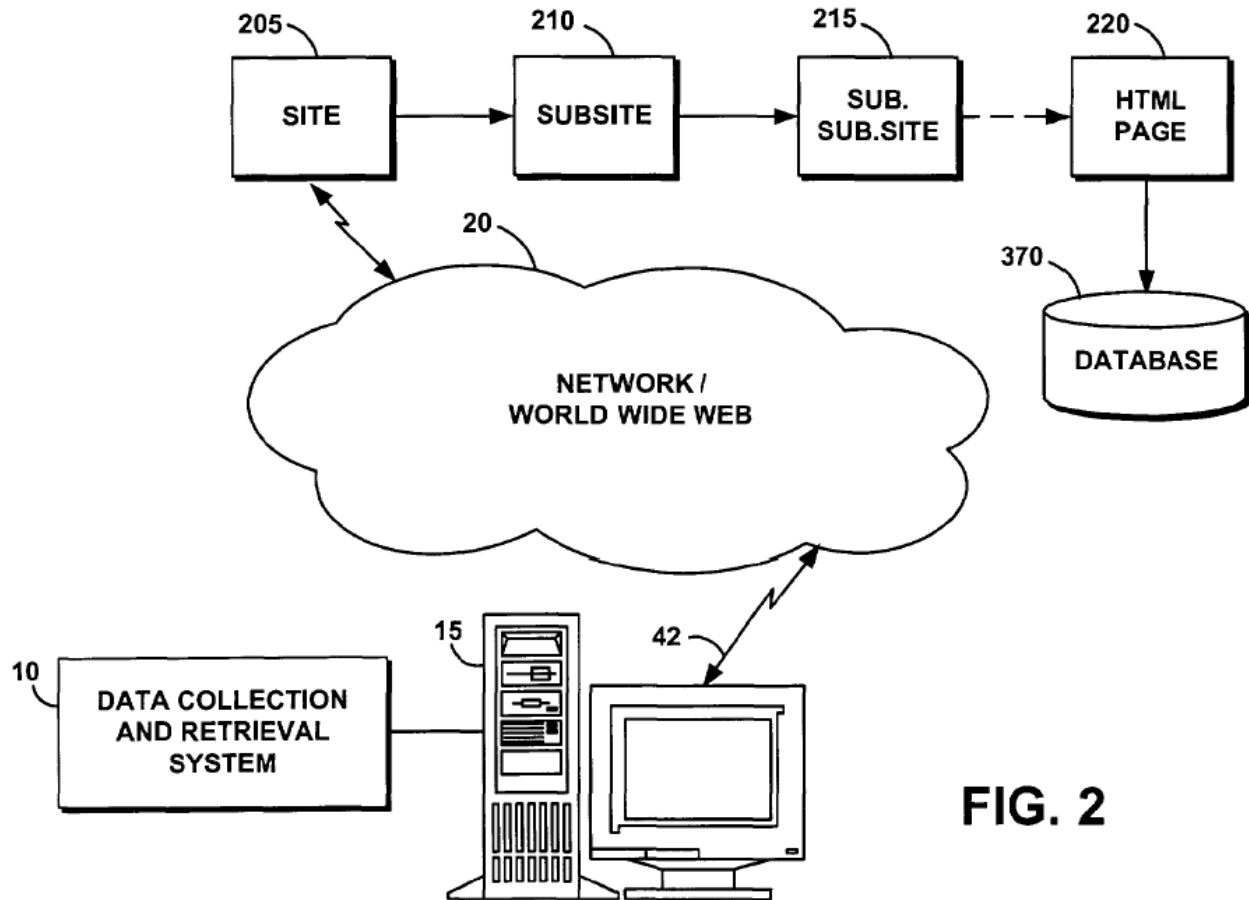


FIG. 2

EX1001, 2:53-59 (showing the manner in which the data collection and retrieval system accesses data residing on various Web pages). As depicted, a user may initiate communication with the world wide web to access a website 205, which can be an internet portal. EX1001, 4:18-26. The user can then navigate through different sites on the internet by selecting different subsites 210, 215 “until a webpage such as webpage 220 is found that contains the data in which the user is interested to transcribe (or copy) to a database or storage 370.” *Id.* Once a user has navigated to (or otherwise identified) the desired webpage, the user can implement the initialization stage 305 and automatic execution stage.

B. Initialization Stage

39. An illustrative method of performing the initialization stage of the system is shown in the flow chart found in FIG. 4 (inset below) of the '667 Patent.

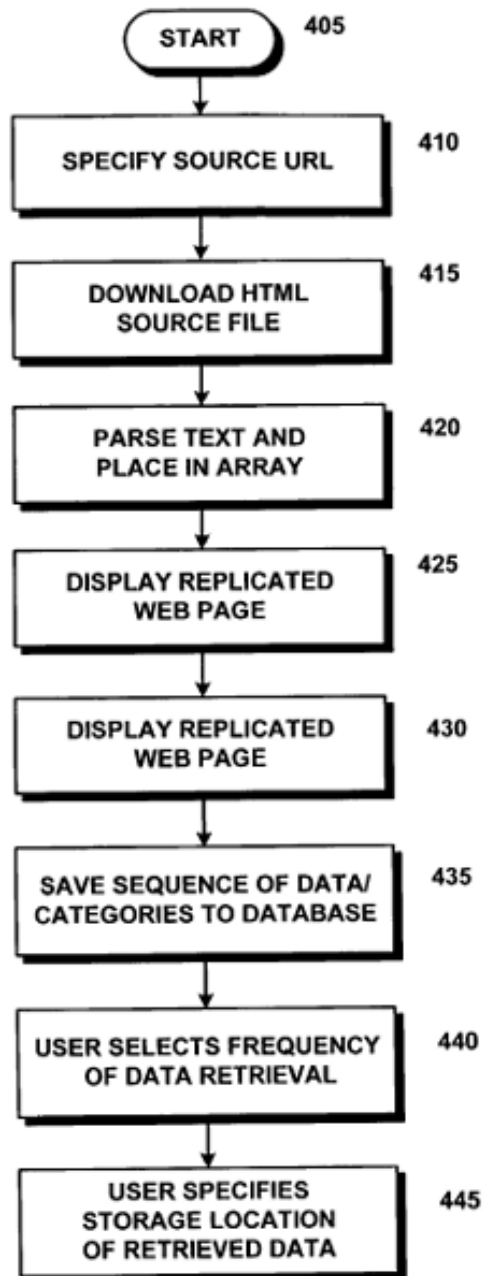


FIG. 4

EX1001, 2:61-62. This illustrative method starts with step 410, where “the user specifies the URL for an HTML page 220 that contains data to be transcribed to the target database 370.” EX1001, 4:46-49. This step corresponds to the navigation to and selection of the web page 220, as shown in FIG. 2. EX1001, 4:46-49 (In step 410 the “user selects the Web page 220, as shown by FIG. 2, and saves the Web page URL in system 10.”).

40. After the user has found the desired webpage (e.g., 220), the system will then “download the HTML source file 320 for the Web page 220 in step 415.” EX1001, 4:56-57. In step 420 a parser 415 will then process the source file 320, tokenize the HTML format of the Web page 220, and place it in a data array 330 in step 420. EX1001, 4:56-59. The system may then replicate and display portions of the webpage in a refined user GUI 340. *Id.*, EX1001, 4:59-66, 5:6-11 (the user can access the “replicated Web page” on the refined user input GUI (340 of FIG. 3)). The user can then select data and or categories of interest (e.g., from the replicated webpage) in step 430 and then, in step 435, the system saves the location of the sequence of data or categories selected by the user and remembers that location so that it may retrieve corresponding data from the same location during successive data transfer. EX1001, 5:12-16. The user can then select “the frequency of data retrieval” at step 440 and specify where to store the retrieved data in step 445. EX1001, 5:17-29.

C. Automatic Execution Stage

41. After the initialization stage 305 is complete, the system can then perform the automatic retrieval of data. EX1001, 5:30-35. An illustrative example of operation of the automatic execution stage is depicted in FIGs. 5A-5B. The data retrieval process can begin after expiration of the time specified by the user in the initialization stage. EX1001, 5:35-37. For example, the system may access the identified webpage (e.g., Web page 220) according to the frequency of data retrieval set by the user. *Id.* The system may then parse the source file of the identified webpage, extract data from the webpage that corresponds to the data selected by the user, and transfer that data to a database (e.g., 370) that was previously specified by the user. EX1001, 5:37-45. This process is then repeated according to the frequency that was previously specified by the user. *Id.*, 5:45-47.

D. Independent Claims 1, 16, and 20

42. Independent claim 1 of the '667 Patent is directed to a “method of selectively retrieving data from a source file ... and automatically collecting retrieved data in a target database.” EX1001, 6:63-65. FIGs. 4 and 5A of the '667 Patent, described above, show partial flowchart diagrams of an illustrative method of retrieving and collecting data for some of the steps covered by independent claim 1.

