[MS-FSWADF]:
WebAnalyzer Data Files Format Specification

Intellectual Property Rights Notice for Open Specifications Documentation

- **Technical Documentation.** Microsoft publishes Open Specifications documentation for protocols, file formats, languages, standards as well as overviews of the interaction among each of these technologies.
- **Copyrights.** This documentation is covered by Microsoft copyrights. Regardless of any other terms that are contained in the terms of use for the Microsoft website that hosts this documentation, you may make copies of it in order to develop implementations of the technologies described in the Open Specifications and may distribute portions of it in your implementations using these technologies or your documentation as necessary to properly document the implementation. You may also distribute in your implementation, with or without modification, any schema, IDL's, or code samples that are included in the documentation. This permission also applies to any documents that are referenced in the Open Specifications.
- **No Trade Secrets.** Microsoft does not claim any trade secret rights in this documentation.
- **Patents.** Microsoft has patents that may cover your implementations of the technologies described in the Open Specifications. Neither this notice nor Microsoft's delivery of the documentation grants any licenses under those or any other Microsoft patents. However, a given Open Specification may be covered by Microsoft's Open Specification Promise (available here: http://www.microsoft.com/interop/osp) or the Community Promise (available here: http://www.microsoft.com/interop/cp/default.mspx). If you would prefer a written license, or if the technologies described in the Open Specifications are not covered by the Open Specifications Promise or Community Promise, as applicable, patent licenses are available by contacting iplg@microsoft.com.
- **Trademarks.** The names of companies and products contained in this documentation may be covered by trademarks or similar intellectual property rights. This notice does not grant any licenses under those rights.
- **Fictitious Names.** The example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted in this documentation are fictitious. No association with any real company, organization, product, domain name, email address, logo, person, place, or event is intended or should be inferred.

Reservation of Rights. All other rights are reserved, and this notice does not grant any rights other than specifically described above, whether by implication, estoppel, or otherwise.

Tools. The Open Specifications do not require the use of Microsoft programming tools or programming environments in order for you to develop an implementation. If you have access to Microsoft programming tools and environments you are free to take advantage of them. Certain Open Specifications are intended for use in conjunction with publicly available standard specifications and network programming art, and assumes that the reader either is familiar with the aforementioned material or has immediate access to it.

Revision Summary

Date	Revision History	Revision Class	Comments
07/13/2009	0.1	Major	Initial Availability
02/19/2010	1.0	Editorial	Revised and edited the technical content
03/31/2010	1.01	Editorial	Revised and edited the technical content
04/30/2010	1.02	Editorial	Revised and edited the technical content
06/07/2010	1.03	Editorial	Revised and edited the technical content
06/29/2010	1.04	Editorial	Changed language and formatting in the technical content.
07/23/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
09/27/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
11/15/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
12/17/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
03/18/2011	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
06/10/2011	1.04	No change	No changes to the meaning, language, or formatting of the technical content.

Table of Contents

1 Introduction

This document specifies the WebAnalyzer Data File Format, which is used to store information in files during anchor analysis.

1.1 Glossary

The following terms are defined in [MS-GLOS]:

Augmented Backus-Naur Form (ABNF)
Coordinated Universal Time (UTC)
little-endian
MD5 hash
UTF-8

The following terms are defined in [MS-OFCGLOS]:

anchor text
content collection
document identifier
equivalence class
hyperlink
item
rank
site

The following terms are specific to this document:

MAY, SHOULD, MUST, SHOULD NOT, MUST NOT: These terms (in all caps) are used as described in [RFC2119]. All statements of optional behavior use either MAY, SHOULD, or SHOULD NOT.

1.2 References

1.2.1 Normative References

We conduct frequent surveys of the normative references to assure their continued availability. If you have any issue with finding a normative reference, please contact dochelp@microsoft.com. We will assist you in finding the relevant information. Please check the archive site, http://msdn2.microsoft.com/en-us/library/E4BD6494-06AD-4aed-9823-445E921C9624, as an additional source.

[MS-FSFDMW] Microsoft Corporation, "FAST Distributed Make Worker Protocol Specification"

[MS-FSIN] Microsoft Corporation, "Input Normalization Data Structure"

[MS-FSWASDR] Microsoft Corporation, "WebAnalyzer/SPRel Data Receiving Protocol Specification"

[MS-FSWASDS] Microsoft Corporation, "WebAnalyzer/SPRel Data Serving Protocol Specification"

[MS-FSWCU] Microsoft Corporation, "WebAnalyzer/Crawler Utility Structure Specification"

[RFC1950] Deutsch, P., and Gailly, J-L., "ZLIB Compressed Data Format Specification version 3.3", RFC 1950, May 1996, http://www.ietf.org/rfc/rfc1950.txt

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, http://www.rfc-editor.org/rfc/rfc2119.txt

[RFC3986] Berners-Lee, T., Fielding, R., and Masinter, L., "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005, http://www.ietf.org/rfc/rfc3986.txt

[RFC5234] Crocker, D., Ed., and Overell, P., "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008, http://www.rfc-editor.org/rfc/rfc5234.txt

1.2.2 Informative References

[MS-GLOS] Microsoft Corporation, "Windows Protocols Master Glossary".

