

Session Identification URI

W3C Working Draft *WD-session-id-960221*

This version:

<http://www.w3.org/pub/WWW/TR/WD-session-id-960221.html>

Latest version:

<http://www.w3.org/pub/WWW/TR/WD-session-id.html>

Authors:

Phillip M. Hallam-Baker <hallam@w3.org>

Dan Connolly <connolly@w3.org>

Status of this document

This is a W3C Working Draft for review by W3C members and other interested parties. It is a draft document and may be updated, replaced or obsoleted by other documents at any time. It is inappropriate to use W3C Working Drafts as reference material or to cite them as other than "work in progress". A list of current W3C working drafts can be found at: <http://www.w3.org/pub/WWW/TR>

Note: since working drafts are subject to frequent change, you are advised to reference the above URL, rather than the URLs for working drafts themselves.

Abstract

A Uniform Resource Identifier for identifying HTTP sessions is described. Session identification URIs permit HTTP transactions to be linked within a limited domain. This provides a balance between the needs of commercial servers for demographic data collection and the privacy concerns of users. In addition session identification URIs may be used as part of a high security authentication mechanism to prevent replay attacks.

Introduction

HTTP is specified as a stateless protocol. This permits HTTP servers to handle a large number of simultaneous requests. The stateless nature of HTTP reduces its utility however. It is not possible to track user reading patterns on a single server nor is it possible for a server to adapt its behavior on the basis of previous interactions.

The ability to trace the path of readers within a Web is important for maintainers of larger sites. Trace information may be used to analyze the efficacy of cross references within the site, and to build profiles of typical users. If it is known for example, that readers of an online newspaper who visit the computer section are very likely to also visit the business section reporters might be asked to provide more cross linkages between these sections. Administrators may also wish to discover the number of users visiting their site rather than the number of visits.

Advertising as Revenue for Web Content Providers.

Many content providers raise revenue through advertising. Advertisers therefore need to know the effectiveness of Web based advertising. Content providers who can provide advertisers with detailed profiles of the readership

of their material will be able to charge higher rates. Reader profiling would permit those advertisements most likely to obtain a response to be chosen.

A distinctive feature of the Web is its interactive nature. Gill [Gill96] points out that the interactive nature of the Web may make traditional models of "targeted" advertising obsolete, replacing them with participatory models. The Web is an information system and users who wish to purchase goods are likely to use it to find out details. It may be unnecessary to target advertising in an intrusive manner (e.g. unsolicited email). As users become accustomed to more participatory modes of advertising intrusive methods may become counter productive.

There are many metrics which an advertiser may wish to use to assess the value of a Web placement. These include:

Hit counts

The number of times an advertisement is downloaded. These roughly correspond to exposures as understood in conventional media.

Referrals

The number of times an advertising hyperlink is followed. This implies that the advertiser also has a Web site.

Hot leads and Sales.

Referrals which result in readers demonstrating a significant level of interest or which generate sales.

Referrals may be determined using the [HTTP referer field](#) which informs a server of the URI of the resource which referred the client to a resource. Unfortunately current log file formats do not include this information. A [companion document](#) describes an extension to the logfile format to record this data.

The number of hot leads and/or sales generated by a placement may be determined by correlating trace data within the advertiser's home Web site with the referer field. This procedure creates an interesting correspondence of interest between the parties which removes the need for conventional auditing. An advertiser might pay the publisher according to the business generated by a placement. It is in the advertiser's interest to be honest in determining the amount paid since the publisher would determine placement frequency according to the rate of return. This mechanism is of particular interest for adverts targeted at a particular readership where auditing may be difficult.

Privacy Concerns

Just because an advertiser is interested in information does not mean that the user is willing to provide it. If care is not taken to protect the privacy of its users the Web could enable more extensive surveillance of its users than has been available to the most ruthless dictatorships.

The Internet has a strongly developed but highly unpredictable ethical sense. It is a medium of active participants, not of passive consumers. Users may complain very publically about perceived wrongs (whether justified or not) via Usenet which has a readership of several millions. Privacy issues in particular are a frequently issues. Consequently it is advisable to approach the issue of personal privacy cautiously.

Users may be prepared to exchange information about themselves in return for access to content. Such systems may provide inaccurate data however. Users who believe their privacy to be threatened may deliberately supply incorrect information, supplying a false address and telephone number to prevent unsolicited mail and phone calls.