43. Claim 1 recites **[1a]** A computer-implemented method of selectively retrieving data from a source file on a public network and automatically collecting retrieved data in a target database, comprising:

[1b] initializing a data retrieval stage by:

[1c] selecting the source file;

[1d] parsing selected data elements in the source file into a list of data elements;

[1e] selecting at least one of the data elements from the list to be stored in the target database;

[1f] identifying source locations of the selected at least one of the data elements;

[1g] selecting the target database and target locations in the target database where the at least one of the data elements is to be stored in the target database;

[1h] saving the source locations and target locations of the at least one of the parsed data elements in a location file in the target database;

[1i] setting a schedule frequency for transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database; and

[1j] automatically, selectively parsing new data elements corresponding to the at least one of the parsed data elements previously selected during the initialization of the data retrieval stage, pursuant to the schedule frequency, by transferring the selectively parsed new data elements from the saved source locations in the source file to the saved target locations in the target database;

EX1001, 6:63-7:25.

44. Claim 16 is directed to **[16a]** a computer program product having instruction codes for selectively retrieving data from a source file on a public network and for automatically collecting retrieved data in a target database that includes limitations similar to the Limitations [1a]-[1j] of claim 1. Claim 16 recites **[16b]** “a first set of instruction codes for initializing a data retrieval stage by:” performing Limitations [1a]-[1i] and **[16j]** “second set of instruction codes for” performing Limitation [1j]. EX1001, 7:61-8:26.

45. Claim 20 is directed to **[20a]** a computer-implemented system for selectively retrieving data from a source file on a public network and for automatically collecting retrieved data in a target database that includes limitations similar to the Limitations [1a]-[1j] of claim 1. Claim 16 recites **[16b]** means for initializing a data retrieval stage and **[16j]** means for automatically, selectively parsing new data elements. EX1001, 8:34-64.

VI. OVERVIEW OF THE PRIOR ART

46. Dr. Williams primarily relies on two cited references, each directed to different problems than those faced by the inventors of the '667 patent. *See* EX1003, ¶62.

A. U.S. Patent No. 7,020,667 (“DaCosta”)

47. U.S. Patent No. 7,020,667 is titled “System for providing database functions for multiple internet sources” and lists as its first inventor Gerson Francis DaCosta. I will refer to this reference as “DaCosta.”

48. DaCosta discloses a system for recording scripts to extract data through the internet. EX10004, Abstract. The system enables a user to record a navigation path to arrive at a target page and a target pattern to analyze information on the target page. *Id.* With these recorded scripts, the system can then automatically access the target page according to the recorded sequence and then process information according to the target pattern. *Id.* DaCosta provides a summary of this process at column 5, lines 10-20:

According to the present invention and referring now to the figures, wherein like reference numerals identify like elements of the various embodiments of the invention, one can automatically navigate to a plurality of web site destinations, extract specified information based upon taught schemas, process the extracted data according to customizable scripts, integrate information from other applications such as Microsoft (“MS”) Word, Excel or Access, view the final output using a browser such as Microsoft Internet Explorer, for example, and automatically repeat these steps in a scheduled manner or when requested, for example.

EX1004, 5:12-19.

49. FIG. 1 of DaCosta (inset below) depicts an embodiment of the invention that includes a navigation API 10 (i.e., navigation module 10) and an extraction API 20 (i.e., extraction module 20). EX1004, 5:37-67. “[T]he navigation API 10 provides two services: recording of navigation paths 12 and playback of that which has been recorded 14.” EX1004, 6:50-53. Additionally, the extraction API 20 “provides two services: recording of extraction patterns 22 and playback of extraction patterns 24.” EX1004, 6:53-55.

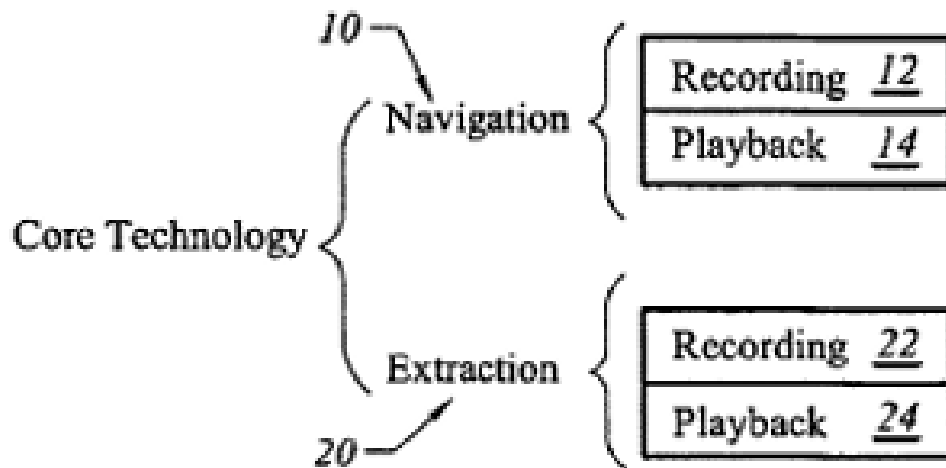


FIG. 1

The navigation module 10 is distinct from the extraction module 20 and each perform different processes. EX1004, 8:67-9:4 (The extraction module 20 “follows the navigation module 10”). The navigation module 10 records scripts and plays back those scripts to navigate to a specified webpage, such as one behind a login screen. EX1004, 5:37-51. The navigation module 10 does not process or identify

any specific information on the specified webpage and functions only to navigate to the specified webpage. On the other hand, the extraction module 20 operates on the detailed information of the specified webpage. *See* EX1004, 5:61-63 (“instead of recording/playing across web pages and web space, the extraction module 20 records and plays across elements within a single page.”), 8:61-65 (“[W]hile the navigation module [10] operates across multiple web pages or web space, the extraction module 20 operates at the HTML or web page or other web-accessible document level.”)

50. These distinct navigation and extraction modules 10 and 20 can be implemented via separate “plug-ins in a browser” such as internet explorer. EX1004, 10:18-20. FIG. 7 depicts an example of an application 220 that could implement these two separate modules (10, 20).