[MS-OFCGLOS] Microsoft Corporation, "Microsoft Office Master Glossary".

1.3 Structure Overview (Synopsis)

This document describes how to store information during **anchor text** relevance analysis in the system. The analysis consists of many stages and every stage uses its own format for the information it processes.

1.4 Relationship to Protocols and Other Structures

The file formats in this document are used by the protocol described in [MS-FSFDMW]. The initial input files in section 2.2 are produced by the protocol described in [MS-FSWASDR]. The output files in section 2.5 are used to implement the protocol described in [MS-FSWASDS].

1.5 Applicability Statement

None.

1.6 Versioning and Localization

None.

1.7 Vendor-Extensible Fields

None.

2 Structures

This section specifies the format for each file type.

2.1 Common file structures

Either a file is empty, or it MUST contain a set of rows. Each row consists of one or more columns terminated with a newline, which is either a carriage return character combined with a line feed character, or is only a line feed character. Columns MUST be separated by a white space delimiter. If the column does not contain binary data, it MUST be encoded in **UTF-8**.

The common structure for a file that does not contain binary data corresponds to the following rules written in **Augmented Backus-Naur Form (ABNF),** as specified in [RFC5234].

```
FILE = *LINE

; The following sections specify the rules for each type of ROW

LINE = ROW NEWLINE

AFREQ = COUNT

ARANK = RANK

LAFREQ = COUNT

LARANK = RANK

TO-RANK = RANK

SITE-RANK = RANK

SITE-OR-TO-URL = SITE / URL

ANCHORTEXT = TOKEN *(SP TOKEN)

BASE64-CHAR = ALPHA / DIGIT / "=" / "+" / "/"

BASE64 = 1*BASE64-CHAR

CLASS = URLHASH

COUNT = 1*DIGIT

EQREPR = %x00 / *(URL %x00) URL

FROM = URLHASH

TO = URLHASH

INTRA = "0" / "1"

MEMBER = URLHASH

RANK = (1*DIGIT "." 1*DIGIT) / 1*DIGIT

SITE = URL / %xc7 %x82
```

TIMESTAMP = 1*DIGIT

TOKEN = 1*(%x21-ff)

URL = 1*(%x21-ff)

URLHASH = 21*21(BASE64-CHAR)

NEWLINE = (CRLF / LF)

Exceptions to this general structure are specified where applicable. Some of the ABNF rules are specified in the following table.

Column name	Description
ANCHORTEXT	The anchor text from a **hyperlink**. The tokens of the anchor text consist of UTF-8 encoded characters, excluding control characters and white space, and they are normalized as specified in [MS-FSIN]. The tokens are separated by one space character.
BASE64	MUST be a sequence of bytes in base 64 encoding.
CLASS	The main **item** of an equivalence class.
EQREPR	A **string** that contains all the items in the **equivalence class** of an item. The items are delimited by a null byte, which is the hexadecimal character 0x00.
INTRA	Specifies the location of the destination URL relative to the source URL of a hyperlink. The value MUST be 1 if the hyperlink points to a URL that is located on the same **site(1)** or site(2) as the source URL; otherwise it MUST be 0.
MEMBER	A member of an equivalence class.
RANK	A quality score assigned to an item or an anchor text during the relevance analysis. The quality score is a measure of the quality and importance for relevancy for the specified item or anchor text. The quality score is part of the **rank** score for an item in the system, and is specified as a floating point decimal number.
SITE	The site(1) or site(2) of an item. If a site is not available, the value MUST be the two hexadecimal bytes 0xc782.
TIMESTAMP	A timestamp that specifies when an event occurred. This is an integer that specifies the time in seconds elapsed after 00:00:00 1970-01-01 **UTC**.
TOKEN	Token encoded in UTF-8 and normalized as specified in [MS-FSIN].
URL	The URL of an item. The complete ABNF rule set for URLs is specified in [RFC3986].
URLHASH	Represents the **document identifier(3)** of an item. This MUST contain the **MD5 hash** value of the URL of the item in base 64 encoding. The length MUST be 21 characters.

The file formats in the following subsections are specified in the order of their use in the analysis process. Information flows from formats in the earlier sections to formats in the later sections.