Personal data is often collected by financial institutions to serve as a means of customer authentication. Disclosure of personal data may therefore increase fraud risks.

Many countries have enacted privacy legislation which controls storage and use of personal data. Sites which are governed by such laws may wish to avoid unnecessary acquisition and recording of personal data.

Although the Web has gained popularity as a publishing medium it was conceived as a collaboration tool. As Turkel points out [Tutel96], a part of the interest of cyberspace may be the ability to take on different personas, the ability to voice unpopular views without risk. Such partitioning of identity requires the ability to separate online activity from offline activity and online activity at one site with activity at another. The Web should therefore permit users to take on new cyberspace identities through use of pseudonyms and the boundaries between these identities must be carefully protected.

Privacy or lack thereof has often been an unanticipated consequence of a particular technology. Early telephone users had little privacy since every conversation could be overheard by the operator. This led directly to the automatic exchange which was invented by an undertaker whose rival was stealing his business by bribing telephone operators.

Transactions in the HTTP 1.0 protocol are disjoint. A single request is made which results in a single response after which the operation is completed and the TCP/IP connection closed. The HTTP/1.1 allows the same TCP/IP connection to be used to perform multiple operations.

Pseudo Session Identifiers.

IP addresses and ports may be used to provide pseudo identifiers for analysis of demographic data. The usefulness of such identifiers is severely limited. It is not possible to differentiate two users timesharing on a single machine. Nor do users necessarily use the same IP address each time. The value of IP addresses for analysis is rapidly decreasing due to the growing use of proxies and dynamic IP address assignment. These trends will be exacerbated by new developments such as mobile IP.

Although these pseudo session identifiers are unreliable and unsatisfactory they should be taken into consideration when considering the privacy issues raised by this proposal. In particular it is unnecessary to provide exhaustive proofs of that certain forms of linkage cannot be achieved where this is possible through similar analysis of IP addresses and ports.

Relationship to State-Info (Cookies)

[State Info](#) [Kristol95] is a proposed extension to the HTTP protocol. It is a refinement of the [Netscape "Cookies" proposal](#) [Netscape95]. This mechanism permits a server to generate a token which a client which is returned with future requests. This mechanism requires clients to store data for every server visited and is consequently unusable with a tracking mechanism unless the number of sites using it is small. In the Session Identifier URI proposal identifiers are generated by clients, not servers. This provides for scalability since a client need only store a fixed amount of identifier information regardless of the number of sites visited.

URI Format

Session IDs have the form:

`SID:type:realm:identifier[-thread][:count]`

Where the fields *type*, *realm*, *identifier*, *thread* and *count* are defined as follows:

type

Type of session identifier. This field allows other session identifier types to be defined. This draft specifies the identifier type "ANON".

realm

Specifies the realm within which linkage of the identifier is possible. Realms have the same format as DNS names.

identifier

Unstructured random integer specific to realm generated using a procedure with a negligible probability of collision. The identifier is encoded using base 64.

thread

Optional extension of identifier field used to differentiate concurrent uses of the same session identifier. The thread field is an integer encoded in hexadecimal.

count

Optional Hexadecimal encoded Integer containing a monotonically increasing counter value. A client should increment the count field after each operation.

Examples

The following example shows a sequence of session identifiers created by the same client. Note that the same counter register is used to generate all the session identifiers within the same thread.

```
SID:ANON:www.w3.org:j6oA0xCWZh/CD723LGeX1f-01:34
SID:ANON:mc.ai.mit.edu:NRviSpoYm7mdkYB4W24711-01:35
SID:ANON:www.w3.org:j6oA0xCWZh/CD723LGeX1f-01:36
SID:ANON:mc.ai.mit.edu:NRviSpoYm7mdkYB4W24711-01:37
SID:ANON:www.w3.org:j6oA0xCWZh/CD723LGeX1f-02:01
SID:ANON:www.w3.org:j6oA0xCWZh/CD723LGeX1f-01:38
```

Limited Linkage of Session Identifiers.

Session Identifier URIs permit linkage of transactions within a single *realm*. A realm may be considered to approximate to a DNS name. DNS names correlate reasonably well with administrative divisions. This allows a content provider to track activities within sites on their network but does not permit data from different sites to be correlated without specific user authorization in advance.