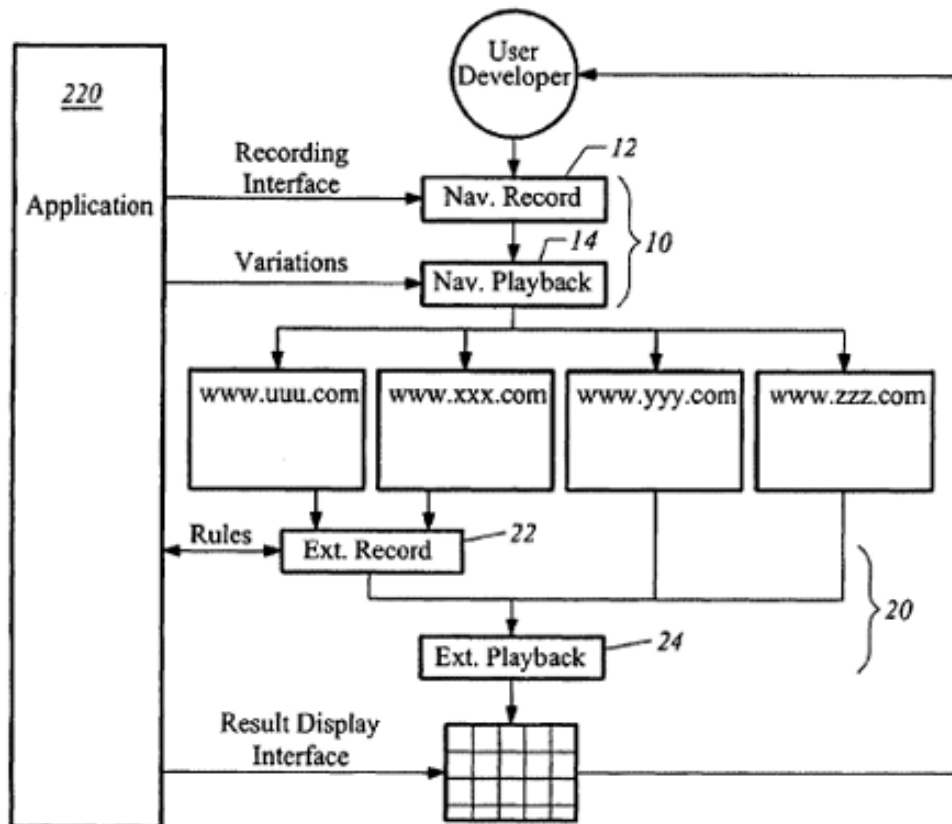


FIG. 7

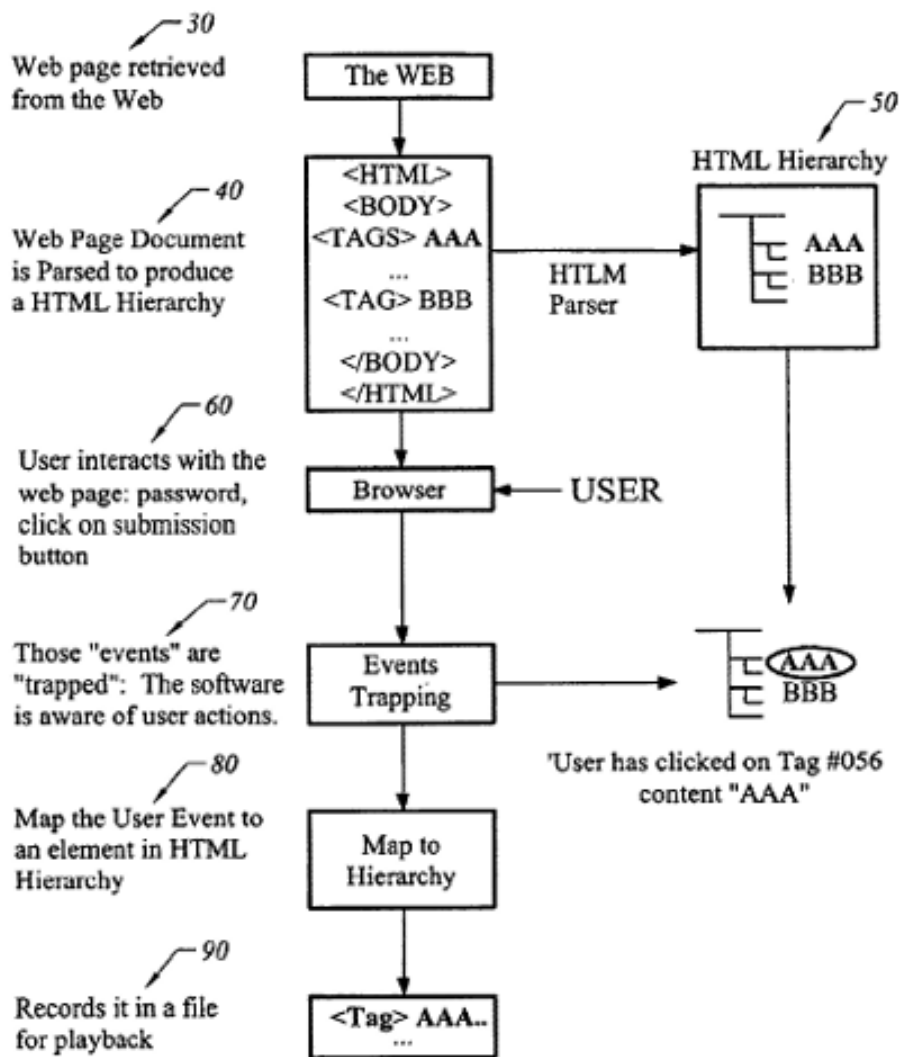
EX1004, 10:55-57. As shown, the navigation module 10 and extraction module 20 are depicted as separate components that perform distinct functions and different times.

1. Navigation Module

51. Navigation module 10 enables “a user to learn and store navigation paths to given web pages” EX1004, 5:37-40. “For example, if a web site requires a user to enter a login name and password to reach an orientation page and then asks for a set of preferences to go to specific web pages or other web-accessible documents of interest, it is an object of the present invention to enable a client

application to record this path once, then play it back many times” EX1004, 5:44-50.

52. FIG. 2 (inset below) depicts a process of recording navigation paths and with the navigation recording module 12. EX1004, 6:57-59.



Recording Navigation Script - Getting to a page

FIG. 2

In this process, a web page is provided to navigation module (e.g., via a web browser) and the web page is parsed to convert the HTML file into an object hierarchy. EX1004, 6:60-7:9. The user then interacts with the web page by clicking on links or buttons or otherwise entering data, which is recorded and saved in an Extensible Markup Language (“XML”) file. EX1004, 7:17-34.

53. The recorded navigation XML file can then be used as an input file to the navigation playback module 14. EX1004, 8:19-22. This “navigation XML file contains details about the recorded navigation in a format that can be read by the navigation playback module 14” and can include a series of steps that the user took to reach the page. EX1004, 7:41-51 (steps include information such as hyperlinks and form fields that are acted upon).

54. Playback module 14 performs “two different types of steps:” (1) form variations and (2) URL variations. These allow a user to change how the navigation file is played back. For example, the form parameters can be changed by “specifying new form parameters to replace the ones originally recorded using the navigation recording module 12.” EX1004, 7:57-63. These form variations may include updating form submissions (e.g., updating a password) or changing queries to access different webpages. DaCosta discloses an example of a form variation where the user changes queries—the title of a book—that are entered into an online bookstore to access webpages corresponding to the new queries. EX1004, 7:58-8:1. Following

this book example, the system allows a user to access several book titles with a single script by entering all the form variations into the playback module (e.g., Title 1, Title 2, Title 3, ..., etc.) so that the system can navigate to the corresponding webpages associated with each book title. *See* EX1004, 8:1-5 (“multiple queries can now be automatically run based upon a single exemplomatic query”).

2. *Extraction Module*

55. Extraction module 20 “follows the navigation module 10 by processing each final or target web page ... to extract relevant or select information or data therefrom” EX1004, 8:67-9:4. DaCosta’s FIG. 4 (inset below) depicts a process for recording an extraction script via extraction recording module 22. EX1004, 4:18-20.

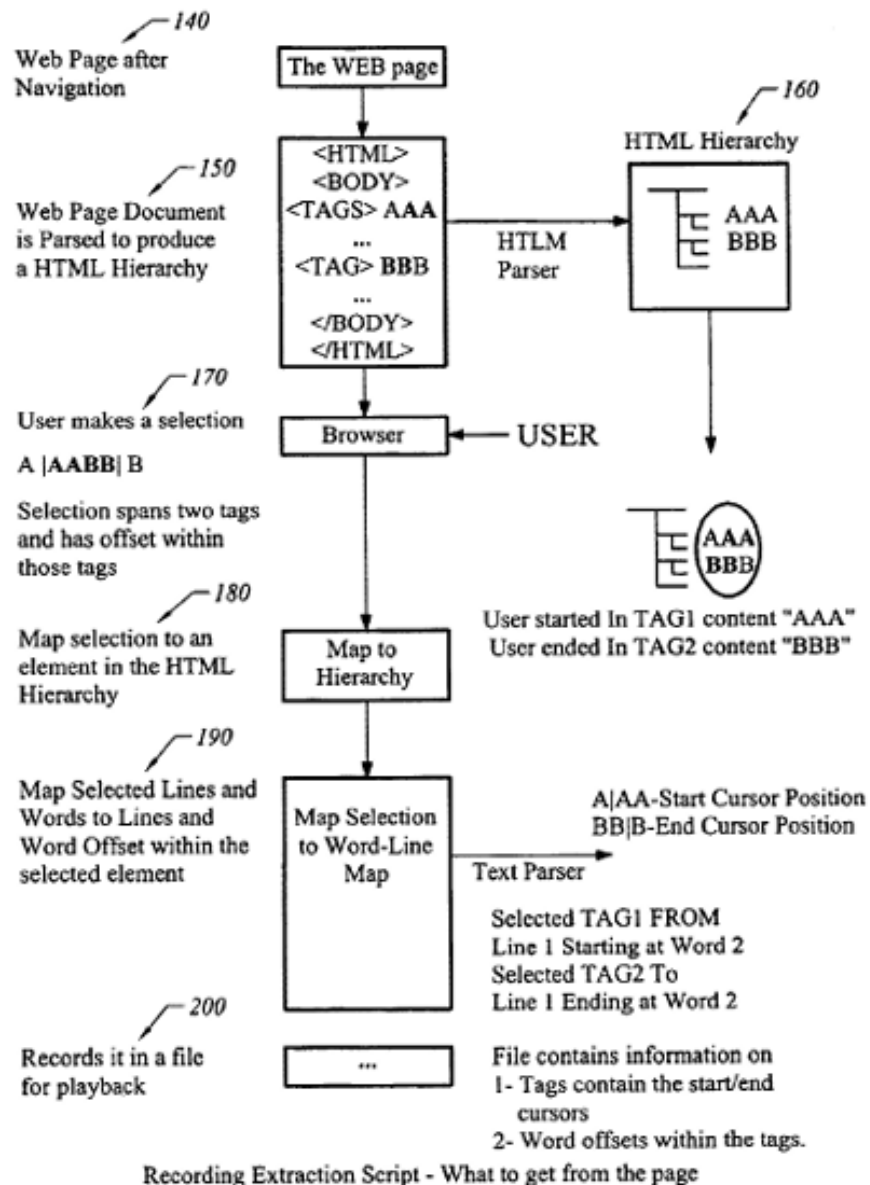


FIG. 4

56. In this process, a web page is parsed to generate the HTML hierarchy. EX1004, 9:53-55. A user may then select text from the web page, such as by highlighting a portion of that web page. EX1004, 9:57-59. This selection is then mapped, and the script is recorded as an XML file that will be input to the extraction

playback module 24. EX1004, 10:1-8. Extraction playback module 24 can then extract actual information from the web page as a text range 210. EX1004, 9:57-59.