The overall pattern is that the input files in section 2.2 are created with the format specified in [MS-FSWASDR]. They are transformed initially using the file formats in section 2.3. The main anchor text relevance analysis, as specified in [MS-FSFDMW], uses the file formats in section 2.4. The analysis process creates two outputs, one of which is the database files that use the formats in section 2.5. The other part is the files used to send partial updates to the system. Their formats are specified in section 2.6.

2.2 Input Files

This section specifies the initial input files for the anchor text relevance analysis, as described in [MS-FSFDMW].

2.2.1 delete

This file specifies rows that represent delete operations submitted to the system. The rows are also described in [MS-FSWASDR] section 2.2.1.3.1.4. Rows are specified by the following ABNF rule.

ROW = URLHASH SP TIMESTAMP

The following table specifies the columns.

Column name	Description
URLHASH	The item to delete from the system. This item, and hyperlinks associated with this item, MUST be excluded from future anchor text relevance analysis, as described in [MS-FSFDMW].
TIMESTAMP	The time that the delete operation was submitted.

2.2.2 eqrepr

This file specifies rows that represent the equivalence class of an item. The equivalence class information is also described in [MS-FSWASDR] section 2.2.1.3.1.3. The anchor text relevance analysis uses the equivalence class to normalize hyperlinks, as described in [MS-FSFDMW]. Rows are specified by the following ABNF rule.

ROW = URLHASH SP URL SP TIMESTAMP SP EQREPR

The following table specifies the columns.

Column name	Description
URLHASH	The item that owns the equivalence class, represented by a hash value that specifies the source or destination of a hyperlink when normalizing hyperlinks. The item is used instead of other members of the equivalence class, as described in [MS-FSFDMW].
URL	The item that owns the equivalence class, represented by an URL.
TIMESTAMP	The time that the equivalence class was submitted to the system.
EQREPR	The equivalence class of the item.

2.2.3 links

The file specifies rows that represent hyperlinks, including the anchor text. The links are the main component of the anchor text relevance analysis, as described in [MS-FSFDMW]. The rows are also described in [MS-FSWASDR] section 2.2.1.3.1.1. Rows are specified by the following ABNF rule.

ROW = FROM SP TO SP INTRA SP TIMESTAMP SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
FROM	The source URL of the hyperlink, represented as a hash value.
TO	The destination URL of the hyperlink, represented as a hash value.
INTRA	Specified in section 2.1.
TIMESTAMP	The time that the link was submitted to the system.
ANCHORTEXT	The anchor text of the hyperlink.

2.2.4 no_links

This file specifies rows that represent an item that has no outgoing hyperlinks. The item is not present in the **FROM** column specified in section 2.2.3. The protocol makes this information consistent for anchor text relevance analysis, as described in [MS-FSFDMW]. The information is described in [MS-FSWASDR] section 2.2.1.3.1.5.

Files of this type contain rows that are specified with the following ABNF rule.

ROW = URL SP URLHASH SP TIMESTAMP

The following table specifies the columns.

Column name	Description
URL	The item that has no hyperlinks.
URLHASH	The hash value of the item that has no hyperlinks.
TIMESTAMP	The time that the item was submitted to the system.

2.2.5 sitemap

This file specifies rows that represent a mapping from an item to the site(2) of the item. The information is described in [MS-FSWASDR] section 2.2.1.3.1.7. During the anchor text relevance analysis it is important to keep track of the site of an item, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URLHASH SP SITE

The following table specifies the columns.

Column name	Description
URLHASH	The hash value of an item.
SITE	The site(2) of the item.

2.2.6 urieq

This file specifies rows that represent a mapping between an item and a member of the equivalence class of the item. There are multiple rows for the same item if the equivalence class contains more than one member. The information is also described in [MS-FSWASDR] section 2.2.1.3.1.2. The

mapping is important for finding the equivalence class of an item during the anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = CLASS SP MEMBER SP TIMESTAMP

The following table specifies the columns.

Column name	Description
CLASS	Represents an item. This item identifies the equivalence class.
MEMBER	A member of the equivalence class. Can be equal to the **CLASS** column.
TIMESTAMP	The time that the mapping was submitted to the system.

2.2.7 urimap

This file specifies rows that represent a mapping from a URL to the hash value of the URL. The information is also described in [MS-FSWASDR] section 2.2.1.3.1.6. This mapping is used to change the representation of an item from an URL to a hash value and back again. The anchor text relevance analysis mainly uses the hash value representation of an item for performance reasons, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URL SP URLHASH

The following table specifies the columns.

Column name	Description
URL	The URL of an item.
URLHASH	The hash value of the URL.

2.3 Initial Processing Files

2.3.1 links_by_to

This file specifies rows that represent a hyperlink including the anchor text. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = FROM SP TO SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
FROM	The source URL of the hyperlink.
TO	The destination URL of the hyperlink.

Column name	Description
ANCHORTEXT	The anchor text of the hyperlink.

