INTERNET DRAFT
Expires in six months

Phillip M. Hallam-Baker, W3C
email: <hallam@w3.org>
Brian Behlendorf
email: <brian@organic.com>
21st February 1996

Extended Log File Format

<draft-hallam-http-logfile-00.txt>

Status of this Memo

Extended Log File Format

W3C Working Draft _WD-logfile-960221_

This version:
http://www.w3.org/pub/WWW/TR/WD-logfile-960221.html

Latest version:
http://www.w3.org/pub/WWW/TR/WD-logfile.html

Authors:

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

Extended Log File Format

Brian Behlendorf <brian@organic.com>

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.

Extended Log File Format

Abstract

An improved format for Web server log files is presented. The format is extensible, permitting a wider range of data to be captured. This proposal is motivated by the need to capture a wider range of data for demographic analysis and also the needs of proxy caches.

Introduction

Most Web servers offer the option to store logfiles in either the common log format or a proprietary format. The common log file format is supported by the majority of analysis tools but the information about each server transaction is fixed. In many cases it is desirable to record more information. Sites sensitive to personal data issues may wish to omit the recording of certain data. In addition ambiguities arise in analysing the common log file format since field separator characters may in some cases occur within fields. The extended log file format is designed to meet the following needs:

* Permit control over the data recorded.

* Support needs of proxies, clients and servers in a common format

* Provide robust handling of character escaping issues

* Allow exchange of demographic data.

* Allow summary data to be expressed.

The log file format described permits customized logfiles to be recorded in a format readable by generic analysis tools. A header specifying the data types recorded is written out at the start of each log.

This work is in part motivated by the need to support collection of demographic data. This work is discussed at greater length in companion drafts describing session identifier URIs [Hallam96a] and more consistent proxy behaviour [Hallam96b].

Format

A extended log file contains a sequence of _lines_ containing ASCII characters terminated by either the sequence CR or CRLF. Log file generators should follow the line termination convention for the platform on which they are executed. Analysers should accept either form. Each line may contain either a _directive_ or an _entry_.

Entries consist of a sequence of _fields_ relating to a single HTTP transaction. Fields are separated by whitespace, the use of tab characters for this purpose is encouraged. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself.

Extended Log File Format

The following directives are defined:

Version: _<integer>_._<integer>_
: The version of the extended log file format used. This draft defines version 1.0.

Fields: [_<specifier>_...]
: Specifies the fields recorded in the log.

Software: _string_
: Identifies the software which generated the log.

Start-Date: _<date>_ _<time>_
: The date and time at which the log was started.

End-Date:_<date>_ _<time>_
: The date and time at which the log was finished.

Date:_<date>_ _<time>_
: The date and time at which the entry was added.

Remark: _<text>_
: Comment information. Data recorded in this field should be ignored by analysis tools.

The directives Version and Fields are required and should preceed all entries in the log. The Fields directive specifies the data recorded in the fields of each entry.

Example

The following is an example file in the extended log format:

```
#Version: 1.0
#Date: 12-Jan-1996
#Fields: time cs-method cs-uri
00:34:23 GET /foo/bar.html
12:21:16 GET /foo/bar.html
12:45:52 GET /foo/bar.html
12:57:34 GET /foo/bar.html
```

Fields

The #Fields directive lists a sequence of _field identifiers_ specifying the information recorded in each entry. Field identifiers may have one of the following forms:

identifier
: Identifier relates to the transaction as a whole.

prefix-_identifier_

: Identifier relates to information transfer between parties defined by the value _prefix_.

prefix(_header_)
: Identifies the value of the HTTP header field _header_ for transfer between parties defined by the value _prefix_. Fields specified in this manner always have the value <string>.

The following prefixes are defined:

cs
: Client to Server.

sc
: Server to Client.

sr
: Server to Remote Server, this prefix is used by proxies.

rs
: Remote Server to Server, this prefix is used by proxies.

x

Application specific identifier.

The identifier cs-method thus refers to the method in the request sent by the client to the server while sc(Content-Type) refers to the content type field of the reply.

Identifiers.

The following identifiers do not require a prefix

date
Date at which transaction completed, field has type <date>

time
Time at which transaction completed, field has type <time>

time-taken
Time taken for transaction to complete in seconds, field has type <fixed>

bytes
bytes transfered, field has type <integer>

cached
Records wether a cache hit occured, field has type <integer> 0 indicates a cache miss.