57. This extracted data is then output to the browser of the user. EX1004, 5:12-20 (The user “view[s] the final output using a browser such as Microsoft Internet Explorer”), 10:45-47 (A calling application arranges the output “in a form suitable for display to the user”). As shown in FIG. 7 (reproduced above), the system that integrates the navigation module and extraction module send the result from the extraction module to the result “display interface.” See EX1004, 4:28-29. Similarly, FIG. 10 depicts an example in which a user may “view” this output on their “browser.”

3. *Smart Server*

58. In addition to the above-described modules, DaCosta discloses a smart server may also be provided to gather data on the Internet. EX1004, 18:8-10. Unlike the user-oriented scripts described above, the role of the “smart server is to fill a knowledge repository in *background mode*.” EX1004, 18:8-12. DaCosta does not disclose the smart server storing data in the knowledge repository other than is this “background mode.” See generally EX1004. Unlike the navigation and extraction modules 10, 20—that require a user to interact with the page and select the text desired to be extracted (EX1004, 7:16-28, 9:57-59)—the smart server operates in the background without the active interaction of a user. After data is stored in the

knowledge repository during this background operation, a user can then access the data stored within the knowledge repository. EX1004, 18:15-23 (the user can readily access the stored data). Thus, a POSITA would understand the recording and playback of user generated navigation and extraction scripts is not directly related to the operation of the smart server.

59. A POSITA would understand this smart server is distinct from the user-oriented processes described in DaCosta's FIG. 1-19. This is explicitly stated in DaCosta as the disclosure transitions from the various embodiments of data extractors described in FIGs. 1-19 to the additional tools that may access existing enterprise processes. EX1004, 15:55-66 ("The description [of FIGs. 1-19] to this point refers to various embodiments developed to aid in data extraction from document sources, typically (though not exclusively) accessed on the Internet through a browser."). These additional tools can include smart clients (e.g., Web Bands) and smart servers.

60. Referring to the smart server, depicted and described with respect to FIG. 22, this server does not employ the same processes disclosed in these earlier figures. *Compare* FIG. 21 (showing smart client/Web Bands implementing navigation and extraction tools for the user) *with* FIG. 22 (depicting a server architecture without the navigation and extraction tools). For example, FIGs. 1-21 describe "a preferred embodiment of the present invention," while FIG. 22

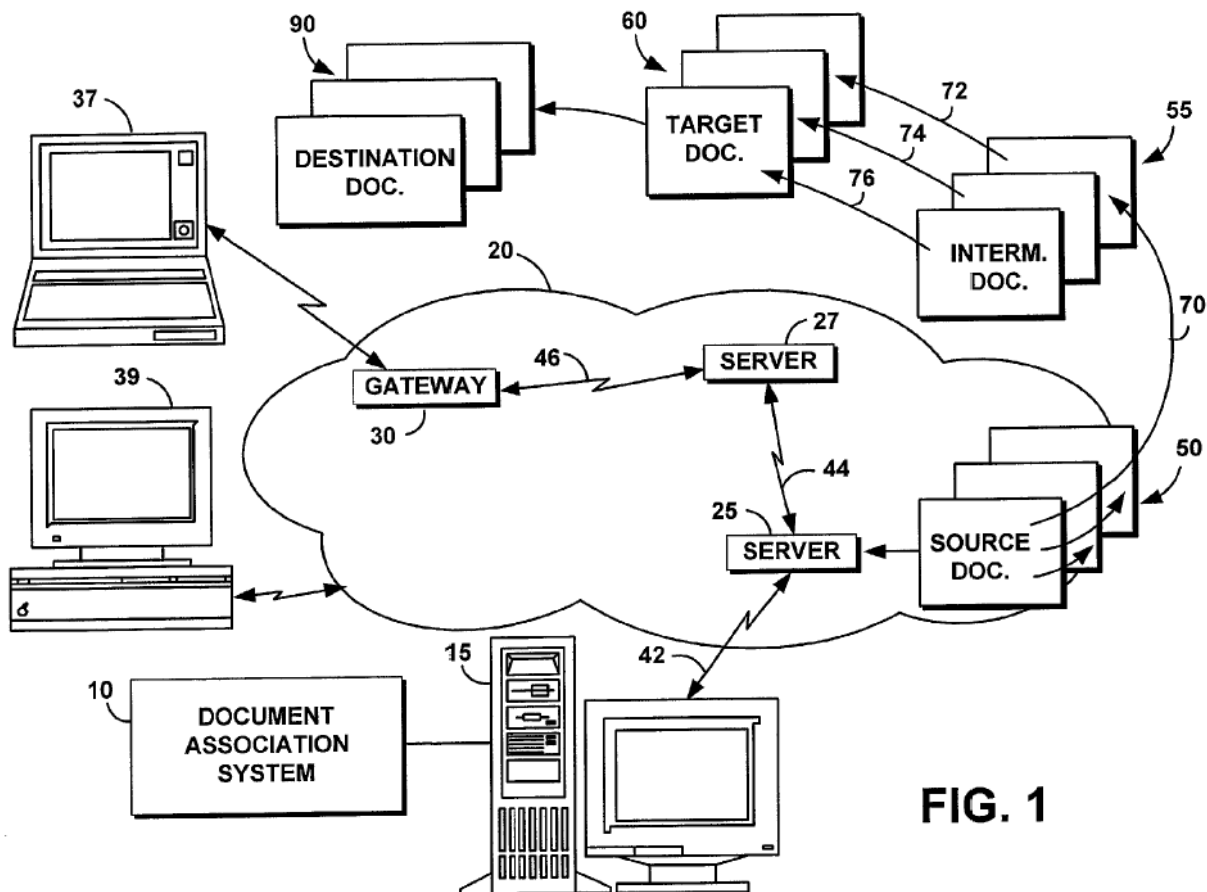
“illustrates an architecture for a smart web server to scour data from the Internet for deposit in a data repository.” *See* EX1004, 4:10-5:3. DaCosta intentionally changes the language in the description of FIG. 22 so to not take antecedent basis from the “preferred embodiment” described in the previous figures. EX1004, 4:10-5:3. A POSITA would also understand that the smart server is functioning as a typical server (EX1004, 18:21-22 (user can readily access the stored data)) that is different from the client-facing interface described in FIGs. 1-15 and 21. Thus, a POSITA would interpret this abrupt and apparent change in terminology to signify that the smart server—acting as a server that is accessed by the client—is distinct from recording of the scripts shown in FIGs. 2 and 4 and the system shown in FIG. 7.

B. Kraft (EX1005)

61. U.S. Pat. Pub. No. 2003/0018669 (“Kraft”) is an application that was filed by the same Applicant as the ’667 Patent. EX1001, 1/13 (Assigned to “International Business Machines Corporation”); EX1005, 1/13 (same). Accordingly, Kraft and the ’667 Patent share some of the same background disclosure, such as the definitions found in paragraphs [0025]-[0039] of Kraft. *See e.g.*, EX1001, 3:4-53.

62. Kraft discloses “a system for saving the content of a target document bundled with contextual metadata, such as the location of a source document, as attributes of the target document.” EX1005, ¶[0001]. Referring to FIG. 1 (inset

below), depicts the system 10 in an exemplary computing environment in which servers 25, 27 provide access to stored information such as hypertext or web documents indicated generally at 50, 55, and 60.



EX1005, ¶¶[0040]-[0042].

63. A user can then navigate between websites or documents until the user “detects the desired target document 60.” *Id.*, ¶¶[0057]-[0059]. After navigating to the target document, the user “issues a save command, and enters the destination address (URL) 176 of the destination documents repository 170 where he or she desires to store the destination document 90.” *Id.* The destination document

corresponds to the entire target document 60 along with additional “context data,” which includes the URL 630 of the source document 50, the URL 640 of the target document, the author, and a revision number. *Id.*, ¶[0060]. Thus, Kraft’s system can save an entire webpage (e.g., target document 60) along with information as to how a user navigated to that webpage (e.g., context data) in a document repository.

VII. GROUND 1: THE PETITION FAILS TO SHOW ANTICIPATION OF CLAIMS 1, 16, AND 20 BY DACOSTA

64. Dr. Williams’ analysis of claims 1, 16, and 20 is identical except for the identification of the first and second set of instructions for claim 16. EX1003, ¶¶135-144 (incorporating discussion from claim 1 for claim 16) and ¶¶148-157 (incorporating discussion from claim 1 for claim 20) . Therefore, while I only explicitly address the deficiencies of Dr. Williams’ analysis of claim 1, these deficiencies must also apply to claims 16 and 20.