2.3.2 links_by_to_raw

This file specifies rows that represent hyperlinks, including the anchor text. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW]. Rows are specified by the following ABNF rule.

ROW = FROM SP SITE SP TO SP INTRA SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
FROM	The source URL of the hyperlink.
SITE	The site of the source URL in the **FROM** column.
TO	The destination URL of the hyperlink.
INTRA	Specified in section 2.1.
ANCHORTEXT	The anchor text of the hyperlink.

2.3.3 urieq_by_class

This file specifies rows that represent the same mapping specified in section 2.2.6, but without the **TIMESTAMP** column. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = CLASS SP MEMBER

2.3.4 eqrepr_by_uri

This file specifies rows that represent an item and its equivalence class. It is an intermediate file used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URL SP EQREPR

The following table specifies the columns.

Column name	Description
URL	The URL of the item.
EQREPR	The equivalence class of the item.

2.3.5 urihash

This file specifies rows that represent an item. It is an intermediate file used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URLHASH

The following table specifies the columns.

Column name	Description
URLHASH	The hash value of the URL of an item.

2.4 Main Processing Files

The file formats in the following subsections are specified in the order of their use in the analysis. Information flows from formats in the earlier sections to formats in the later sections.

2.4.1 rank_links_by_src

This file specifies rows that represent a hyperlink between two items, and the number of links total from the source item. It is an intermediate file used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = FROM SP TO SP COUNT

The following table specifies the columns.

Column name	Description
FROM	The source of the hyperlink.
TO	The destination of the hyperlink.
COUNT	The number of hyperlinks from the source.

2.4.2 rank_by_uri

This file specifies rows that contain the quality score for an item during anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URLHASH SP RANK

The following table specifies the columns.

Column name	Description
URLHASH	The hash value of the URL of an item.
RANK	The quality score specified for an item.

2.4.3 linkscore_by_dst

This file specifies rows that represent a hyperlink and the quality score that the anchor text relevance analysis process associates with the destination item of the hyperlink, as described in [MS-FSFDMW]. Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = FROM SP TO SP RANK

The following table specifies the columns.

Column name	Description
FROM	The source of the hyperlink.
TO	The destination of the hyperlink.
RANK	The quality score associated with the destination item.

2.4.4 links_norm_with_fromrank_by_anchor

This file specifies rows that represent a hyperlink, and the quality score associated with the source item of the hyperlink during anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = RANK SP FROM SP TO SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
RANK	The quality score associated with the source item.
FROM	The source of the hyperlink.
TO	The destination of the hyperlink.
ANCHORTEXT	The anchor text of the hyperlink.

2.4.5 anchor_freqs_by_anchor

This file specifies rows that contain information about the frequency and quality score associated with the specified anchor text. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = AFREQ SP ARANK SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
AFREQ	The frequency count of the anchor text in hyperlinks in the system.
ARANK	The calculated quality score of the anchor text in the system.

Column name	Description
ANCHORTEXT	The anchor text.

2.4.6 links_with_freqs_by_to

This file specifies rows that contain information about the frequency and quality score of the anchor text in a hyperlink. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = RANK SP AFREQ SP ARANK SP TO SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
RANK	The quality score of the source of the hyperlink.
AFREQ	The frequency count of the anchor text across all hyperlinks in the system.
ARANK	The calculated quality score of the anchor text across all hyperlinks in the system.
TO	The destination of the hyperlink.
ANCHORTEXT	The anchor text from the hyperlink.

2.4.7 uri_anchors_by_urihash

This file specifies rows that contain information about the frequency and quality score associated with anchor text across all hyperlinks that point to a specified destination. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = TO-RANK SP LAFREQ SP LARANK SP AFREQ SP ARANK SP TO SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
TO-RANK	The quality score of the destination of the hyperlinks.
LAFREQ	The frequency count of the anchor text across the hyperlinks pointing to the destination, the **TO** column.
LARANK	The calculated quality score of the anchor text across all hyperlinks that point to the destination, the **TO** column.
AFREQ	The frequency count of the anchor text across all hyperlinks in the system.
ARANK	The calculated quality score of the anchor text across all hyperlinks in the system.
TO	The common destination of the hyperlinks.
ANCHORTEXT	The anchor text from the hyperlinks.

2.4.8 anchor_by_to

This file specifies rows where the columns in the row are the same as in section 2.4.7, except for the **TO** column which was replaced by the actual URL in the **URL** column. In addition, the **SITE** column contains the site(2) of the item identified by the **URL** column. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = TO-RANK SP LAFREQ SP LARANK SP AFREQ SP ARANK SP URL SP SITE SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
TO-RANK	The quality score of the destination of the hyperlinks.
LAFREQ	The frequency count of the anchor text across all hyperlinks that point to the destination, the **URL** column.
LARANK	The calculated quality score of the anchor text across all hyperlinks that point to the destination, the **URL** column.
AFREQ	The frequency count of the anchor text across all hyperlinks in the system.
ARANK	The calculated quality score of the anchor text across all hyperlinks in the system.
URL	The common destination of the hyperlinks.
SITE	The site of the item identified by the **URL** column. If a site is not available, the value MUST be the two hexadecimal bytes 0xc782.
ANCHORTEXT	The anchor text from the hyperlinks.