Prevention of Replay Attacks

Session Identifiers may also be used within a strong authentication scheme to prevent replay attacks. A replay attack involve the recording of authentic traffic then replaying it at a later date. For example Mallet might intercept Alice's request to download her mail file on Monday and then replays it each day to receive the mail for the rest of the week.

Replay attacks may be prevented by checking message timestamps. Unfortunately this requires accurate and secure synchronisation of clocks at both ends of the communication which is difficult. Alternatively a challenge/response sequence may be employed. This introduces an additional round trip delay into the transaction and requires the server to maintain a check of which challenges have already been responded to.

The session identifier URI may be used to prevent replay attacks in combination with a timestamp. The server maintains a record of each identifier used and checks that subsequent requests with that identifier have a higher count field. The volume of data storage required may be minimized by checking that the timestamp falls within an acceptable validity interval.

Implementation Issues

A standardized method of constructing session identifiers would permit users to use the same session identification information on different machines avoiding the need to re-register with content providers. This would also be convenient for content providers, avoiding a user with more than one machine being counted twice. The nature of the session identifiers prevents enforcement of such a policy however and the following construction method is therefore only advisory.

A convenient method of constructing session identifiers which does not require separate storage for each realm visited is to use a Message Authentication Code (MAC) based upon a cryptographically secure one way function such as MD5 [Rivest92].

On initialization the client obtains a value *key*. This value should be selected in a random manner so as to provide at least 128 bits of ergodicity. When a realm is visited the value of the identifier field is created using the formula

identifier = MD5 (*realm* + *key*).

The client should store the value of *key* and the counter value associated with each thread.

HTTP Integration

Session identifiers may be incorporated in HTTP messages using the Session-Id header. The existing WWW-Authenticate header is extended to permit use of session identifiers as a lightweight authentication mechanism.

Session-Id

Session-Id: *URI*

The Session-Id header may be incorporated in a http request or response. The header accepts a single parameter, the identifier URI.

Session identifiers are only created by clients. A Session-Id header should only be present in a response if one was specified in the corresponding request and should return the same session identifier value as the request.

Example

The following example shows a HTTP request incorporating a session identifier.

```
GET / HTTP/1.0
Accept: text/plain
Accept: text/html
Session-Id: SID:ANON:w3.org:j6oA0xCWZh/CD723LGeXlf-01:034
User-Agent: libwww/4.1
```

A client supporting session identifier URIs should by default attach a session identifier to every request using the DNS name of the server as the realm. Clients must provide users with an option to disable session identifier generation. Clients are encouraged to provide a means of selecting the *realm* -> *identifier* mapping.

WWW-Authenticate

WWW-Authenticate: 1#*challenge*

The WWW-Authenticate header is used by a server to request that a client to provide a session identifier where none was given or to specify one for an alternative realm. This mechanism permits linkage of identifiers across realms, but only under user control.

Example

The following data shows a server requesting an identifier for the realm "w3.org".

HTTP/1.1 401 Unauthorized
WWW-Authenticate: Session, realm=w3.org
Server: libwww/4.1

Clients must not automatically respond to a WWW-Authenticate challenge without user direction.

A client may offer the user a facility whereby requests for session identifiers in alternative names are automatically accepted provided they are compatible. Realms may be considered compatible provided they are a non trivial prefix of the server dns name. For example a server `www.w3.org` request for the session identifier in the realm `w3.org` would be regarded as compatible but requests for `w3.com`, `mit.edu` or `org` would not. DNS names in the toplevel domains `com`, `edu`, `gov`, `mil` and `org` may generally be considered non trivial prefixes (the exclusion of `net` from this list is intentional. Other DNS domains may be considered non trivial prefixes if they are below the second level of the DNS hierarchy.

Security Considerations

Security considerations are discussed throughout this paper in addition to this section.

Unintended Linkage

Collusion between sites may permit linkage of session identifiers between realms. A server may permit linkage between identifiers within its own realm and another by incorporating the identifier component in a URI. The server `www.w3.org` receiving the session identifier `SID:ANON:www.w3.org:j6oA0xCWZh/CD723LGeX1f-01:34` could construct an identifier `http://ai.mit.edu/link/j6oA0xCWZh/CD723LGeX1f`. If the link was followed the server `ai.mit.edu` would be able to track the user's activity across both realms.