A. Claims 1, 16, and 20 are Novel Over DaCosta

65. For the reasons explained below, Petitioner fails to establish DaCosta anticipates claims 1, 16, and 20.

1. [1g], [16g], [20g]: “selecting the target database and target locations in the target database where the at least one of the data elements is to be stored in the target database”

66. The claims require “selecting the target database and target locations where the at least one of the data elements are to be stored” (Limitation [1g]) and

the at least one of the data elements refers to the data elements selected from the list of data elements parsed from the source file (Limitation [1d]-[1e]). *Supra*, §V.D

67. Dr. Williams identifies a user selecting text from a web page via the extraction script (e.g., 22) as corresponding to the claimed at least one of the data elements for Limitation [1e]. EX1003, ¶73 (identifying “the user would then make a selection from the web page” at step 170 of FIG. 4) and ¶74 (“the user then highlights (i.e., using a mouse, drags over) the desired information”).

68. For the claimed target database, Dr. Williams identifies the “smart web server” depositing data into a knowledge repository and contends the smart server has the ability to use a distributed repository or a centralized depository. EX1003, ¶78-79. Petitioner then contends that DaCosta discloses selecting a particular target database and location “through a navigation script.” EX1003, ¶81 (citing EX.1004, 5:12-20 and FIG. 2). Dr. Williams’ theory is inconsistent with both its prior mapping and the teachings of DaCosta.

a) Petitioner’s treatment of DaCosta’s repositories is misguided

69. As explained in Section VI.A, DaCosta’s smart server operates in the background mode to fill a knowledge repository with data. EX1004, 18:8-12. This data stored within the knowledge repository is different from the data selected from the web page via the extraction module 20. *see supra*, §VI.A.3.

70. When filling the knowledge repository with its smart server, DaCosta states:

Smart servers provide distributed, multiple dynamic virtual data feeds. The data feeds Web Folders that comprise a knowledge repository accessible by smart clients. *In an alternative embodiment* of the present invention, *extracted data can be stored* in a conventional repository, including Microsoft Exchange 2000 or any other database that constitutes a knowledge repository.

EX1004, 18:14-23. That is, DaCosta discloses two “alternative embodiment[s]” for this knowledge repository: (1) the *alleged* distributed repository, where the smart server “provides distributed, multiple dynamic virtual data feeds” to “Web Folders” and (2) the *alleged* conventional repository, where “data can be stored in a conventional repository.” EX1004, 18:14-23. In addition to the knowledge repository, DaCosta discloses another embodiment with a distinct, centralized repository that stores *scripts*. EX1004, 18:34-39. A POSITA would understand that DaCosta discloses two alternative server architectures for storing extracted data in its knowledge repository and that its scripts may be stored in a centralized repository.

71. Contrary to this understanding, Dr. Williams’ mixes and matches the functionality of these two distinct repositories in an attempt to arrive at the claimed features. *See e.g.*, EX1003, ¶79 (“DaCosta further discloses the ability to have distributed repositories or a centralized depository”), ¶¶86-87 (“the target database

is a repository where that repository may be a conventional repository or a centralized repository”). However, this mixing and matching of alternative embodiments and other district repositories is not disclosed in DaCosta and would have been inconsistent with a POSITA’s understanding. As explained in more detail below, none of these repositories are selected to store the data elements identified by Dr. Williams.

b) DaCosta’s knowledge repository does not teach the claimed feature

72. DaCosta’s smart server is filled by the smart server during its operation in the *background mode*. EX1004, 18:8-12. DaCosta does not disclose the smart server storing data in the knowledge repository other than is this “background mode.” There is nothing in DaCosta that would have led a POSITA to understand that the user-selected text identified in paragraphs 73-74 of Dr. Williams’ declaration would have been stored in the smart server’s knowledge repository.

73. A POSITA would have understood this “background” operation to be operating without user interaction or intervention. EX2005, 5/6 (background processes “run behind the scenes (i.e., in the background) *without user intervention*”); EX2006 1/7 (“A background process, as the name suggests, operates behind the scenes. It is a computer process that runs independently of the user

interface, meaning it *does not require user intervention* to function”)¹ That is, the user does not select the specific text to be extracted during operation of the smart server in the background mode.

74. On the other hand, the scripts shown in FIGs. 2 and 4 (navigation and extractions scripts, respectively) require a user to “interact with the loaded page by clicking on or activating links or buttons, entering data and so on.” EX1004, 7:16-28; *see also id.* (“user-generated event 60”), EX.1004, 9:56-65 (“the user would then make a selection”), 11:12-17 (“the user then highlights (i.e., using a mouse, drags over) the desired information”). A POSITA would understand this selection of text as requiring user interaction or intervention. Such a process would have been understood to be a “foreground” operation, rather than the “background” operation performed by the smart server.

¹ The definition of a background process provided by the Linux (EX2005) and OneNinja (EX2006) references are consistent with a POSITA’s understanding at the time of the invention (July 18, 2002). As described in these references, the understanding of a background process did not meaningfully change between 2004 (Linux) to 2023 (OneNinja) and, similarly, that same understanding did not change as of 2002.

75. DaCosta does not disclose storing the user-selected text in the knowledge repository. *see supra*, §§VI.A.2-3. Instead, the user-selected text is output to the browser of the user. EX1004, 5:12-20 (The user “view[s] the final output using a browser such as Microsoft Internet Explorer”); *see id.*, FIG. 7 and FIG. 10 (describing to “view output on your browser”). Thus, a POSITA would have understood that the *alleged* data elements are output to a browser, not sent to the knowledge repository of the smart server. Further, although not alleged by Dr. Williams, a POSITA would not have understood the browser (e.g., internet explorer) as being the target database. *See* EX1001, 3:8-9, 37-40 (defining database and web browser as separate components). For the reasons explained above, Dr. Williams’ theory that the user-selected data is stored in the knowledge repository is unsupported and inconsistent with DaCosta’s teachings.

76. Further, to the extent Petitioner is mapping DaCosta’s disclose of two alternative embodiments for the knowledge repository as selecting the repository, this is incorrect. The disclosure of two separate, “alternative” embodiments for the architecture of the smart server does not disclose a selection between two repositories. A POSITA would have understood that the “alternative embodiment” of a conventional repository is discussing how the server architecture would have been implemented within DaCosta’s system. EX1004, 18:16-23. That is, the smart server could have been set up to store data in the “distributed repository,” as is

depicted in FIG. 22. *See* EX1004, FIG. 22, 5:1-3 (FIG. 22 illustrates “an architecture” for a smart server). Alternatively, in a distinct embodiment, the smart server could have been set up to store data in the conventional repository. EX1004, 18:16-23. DaCosta does not disclose that the smart server would select between a distributed repository and a conventional repository when storing data. Thus, to the extent Dr. Williams intends to rely on the “alternative embodiments” of the smart server as selecting between repositories, this is inconsistent with DaCosta’s teachings.

c) DaCosta’s centralized repository or navigation scripts do not teach the claimed feature

77. For Limitation [1e], Dr. Williams contends that step 170 of DaCosta’s FIG. 4 discloses selecting the data elements from the list. EX1003, ¶73. Here, Petitioner now identifies a centralized repository that is used to store scripts. Pet., 20; EX1004, 18:33-36 (“scripts are stored at a centralized repository”). The scripts are distinct from the data extracted from the webpage. Thus, DaCosta fails to disclose storing the user-selected text—the alleged data elements—in the centralized repository—the alleged target database.

78. Dr. Williams’ assertion that “the navigation script also controls the storage of the extracted data in a repository, e.g., target database, as shown in element 90 of Figure 2” is also unsupported by DaCosta. EX1003, ¶82. As illustrated in DaCosta’s FIG. 2 (produced in Section VI.A.1), “the navigation

recording module 12 memorizes which anchors were clicked and which forms were submitted 70 and maps 80 each user event to an element in the recorded HTML object or element hierarchy” and “[t]his information is preferably stored 90 in an [XML] file.” EX1004, 7:29-34. Storing the navigation recording script in an XML file at step 90 of FIG. 2 does not disclose that this navigation script controls the storage of the extracted data in a repository. Further, these recorded “anchors” that were clicked and “forms” that were submitted are distinct from the text selected in step 170 of the extraction script depicted in FIG. 4. Dr. Williams haphazard mixing of various steps in the different navigation and extraction scripts is inconsistent with DaCosta’s teachings and a POSITA would not understand the cited portions of DaCosta as teaching the claimed features. In fact, a POSITA would not understand the navigation script to have access to the extracted data and, therefore, the navigation script could not control the storage of this data. EX1004, 5:61-63 (“instead of recording/playing across web pages and web space, the extraction module 20 records and plays across elements within a single page.”), 8:61-65 (“[W]hile the navigation module [10] operates across multiple web pages or web space, the extraction module 20 operates at the HTML or web page or other web-accessible document level.”); *see also* EX1004, FIG. 7 (depicting the navigation module 10 as separated from the data handled by extraction module 20).