2.4.9 rank_by_site

This file specifies rows that contain the calculated quality score for an item. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = SITE SP RANK SP URL

The following table specifies the columns.

Column name	Description
SITE	The site part of the URL of the item.
RANK	The calculated quality score for the item.
URL	The item.

2.4.10 siterank_by_uri

This file specifies rows that contain the calculated quality score for a specified site. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = SITE-OR-TO-URL SP SITE-RANK

The following table specifies the columns.

Column name	Description
SITE-OR-TO-URL	A site, or the URL of an item.
SITE-RANK	The calculated quality score for the site, or the site part of the URL if the **SITE-OR-TO-URL** column contains an item rather than a site.

2.4.11 anchor_by_uri

This file specifies rows that contain information about the frequency and quality scores associated with anchor text across all hyperlinks that point to a specified destination. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = SITE-RANK SP TO-RANK SP LAFREQ SP LARANK SP AFREQ SP ARANK SP SITE-OR-TO-URL SP ANCHORTEXT

The following table specifies the columns.

Column name	Description
SITE-RANK	The calculated quality score for the site specified by the **SITE-OR-TO-URL** column.
TO-RANK	The quality score of the destination of the hyperlinks.
LAFREQ	The frequency count of the anchor text across all hyperlinks that point to the destination, the **SITE-OR-TO-URL** column.
LARANK	The calculated quality score of the anchor text across all hyperlinks that point to the destination, the **SITE-OR-TO-URL** column.
AFREQ	The frequency count of the anchor text across all hyperlinks in the system.
ARANK	The calculated quality score of the anchor text across all hyperlinks in the system.
SITE-OR-TO-URL	The common destination of the hyperlinks. Either the site or the URL of an item.
ANCHORTEXT	The anchor text from the hyperlinks.

2.4.12 anchor_by_uri_with_repr

This file specifies rows where the columns in the row are the same as specified in section 2.4.11, with the addition of the leading **EQREPR** column. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = EQREPR SP SITE-RANK SP TO-RANK SP LAFREQ SP LARANK SP AFREQ SP ARANK SP SITE-OR-TO-URL SP ANCHORTEXT

The following table specifies the **EQREPR** column.

Column name	Description
EQREPR	The equivalence class of the item in the **SITE-OR-TO-URL** column if that column contains an item. If the **SITE-OR-TO-URL** column contains a site, the **EQREPR** column MUST contain the null byte 0x00.

2.4.13 anchor_info_new

This file specifies rows that contain information about the frequency and quality score for all anchor text that point to a specified destination. It is an intermediate format used for anchor text relevance analysis, as described in [MS-FSFDMW].

Files of this type MUST contain rows that are specified with the following ABNF rules.

ROW = SITE-OR-TO-URL-HASH SP ANCHORINFO

SITE-OR-TO-URL-HASH = 1*39DIGIT

ANCHORINFO = BASE64

The following table specifies the columns.

Column name	Description
SITE-OR-TO-URL-HASH	This field is computed from the **SITE-OR-TO-URL** column from section 2.4.12. It MUST contain the 128-bit MD5 hash of the URL, which is a big-endian 128-bit unsigned hexadecimal integer. The integer MUST be converted to base 10 and encoded as an ASCII **string**.
ANCHORINFO	This column contains a dictionary, composed of key-value pairs. The dictionary MUST contain the keys "anchors", "queries", "contentid", "rank", "siterank" and "urieqs". The key fields contain values as specified in [MS-FSFDMW]. The dictionary is serialized as specified in [MS-FSWCU], and then encoded using base 64 encoding.

2.5 Database Files

These three file formats create an in-memory lookup database. This database is used as a back end for an implementation of the protocol specified in [MS-FSWASDS]. The database is the final output of the anchor text relevance analysis, and files using the three file formats are produced as specified in [MS-FSFDMW].

2.5.1 bin

Files of this type contain binary data that represents a set of records. The record size MUST be a multiple of 32, specified as a 32-bit signed integer in little-endian order before each record. If the record size is not a multiple of 32, the record MUST be padded with zeros.

A record contains a dictionary, composed of key-value pairs. The dictionary MUST be serialized as specified in [MS-FSWCU]. A key-value pair MUST be specified in two **strings**. The first **string** specifies the key, and the second **string** specifies the value. There is one exception for the type of

the value: In the first record, if the key is the string "offset_step", then the value MUST be the integer 32. Key-value pairs MUST be serialized as specified in [MS-FSWCU].