Unsafe Construction Techniques.

Care must be taken in constructing session identifiers. A keyed digest technique known to be cryptographically sound is recommended. In particular implementors should note that a number of techniques for constructing MACs from ciphers using XOR functions are insecure for this application.

Further Work

Data Escrow Agents.

The method for constructing session identification URIs described provides only one possible compromise between privacy and tracking. In particular no provision is made for supporting joint registration services. Such services would permit a user to register demographic details (age, sex, interests etc.) with a single server

Data Escrow Agents support Joint registration services without compromising user privacy. A data escrow agent would capture demographic data at a central location, and analyze content providers log files on their behalf. Escrow agents would be responsible for preventing content providers receiving data detailed enough to compromise user privacy.

In order to protect user privacy session identifiers must only be linkable by the data escrow agent. This may be achieved using either public key cryptography or message authentication codes.

In an implementation of a data escrow agent using public keys the data escrow agent provides each content provider with the public component of a public key pair. A user visiting a content provider's site first creates a session identifier as if the escrow agent's realm were to be visited then encrypts it using the content provider's

public key to create a session identifier specific to the content provider. In order to analyze a log file the escrow agent decrypts the session identifiers using the private portion of the key.

In an implementation of a data escrow agent using a MAC, the user provides the data escrow agent with demographic data indexed by a session identifier keyed to the agent's realm. When contacting a content provider the client constructs a session identifier using a MAC of the session identifier keyed by the provider's realm. The escrow agent may construct a linkage between the provider's logfiles and entries in the escrowed database by calculating a MAC for every entry in the database. Although this technique involves a larger number of operations than the public key based scheme, these operations are approximately four orders of magnitude faster.

Interaction with Proxies and Caches.

Many Web users browse the Web through a caching proxy. In many countries this mode of operation is essential due to saturation of international network connections. When a proxy serves a user from a local cache the originating server has no knowledge of the transaction. Consequently logfiles may be incomplete. This problem is most serious for commercial sites which use hit counts as a measure of readership.

A number of techniques may be used to prevent proxies from caching data. This permits demographic data to be collected at the cost of severely reducing network response. In a significant number of cases this will prevent a user from receiving any data at all [Smith96].

A better solution is to provide a mechanism whereby a proxy supplies a server on request with a log of hits served from the cache. Such logs are potentially of value as an indication of audited circulation, particularly if they were to be authenticated using a digital signature technique. In some circumstances it may be desirable for providers of such information to mask usernames by using session identifiers. It is intended to address these issues in a separate document.

Acknowledgments

Dave Raggett made the original proposal to use an anonymous session identifier for capture of demographic data. Rohit Khare and Dan Connolly helped refine many of the ideas. Roger Hurwitz and John Mallery made many helpful comments on early versions of this draft.

References

- [Netscape95]
Netscape Communications Corp. [Persistent client State HTTP Cookies](#)
- [Hallam96]
Phillip M. Hallam-Baker [Extended Log File Format](#)
- [Kristol95]
Kristol, D. [Proposed HTTP State-Info Mechanism](#)
- [Connolly96]
Dan Connolly [Proposals for Gathering Consumer Demographics](#)
- [Hallam93]
Phillip M. Hallam-Baker *Design note on HTTP referer field*. Memo to Tim Berners-Lee.
- [Smith96]
Neil Smith, Address at MIT [Workshop on Internet Survey Methodology and Web Demographics](#) January 29-30 1996. Cambridge Ma.
- [Rivest92]
Rivest, R., ["The MD4 Message-Digest Algorithm"](#), RFC 1321, MIT and RSA Data Security, Inc., April 1992
- [Berners-Lee96]
Tim Berners-Lee, Roy T. Fielding, and Henrik Frystyk Nielsen. [Hypertext Transfer Protocol -- HTTP/1.0](#)

[Gill96]

Neil Smith, Address at MIT [*Workshop on Internet Survey Methodology and Web Demographics*](#) January 29-30 1996. Cambridge Ma.

[RFC1034]

P. Mockapetris. [*Domain Name System*](#) . (RFC1034, RFC1035) November 1987

[Hallam-Baker94]

Phillip M. Hallam-Baker *Shen Secure Hypertext Environment, Design Notes*. CERN Programming Techniques Group.