79. Further, to the extent Dr. Williams intends to assert that DaCosta's scripts select a particular target database, this is unsupported and inconsistent with the teachings of DaCosta. Dr. Williams cites solely to the generic summary of DaCosta (EX1004, 5:12-20) to support the assertion that "DaCosta discloses selecting a particular target database and location through the use of scripts." EX1003, ¶81. However, this cited portion of DaCosta recites "process[ing] the extracted data according to customizable scripts ... [and] view the final output using a browser such as Microsoft Internet Explorer," which is **not** selecting a target database. A POSITA would have understood that sending information to a browser (or viewing the output via a browser) does not teach nor disclose selecting a database.

2. *[1i], [16i], [20i] – "setting a schedule frequency for transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database"*

80. Dr. Williams cites to the generic summary of DaCosta's system (EX1004, 5:12-20), the practice of utilizing form variations during with the playback module 14 of the navigation module 20 (EX1004, 7:62-65), and the smart server (EX1004, 18:23-25). EX1003, ¶¶88-89. Dr. Williams does not connect these three different aspects of DaCosta, and a POSITA would not understand that Dr. Williams picking and choosing of various functionalities of different embodiments within DaCosta as disclosing the claimed feature.

a) DaCosta does not set a schedule frequency for transferring data elements to the target database

81. Dr. Williams first cites to the generic summary (EX1004, 5:12-20) to support the assertion that “DaCosta discloses repeating its method of extracting, processing, integrating.” Importantly, neither “extracting,” “processing,” nor “integrating” disclose transferring the parsed data element to the target database. Further, none of the steps disclosed in DaCosta’s generic summary disclose transferring data to a database. EX1004, 5:12-20. In fact, DaCosta discloses transferring the output to “a browser such as Microsoft Internet Explorer,” instead of transferring to a database. EX1004, 5:12-20 (disclosing to “view the final output using a browser”). Thus, it is irrelevant that any of these steps are repeated as they do not disclose transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database.

82. Next, Dr. Williams cites to the form parameters as allegedly controlling the setting of a schedule frequency, but this assertion is also inconsistent with DaCosta’s teachings. See EX1003, ¶89. As explained in Paragraph 54, the form parameters relate to the form fields in the web page and form variations can be used to change the data entered into the form fields. For example, the navigation script records “which forms were submitted” to navigate to a desired page (EX1004, 7:29-31) and the playback module 14 can change this script by “specifying new form parameters to replace the ones originally recorded.” EX1004, 7:55-64.

83. This specifying of new form parameters enables the system to repeatedly query a web site while introducing variations. EX1004, 7:64-65. DaCosta provides an example of this form variation where “a form can be filled in to get pricing information from an online bookstore for different titles by changing a single parameter in a query form” (e.g., the ‘title’ form field). EX1004, 7:65-8:3. That is, the playback module 14 can modify the original script—which included a single query that only retrieved information for a single book title—to access information on multiple other book titles by “varying” the input to the “book title” form field. The user may then enter multiple form variations (e.g., titles of books) and the script will run multiple queries based on the original exemplar query, instead of the user having to create new scripts to run separate queries for every variation. EX1004, 8:3-5. This varying of queries submitting into a form field has nothing to do with setting a schedule frequency to transfer data to a target database.

84. Finally, Dr. Williams’ example of a smart server looping “after a *preset* delay” does not disclose setting a schedule frequency. EX1004, 18:23-25. As explained in Sections VI.A.3 and VII.A.1.a), the smart server is a distinct embodiment that stores data in a background mode and is unrelated to the user-oriented processes that Dr. Williams maps to as corresponding to the parsed data elements. See EX1003, ¶73. Thus, the operation of the smart server cannot disclose the claimed feature. Further, a POSITA would not understand a “*preset* delay” as

disclosing setting a schedule frequency. A POSITA would have understood a preset delay to mean that this delay was set in advance. EX2007, 1 (preset means “set in advance”).² The fact that a delay was previously set does not disclose setting the delay or a schedule frequency.

3. *[1h], [16h], [20h] “saving the source locations and target locations of the at least one of the parsed data elements in a location file in the target database;”*

85. Dr. Williams identifies the URL or metabookmarks as corresponding to the source location (EX1003, ¶76) and identifies two distinct repositories—the knowledge repository and centralized repository—as corresponding to the target location and target database. EX1003, ¶¶78-79. Dr. Williams’ analysis of Limitation [1h] introduces further inconsistencies that diverge from its previous mappings and are inconsistent with the teachings of DaCosta.

a) **The Navex scripts do not disclose a target location in a target database**

86. Dr. Williams contends that DaCosta’s navigation and extraction scripts (Navex scripts) “include source and destination information.” EX1003, ¶83. Dr. Williams cites only to FIG. 10 of DaCosta to support this assertion (EX1003, ¶¶83-85), but this reliance is incorrect.

² This definition is consistent with a POSITA’s understanding of the term “preset” at the time of the invention.

87. Referring to DaCosta's FIG. 10 (inset below), the Navex scripts can "extract numerous records 260" based on a pattern that was derived by the extraction recording module 22. EX.1004, 11:19-24.

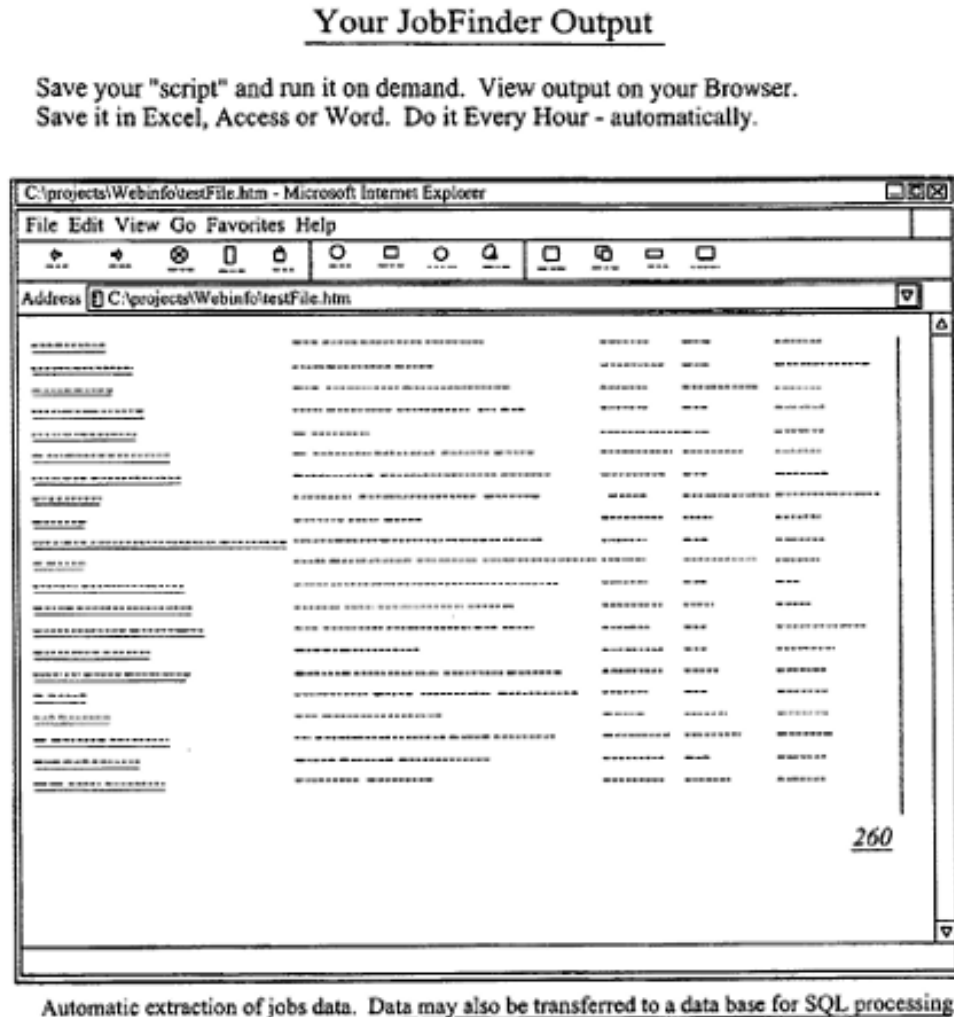


FIG. 10

A POSITA would not understand the identification of this pattern (or the extraction of data based on the pattern) to disclose "destination information" or a target location in the target database. The pattern is identified by the extraction playback module

24 and is used to determine which information to extract from a webpage. EX1004, 9:7-14 (“a pattern according to which data can be automatically extracted from a web page”). There is nothing found within DaCosta’s description of FIG. 10 that would have led a POSITA to understand that the script necessarily determines where to send or store that information. And, notably, the script does not determine to save or store the records in a target database. To illustrate, the data 260 from FIG. 10 that matches the pattern is “output on your browser.” EX1004, FIG. 10. Thus, the Navex script that Dr. Williams identifies sends the data to a browser, not the target databases identified by Petitioner. *See* Pet., 20.