All records except the first record MUST be compressed using the zlib format, as specified in [RFC1950]. For each compressed record, the compression method and flags header MUST be removed. This means the protocol removes the two first bytes 0x789C from every compressed record.

The first record is a header record whose size MUST be set to 124. This record MUST contain the following dictionary.

Key	Value
offset_step	32
len_field_type	I
serializer	pyfastmarshal
compression_type	gzip

The remaining records are compressed and contain dictionaries that contain key-value pairs as specified for the **ANCHORINFO** column in section 2.4.13.

2.5.2 idx

Files of this type contain binary data that represents a set of hash values. Each hash value MUST be computed with the 32 most significant bits of a 128-bit MD5 hash. The 4-byte hash value MUST be specified in **little-endian** order. The 128-bit MD5 hash is calculated from the URLs in the **SITE-OR-TO-URL-HASH** column specified in section 2.4.13. The hash values MUST be in the same order as the record entries, and they each correspond to a dictionary record specified in section 2.5.1.

2.5.3 idx.ofs

Files of this type contain binary data with file offsets to record entries in files with the format specified in section 2.5.1. A record file offset MUST be calculated by subtracting the header length, 128 bytes, from the file offset for the record, and dividing the remaining value by 32. The result is stored as a little-endian ordered 4-byte integer.

A file offset MUST NOT be calculated for the header record.

2.6 Index Update Files

2.6.1 feeduris

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URL SP COLLECTION

COLLECTION = 1*ALPHA

The row contains an item to update in the index, which includes the **content collection** of the item. The following table specifies the columns.

Column name	Description
URL	The URL of an item to update in the index.
COLLECTION	The content collection of the item.

2.6.2 pupdateuris_by_uri

Files of this type MUST contain rows that are specified with the following ABNF rule.

ROW = URL

The row contains items to update in the index. It is an intermediate format used to produce files with the format specified in section 2.6.1. The transformation is specified in [MS-FSFDMW]. The following table specifies the columns.

Column name	Description
URL	The URL of an item to update in the index.

3 Structure Examples

3.1 Input Files

3.1.1 links

Example for the file format described in section 2.2.3 is as follows.

```
Fe8Dff3DFxc+Dy8Q7vFzB 5ClAoACOAY2/Wx8KQiu3h 0 1261403658 example4
UwX+GSFpNgCh/xqPs5+pF 5ClAoACOAY2/Wx8KQiu3h 0 1261403672 example4
ccfq2he9YKIihd9szQ01l 5ClAoACOAY2/Wx8KQiu3h 0 1261403681 example4
hjLsdjaTfOvlbOt6kFImU E90S080WDLdHl81dk8Vak 0 1261403690 example5
5Roj9xC5lItjtuLA5FzkH e+toML6Kvv4qputJZXm3V 0 1261403699 none
```

3.1.2 urimap

Example for the file format described in section 2.2.7 is as follows.

```
http://www.cohowinery.com/example1.html 4nv9pjGzJ9r8bUn0ethTL
http://www.cohowinery.com/example2.html mt94KhEMAAzxJi1BxvzVx
http://www.cohowinery.com/example3.html k8EyJ1bIdfG9y/mME5YDZ
http://www.cohowinery.com/example4.html Rha5OAU0XwVUB5HuG5EvL
http://www.cohowinery.com/example5.html Z1GjJ824myhwBL0F+q2HA
http://www.cohowinery.com/none.html EM9KmjKegTHYQsDC6a485
```

3.2 Initial Processing Files

3.2.1 links_by_to

Example for the file format described in section 2.3.1 is as follows.

```
Z1GjJ824myhwBL0F+q2HA EM9KmjKegTHYQsDC6a485 none
4nv9pjGzJ9r8bUn0ethTL Rha5OAU0XwVUB5HuG5EvL example4
k8EyJ1bIdfG9y/mME5YDZ Rha5OAU0XwVUB5HuG5EvL example4
mt94KhEMAAzxJi1BxvzVx Rha5OAU0XwVUB5HuG5EvL example4
Rha5OAU0XwVUB5HuG5EvL Z1GjJ824myhwBL0F+q2HA example5
```

3.3 Main Processing Files

3.3.1 rank_links_by_src

Example for the file format described in section 2.4.1 is as follows.

```
4nv9pjGzJ9r8bUn0ethTL Rha5OAU0XwVUB5HuG5EvL 1
Rha5OAU0XwVUB5HuG5EvL Z1GjJ824myhwBL0F+q2HA 1
Z1GjJ824myhwBL0F+q2HA EM9KmjKegTHYQsDC6a485 1
k8EyJ1bIdfG9y/mME5YDZ Rha5OAU0XwVUB5HuG5EvL 1
mt94KhEMAAzxJi1BxvzVx Rha5OAU0XwVUB5HuG5EvL 1
```