The following identifiers require a prefix

ip

IP address and port, field has type <address>

dns
DNS name, field has type <name>

status
Status code, field has type <integer>

comment
Comment returned with status code, field has type <>

method
Method, field has type <name>

uri

URI, field has type <uri>

uri-stem
Stem protion alone of URI (omitting query), field has type <uri>

uri-query
Query portion alone of URI, field has type <uri>

host
DNS hostname used, field has type <name>

Special fields for log summaries.

Analysis tools may generate log summaries. A log summary lists the number of operations performed on a URI but omits information specific to a particular transaction.

The following field is mandatory and must preceed all others:

count
The number of entries for which the listed data, field has type <>

The following fields may be used in place of time to allow aggregation of log file entries over intervals of time.

time-from
Time at which sampling began, field has type <time>

time-to
Time at which sampling ended, field has type <time>

interval
Time over which sampling occurred in seconds, field has type <integer>

Entries

Extended Log File Format

This section describes the data formats for log file field entries. These formats are chosen so as to avoid ambiguity, minimize the difficulty of generation and parsing and provide for human readability.

Each logfile entry consists of a sequence of fields separated by whitespace and terminated by a CR or CRLF sequence. The meanings of

the fields are defined by a preceeding #Fields directive. If a field is ommitted for a particular entry a single dash "-" is substituted.

Log file parsers should be tolerant of errors. If an entry contains corrupt data or is terminated unexpectedly the parser should resynchronize using the end of line marker and continue to parse the following entries. Entries must not contain any ASCII control characters.

```
<entry> = <field>* <end-of-line>

<field> = <integer> | <fixed> | <uri> | <date> | <time> | <string>

<digit = "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
```

Integer

```
<integer>    = <digit> +
```

Integers are represented as a sequence of digits.

Fixed Format Float

```
<float>      = <digit> [. <digit>*]
```

URI

A URI as specified by RFC1738, relative URIs are specified by RFC1808. URIs cannot by definition include whitespace or ASCII control characters. Consequently no ambiguity arises from their use.

Date

```
<date>       = <digit> <digit> <digit> <digit>
                    "-" <digit> <digit> "-" <digit> <digit>
```

Dates are recorded in the format YYYY-MM-DD where YYYY, MM and DD stand for the numeric year, month and day respectively. This format

is chosen to assist collation using sort.

Time

```
<time>       = <digit> <digit> ":" <digit> <digit>
                      [":" <digit> <digit> ["." <digit>*]
```

Times are recorded in the form HH:MM, HH:MM:SS or HH:MM:SS.S where HH is the hour in 24 hour format, MM is minutes and SS is seconds.

String

```
<string>     = '"' <schar>* '"'

<schar>      = xchar | '"' '"'
```

Strings are output in quoted form. If a string contains a quotation character the character is repeated. This format is unambiguous since fields are by definition separated by whitespace.

No mechanism for incorporating control characters is defined.

Text

```
<text>       =  <char>*
```

The text field is used only by directives.

Name

```
<name>       =  <alpha> [ "." <alpha> * ]
```

DNS name.

Address

```
<name>       =  <integer> [ "." <integer> * ] [ ":" <integer> ]
```

Numeric IP address and optional port specifier.

Acknowledgements.

Extended Log File Format

Robert Thau provided usefull advice and some code. John Mallery and Roger Hurwitz helped develop many of the ideas.

Phillip M. Hallam-Baker
hallam@w3.org
World Wid Web Consortium
Cambridge MA

Brian Behlendorf
brian@organic.com
Organic Online

References.

[RFC1808]
R. Fielding _ Relative Uniform Resource Locators_, June 1995

[RFC1738]
T. Berners-Lee, L. Masinter, _ Uniform Resource Locators (URL)_, December 1994

[Luotonen95]
A. luotonen, _ The Common Logfile Format_, 1995, http://www.w3.org/pub/WWW/Daemon/User/Config/Logging.html

[Hallam96a]
P. M. Hallam-Baker _ Session Identification URI_ World Wide Web Consortium Working Draft, WD-session-id.

[Hallam96b]
P. M. Hallam-Baker _ Notification for Proxy Caches_ World Wide Web Consortium Working Draft, WD-proxy.