88. Dr. Williams also quotes the statement that “data may also be transferred to a data base for SQL processing” along with all the other text found in FIG. 10. However, as explained in Sections VII.A.1.b)-c) (above) and Section VII.A.3.b) (below), DaCosta does not disclose that this data would be transferred to either the knowledge repository or centralized repository identified by Petitioner. And to the extent Petitioner intends to assert that the browser or the database for SQL processing is the same as the knowledge repository, this is neither taught nor disclosed in DaCosta. Accordingly, a POSITA would not understand the identified portions of DaCosta as disclosing the claimed feature.

b) Dr. Williams' mapping of the Databases is inconsistent with DaCosta

89. Dr. Williams' analysis of the conventional, centralized, and decentralized, repositories is incompatible with DaCosta's teachings. EX1003, ¶¶86-87 (combining features from the distinct repositories); *see supra*, §§VI.A, VII.A.1.a). As explained above, DaCosta discloses a knowledge repository that is filled by "a smart server" operating in "background mode." EX1004, 18:10-12. This knowledge repository can be set up in the form of: (1) multiple dynamic virtual data feeds—the distributed repository—or, alternatively, (2) a conventional repository. EX1004, 18:14-23. In addition to the knowledge repository—which stores extracted data—a preferred embodiment of DaCosta can store scripts in a centralized repository. EX1004, 18:33-36.

90. Dr. Williams' assertion that the target database "may be a conventional repository or a centralized repository" improperly misconstrues the functions of the two separate databases to create something that is not disclosed in DaCosta. The conventional repository is an alternative embodiment for storing ***extracted data*** from the smart server and the centralized repository is used for storing scripts. EX1004, 18:8-23, 34-40; *see supra*, §VII.A.1.a). Dr. Williams' assertion that "scripts may also be stored in a decentralized target database" (EX1003, ¶87) is similarly misplaced. Again, this is misrepresenting the disclosure of DaCosta and how a POSITA would have understood DaCosta's system. DaCosta discloses a first

database for storing scripts (the centralized repository) and a second, distinct database for storing data extracted by the smart servers (the knowledge repository). A POSITA would not understand that scripts (or other information) would be stored in the knowledge repository, or that extracted data would be stored in the centralized repository.

91. In fact, Dr. Williams’ contention is nonsensical when viewed next to the quoted portion of DaCosta (EX1004, 18:17-23), which states that “*extracted data* can be stored in a *conventional repository*”—**not** that *scripts* can be stored in a *decentralized* database. This is similarly nonsensical from the perspective of a POSITA. DaCosta states that multiple users can utilize the same script and this script would be stored in a centralized server. EX1004, 18:34-41. “In this way, if there are multiple users of a script, should that script fail it is easy to ensure each of the users has a corrected script as soon as possible, i.e., as soon as it is downloaded to the central repository.” *Id.* This enables the script to be corrected at a single source, rather than having to correct the script at multiple different sources for each user. On the other hand, the data extracted by a user is specific to that user’s needs and would not be stored in a central repository for everyone to access. Thus, Petitioner’s contention that scripts may be stored in a decentralized target database is contrary to the explicit teachings of DaCosta.

4. [16b], [16j]: First and second sets of instruction codes

92. Dr. Williams contends that DaCosta’s navigation script and extractions script correspond to the claimed first and second sets of instruction codes. EX1003, ¶136 (“a navigation script e.g., a first set of instructions, and an extraction script, e.g., a second set of instructions”). Dr. Williams then incorporates his analysis of claim 1 for the remaining limitations of claim 16 ([16c]-[16j]). EX1003, ¶¶136-144. But this incorporation creates internal inconsistencies that are fatal to Ground 1.

93. Although Dr. Williams identifies DaCosta’s navigation script as the first set of instructions, Dr. Williams relies on the extraction script in several parts of his analysis of the first set of instructions. In fact, the Dr. Williams cites to the extraction script (or extraction module 20) as disclosing portions of Limitations [1c], [1d], [1e], [1g], and [1h]—relying exclusively on the extraction module for Limitation [1e]. EX1003, ¶¶69, 71, 73-75, 81, 84. Dr. Williams' reliance of the navigation script as teaching the first set of instructions while incorporating his discussion of the extraction script—the alleged second set of instructions—renders his theory internally inconsistent.

VIII. GROUND 2: THE PETITION FAILS TO SHOW ANTICIPATION OF CLAIMS 1, 16, AND 20 BY KRAFT

A. Claims 1, 16, and 20 are Novel Over Kraft

94. Dr. Williams’ analysis of claims 16 and 20 is the exact same as that for claim 1. *See* EX1003, ¶¶223-231, 236-244. Therefore, while I only explicitly

address the deficiencies of Dr. Williams’ analysis of claim 1, these deficiencies must also apply to claims 16 and 20.

1. [16b], [16j]: *First and second sets of instruction codes*

95. Unlike Ground 1, Dr. Williams fails to address the claimed first and second sets of instructions that are recited in claim 16. *See e.g.*, EX1003, ¶223. Dr. Williams provides no analysis for the claimed first set of instruction codes or second set of instruction codes and merely states: “This claim limitation closely matches a limitation present in ground 2, claim 1. In my opinion, for the same reasons discussed above regarding claim 1, Kraft discloses this claim limitation.” EX1003, ¶223. However, Dr. Williams fails to identify and “set of instruction code” in its analysis of claim 1. EX1003, ¶¶160-161. As Dr. Williams fails to identify anything in Kraft as allegedly disclosing the claimed first and second sets of instruction codes, this analysis is deficient and cannot show that Kraft discloses each and every element of claim 16.

2. [1i], [16i], [20i]: – “*setting a schedule frequency for transferring the at least one of the parsed data elements from the source locations to the target locations in the location file in the target database;*”

96. Dr. Williams provides three theories for this claimed feature—all of which are deficient. EX1003, ¶¶178-179 (identifying automatic synchronization application), ¶182 (identifying periodic operation of system in the background and web crawler of prior art system)

97. First, Dr. Williams alleges that “Kraft discloses setting a schedule frequency where ‘users will be able to return to the source document, and optionally to use applications for automatically synchronizing a destination document to the target document.’” EX1003, ¶178 (quoting EX1005, ¶[0009]). A user returning to a source document (e.g., web page) fails to disclose setting a schedule frequency. The ’667 Patent describes an example of setting a schedule frequency where (“the user sets the timer 350 by selecting the frequency of data retrieval; i.e., every 5 minutes, 15 minutes, at a specific time such as 12:00 p.m., etc.” EX1001, 5:17-19. Thus, a POSITA, viewing the specification, would not understand the user returning to a source document as setting a schedule frequency.

98. Additionally, the use of applications for automatically synchronizing a destination document to the target document fails to disclose setting a schedule frequency. The synchronization application “performs a comparison of the destination document with the target document to detect changes and to automatically update the destination document.” EX1005, ¶[0016]. Kraft does not disclose setting a schedule frequency to perform this update/comparison. Indeed, a POSITA would understand that the synchronization application is activated based on “the user [requesting] to update the destination document.” EX1005, ¶[0016]. To the extent that Dr. Williams is relying on the term “automatically,” this term is simply referring to the user not having to do anything *after* requesting for the

destination document to be updated. A POISTA would not understand after the synchronization application performs a comparison of the two documents (initiated by a user request), the application will then automatically update the destination document if a difference is found. EX1005, ¶[0016]. Regardless, Kraft unequivocally does not disclose that this synchronization application sets a schedule frequency and, therefore, the synchronization application cannot anticipate the claimed feature.

99. Next, Dr. Williams states that Kraft’s “system 10, transparently to the user, continuously or periodically operates in the background.” EX1003, ¶180 (quoting EX1005, ¶[0044]). This quoted portion fails to disclose setting a schedule frequency. The disclosure that a computing system may operate in the background does not teach the claimed features.

100. Finally, Dr. Williams cites to “the background section” of Kraft that states “[t]he crawler is run periodically to update previously stored data and to append information about newly created web pages” and that “[t]he information compiled by the crawler is stored in a Metadata repository or database.” EX1003, ¶180. This does not disclose setting a schedule frequency.

Additionally, this citation to Kraft’s background section creates inconsistencies in Dr. Williams’ positions that render Ground 2 deficient. None of Dr. Williams’ previous mappings relate to the crawler or Metadata repository.