3.3.2 anchor_freqs_by_anchor

Example for the file format described in section 2.4.5 is as follows.

```
3 0.01014 example4
1 0.1755 example5
1 1.04 none
```

3.3.3 uri_anchors_by_urihash

Example for the file format described in section 2.4.7 is as follows.

```
2.22 1 1.04 1 1.04 EM9KmjKegTHYQsDC6a485 none
0.1755 3 0.01014 3 0.01014 Rha5OAU0XwVUB5HuG5EvL example4
1.04 1 0.1755 1 0.1755 Z1GjJ824myhwBL0F+q2HA example5
```

3.3.4 anchor_by_to

Example for the file format described in section 2.4.8 is as follows.

```
0.1755 3 0.01014 3 0.01014 http://www.cohowinery.com/example4.html ǂ example4
1.04 1 0.1755 1 0.1755 http://www.cohowinery.com/example5.html ǂ example5
2.22 1 1.04 1 1.04 http://www.cohowinery.com/none.html ǂ none
```

The following is the same information shown in hexadecimal format.

```
00000000  302e 3137 3535 2033 2030 2e30 3130 3134 0.1755 3 0.01014
00000010  2033 2030 2e30 3130 3134 2068 7474 703a  3 0.01014 http:
00000020  2f2f 7777 772e 636f 686f 7769 6e65 7279 //www.cohowinery
00000030  2e63 6f6d 2f65 7861 6d70 6c65 342e 6874 .com/example4.ht
00000040  6d6c 20c7 8220 6578 616d 706c 6534 0d0a ml G. example4..
00000050  312e 3034 2031 2030 2e31 3735 3520 3120 1.04 1 0.1755 1
00000060  302e 3137 3535 2068 7474 703a 2f2f 7777 0.1755 http://ww
00000070  772e 636f 686f 7769 6e65 7279 2e63 6f6d w.cohowinery.com
00000080  2f65 7861 6d70 6c65 352e 6874 6d6c 20c7 /example5.html G
00000090  8220 6578 616d 706c 6535 0d0a 322e 3232 . example5..2.22
000000a0  2031 2031 2e30 3420 3120 312e 3034 2068  1 1.04 1 1.04 h
000000b0  7474 703a 2f2f 7777 772e 636f 686f 7769 ttp://www.cohowi
000000c0  6e65 7279 2e63 6f6d 2f6e 6f6e 652e 6874 nery.com/none.ht
000000d0  6d6c 20c7 8220 6e6f 6e65 0d0a           ml G. none..
```

3.3.5 anchor_by_uri_with_repr

Example for the file format described in section 2.4.12 is as follows.

```
1.14516666667 0.1755 3 0.01014 3 0.01014 http://www.cohowinery.com/example4.html example4
1.14516666667 1.04 1 0.1755 1 0.1755 http://www.cohowinery.com/example5.html example5
1.14516666667 2.22 1 1.04 1 1.04 http://www.cohowinery.com/none.html none
1.14516666667 0 0 0 0 0 site://www.cohowinery.com -
```

The following is the same information shown in hexadecimal format.

```
00000000  0020 312e 3134 3531 3636 3636 3636 3720 . 1.14516666667
00000010  302e 3137 3535 2033 2030 2e30 3130 3134 0.1755 3 0.01014
00000020  2033 2030 2e30 3130 3134 2068 7474 703a  3 0.01014 http:
00000030  2f2f 7777 772e 636f 686f 7769 6e65 7279 //www.cohowinery
00000040  2e63 6f6d 2f65 7861 6d70 6c65 342e 6874 .com/example4.ht
00000050  6d6c 2065 7861 6d70 6c65 340a 0020 312e ml example4.. 1.
00000060  3134 3531 3636 3636 3636 3720 312e 3034 14516666667 1.04
00000070  2031 2030 2e31 3735 3520 3120 302e 3137  1 0.1755 1 0.17
00000080  3535 2068 7474 703a 2f2f 7777 772e 636f 55 http://www.co
00000090  686f 7769 6e65 7279 2e63 6f6d 2f65 7861 howinery.com/exa
000000a0  6d70 6c65 352e 6874 6d6c 2065 7861 6d70 mple5.html examp
000000b0  6c65 350a 0020 312e 3134 3531 3636 3636 le5.. 1.14516666
000000c0  3636 3720 322e 3232 2031 2031 2e30 3420 667 2.22 1 1.04
000000d0  3120 312e 3034 2068 7474 703a 2f2f 7777 1 1.04 http://ww
000000e0  772e 636f 686f 7769 6e65 7279 2e63 6f6d w.cohowinery.com
000000f0  2f6e 6f6e 652e 6874 6d6c 206e 6f6e 650a /none.html none.
00000100  0020 312e 3134 3531 3636 3636 3636 3720 . 1.14516666667
00000110  3020 3020 3020 3020 3020 7369 7465 3a2f 0 0 0 0 0 site:/
00000120  2f77 7777 2e63 6f68 6f77 696e 6572 792e /www.cohowinery.
00000130  636f 6d20 2d0a                         com -.
```