EX1003, ¶¶172-173 (identifying destination documents repository as the target database). Petitioner first maps to destination documents repository 170 of Kraft's system as the claimed target database and now relies on the Metadata repository of a prior art system. Yet, aPOSITA would have understood the destination documents repository 170 as being wholly unrelated to the Metadata repository. To the extent this Metadata repository would have been implemented in Kraft's system aPOSITA would understand this repository to correspond to the search engine repository 210. *Compare* EX1003, ¶[0003] ("information compiled by the crawler is stored in a Metadata repository") *with* EX1005, ¶[0053] ("crawler 150 ... downloads web documents to the search engine repository 210"). Thus, Dr. Williams' theory is inconsistent with the teachings of Kraft and Kraft does not disclose the claimed feature.

3. [1e], [16e], [20e]: – “selecting at least one of the data elements from the list to be stored in the target database”

101. Dr. Williams cites solely to paragraph [0059] of Kraft as allegedly disclosing this feature. EX1003, ¶167. Dr. Williams provides no analysis and only states that “Kraft discloses” while quoting specific portions of paragraph [0059]. *Id.* Despite this concise analysis, aPOSITA would not recognize the quoted portions of Kraft as teaching the claimed limitation.

102. Kraft discloses navigating to a target document 60 (e.g., a web page) and then saving the entire document in the destination documents repository 170.

EX1005, ¶[0059]. There is no teaching of selecting data elements and, therefore, the cited portion of Kraft cannot disclose “selecting at least one of the data elements from the list to be stored in the target database.”

103. To the extent that Petitioner intends to argue that the navigation to the target document discloses selecting at least one data element, this is unsupported and inconsistent with the understanding of a POSITA. The '667 Patent clearly distinguishes between selecting a web page (e.g., source file) and selecting of data elements. EX1001, 2:20-22 (System “allows the user to [(1)] select one or more Web pages containing data of interest; [and (2)] to specify exactly which data within any page is to be captured”); EX1001, 4:44-55 (describing selecting a web page at step 410), 5:12-13 (describing selecting data in step 430). Thus, the quoted portion of Kraft cannot disclose the claimed feature. For these reasons, Petitioners Ground 2 must fail.

4. [1d], [16d], [20d]: “parsing selected data elements in the source file into a list of data elements;”

104. Dr. Williams cites to two paragraphs of Kraft as allegedly disclosing the claimed feature. Neither paragraph teaches the limitation. In fact, the two paragraphs of Kraft cited by Dr. Williams are also found in the '667 Patent and are described as being distinct from the claimed feature. *Compare* EX1005, ¶¶[0028], [0058]) *with* EX1001, 3:13-22, 4:6-16.

105. Dr. Williams' first theory cites to the definition of “HTML (Hypertext Markup Language)” that is found in the “**background information** pertaining to the technical field of Kraft’s invention.” EX1005, ¶¶[0025], [0028]. A POSITA would understand this definition as disclosing that HTML is used for including attributes within web pages by the owner of the web page. EX1005, ¶[0028]. And that these attributes will be interpreted by a user’s web browser to display the document to the user. *Id.*

106. A POSITA would not understand this background definition as disclosing the claimed feature. This exact same definition is provided in the ’667 Patent. EX1001, 3:13-22 (providing same definition of HTML). A POSITA—viewing the specification—would understand this background definition of HTML to be distinct from the claimed feature. Dr. Williams provides no analysis of the claim and appears to rely on this definition because it is the only paragraph in Kraft that includes the term “parse.” *See generally*, EX1005. However, the fact that web pages are “interpreted by the browser and used to parse and display the document” (EX1005, ¶[0028]) does not disclose the claimed step of “parsing selected data elements in the source file into a list of data elements.” The ’667 Patent discloses accessing web documents (e.g., webpages) 50, 55, 60 that would require the user’s browser to render and display the webpages. EX1001, 4:6-26. The navigating and

displaying of webpages by a web browser is distinct from the parsing step of FIG. 4 that corresponds to the claimed feature. *See* EX1001, 4:46-59.

107. Next, Dr. Williams contends that Kraft “discloses that the list of data elements may include such hyperlinks.” EX1003, ¶166 (citing EX1005, ¶[0058]). Dr. Williams provides no explanation as to how these hyperlinks relate to the claim language and a POSITA would not the use of hyperlinks as disclosing this claimed feature. EX1001, 4:6-26, 4:46-59 (disclosing hypertext links 70-76 as distinct from the claimed feature).

108. To illustrate, steps 320-330 of FIG. 3 described navigating between web pages using hyperlinks, such as going from source document 50 to documents 55, as illustrated in Kraft’s FIG. 1 (inset below).

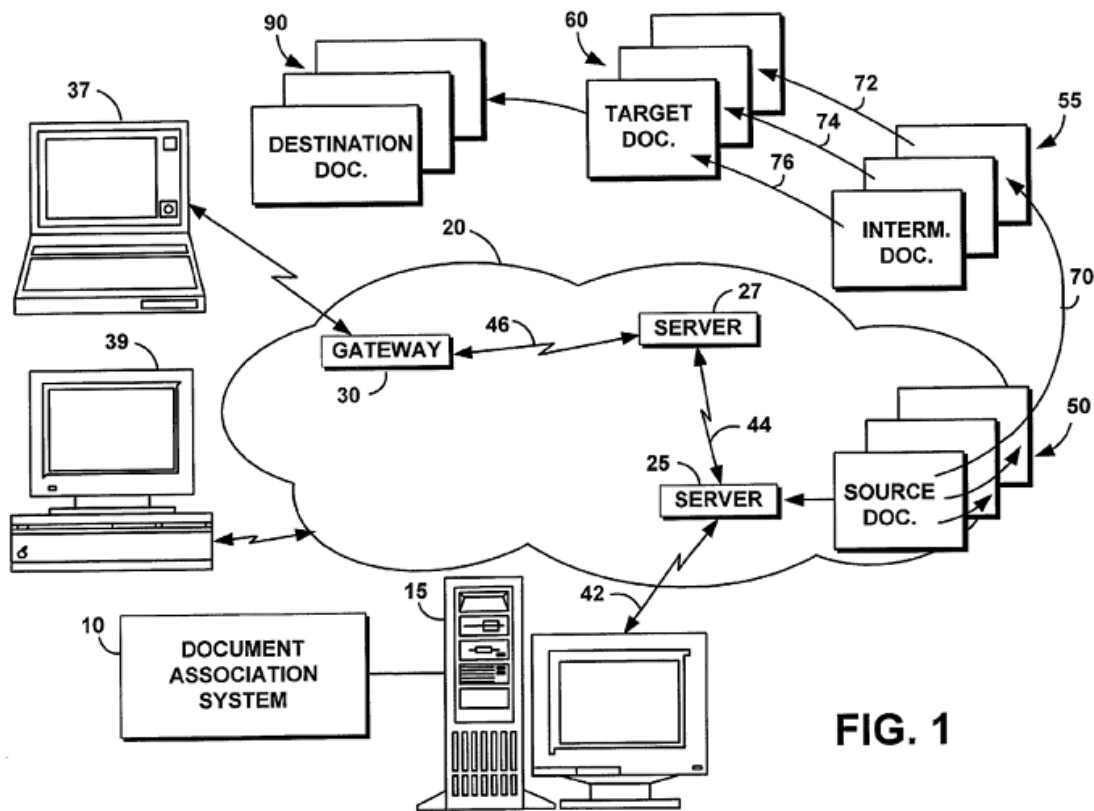


FIG. 1

EX1005, ¶[0058]. This is the same type of navigation disclosed in the '667 Patent, where “servers 25, 27 provide access to stored information such as hypertext or Web documents indicated generally at 50, 55, and 60 ... [that include] hypertext links 70, 72, 74, 76 to other webs sites or documents 55, 60.” EX1001, 4:9-15. Yet, the '667 Patent distinguishes between navigating to a HTML page (e.g., page 220 of FIG. 2) and parsing selected data elements in the source file into a list of data elements. EX1001, 4:44-66 (selecting web page 220 in step 410 and parsing the source file in step 420). In fact, a POSITA reading the '667 Patent would understand that the navigation to a desired HTML page (e.g., 220) would correspond to selecting the source file. EX1004, 44-59 (specifying the source file in step 410 “as shown by FIG.

2” by navigating to the webpage and then later downloading and parsing the source file in steps 415 and 420). This is also confirmed by the Dr. Williams, which cites to this same disclosure (EX1005, ¶[0058]) as disclosing Limitation [1c]. *See* EX1003, ¶163.

109. For the reasons above, a POSITA would not understand either of the quoted portions of Kraft as disclosing the claimed feature.

IX. CONCLUSION

110. All statements made herein of my own knowledge are true, all statements made herein on information and belief are believed to be true, and these statements were made with the knowledge that willful false statement and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001.

Respectfully submitted,

Dated: January 17, 2024



Kevin C. Almeroth, Ph.D.