3.3.6 anchor_info_new

Example for the file format described in section 2.4.13 is as follows.

```
9316394691999639158375674834024144465 2
e3MJAAAAY29udGVudGlkcycAAABodHRwOi8vd3d3LmNvaG93aW5lcnkuY29tL2V4YW1wbGU0Lmh0bWxzCAAAAH
NpdGVyYW5rcw0AAAAxLjE0NTE2NjY2NjY3cwcAAABhbmNob3JzWwEAAAAoBQAAAHMIAAAAZXhhbXBsZTRzAQAAADNzBwA
AADAuMDEwMTRzAQAAADNzBwAAADAuMDE
wMTRzBAAAAHJhbmtzBgAAADAuMTc1NXMGAAAAdXJpZXFzWwAAAAAw
13733436879771899130252258982736517708 8
e3MJAAAAY29udGVudGlkcycAAABodHRwOi8vd3d3LmNvaG93aW5lcnkuY29tL2V4YW1wbGU1Lmh0bWxzCAAAA
HNpdGVyYW5rcw0AAAAxLjE0NTE2NjY2NjY3cwcAAABhbmNob3JzWwEAAAAoBQAAAHMIAAAAZXhhbXBsZTVzAQAAADFzBg
AAADAuMTc1NXMBAAAAMXMGAAAAMC4xNz
U1cwQAAAByYW5rcwQAAAAxLjA0cwYAAAB1cmllcXNbAAAAADA=
22343966497388931581202431452872850660 2
e3MJAAAAY29udGVudGlkcyMAAABodHRwOi8vd3d3LmNvaG93aW5lcnkuY29tL25vbmUuaHRtbHMIAAAAc2l0ZX
JhbmtzDQAAADEuMTQ1MTY2NjY2NjdzBwAAAGFuY2hvcnNbAQAAACgFAAAAcwQAAABub25lcwEAAAAxcwQAAAAxLjA0cwE
AAAAxcwQAAAAxLjA0cwQAAAByYW5rcwQ
AAAAyLjIycwYAAAB1cmllcXNbAAAAADA=
18018513744918695430802156541584513878 2
e3MJAAAAY29udGVudGlkcxkAAABzaXRlOi8vd3d3LmNvaG93aW5lcnkuY29tcwgAAABzaXRlcmFua3MNAAAAMS
4xNDUxNjY2NjY2N3MHAAAAYW5jaG9yc1sBAAAAKAUAAABzAQAAAC1zAQAAADBzAQAAADBzAQAAADBzAQAAADBzBAAAAHJ
hbmtzAQAAADBzBgAAAHVyaWVxc1sAAAA
AMA==
```

3.4 Database Files

3.4.1 bin

Example for the file format described in section 2.5.1 is as follows.

```
  Addr     0 1  2 3  4 5  6 7  8 9  A B  C D  E F 0 2 4 6 8 A C E
--------  ---- ---- ---- ---- ---- ---- ---- ---- ----------------
```

```
00000000  7c00 0000 7b73 0b00 0000 6f66 6673 6574 |...{s....offset
00000010  5f73 7465 7069 2000 0000 730e 0000 006c _stepi ...s....l
00000020  656e 5f66 6965 6c64 5f74 7970 6573 0100 en_field_types..
00000030  0000 4973 0a00 0000 7365 7269 616c 697a ..Is....serializ
00000040  6572 730d 0000 0070 7966 6173 746d 6172 ers....pyfastmar
00000050  7368 616c 7310 0000 0063 6f6d 7072 6573 shals....compres
00000060  7369 6f6e 5f74 7970 6573 0400 0000 677a sion_types....gz
00000070  6970 3000 0000 0000 0000 0000 0000 0000 ip0.............
00000080  7c00 0000 ab2e e664 6060 48ce cf2b 49cd |...+.fd``HNO+IM
00000090  2bc9 4c29 9604 f28a 334b 52ad f4f5 cbcb +IL)..r.3KR-tuKK
000000a0  cbf5 92f3 33f2 cb33 f352 8b2a 81cc dc62 Ku.s3rK3sR.*.L\b
000000b0  0ea8 6c51 625e 7631 2f90 63a8 6768 626a .(lQb^v1/.c(ghbj
000000c0  6806 06e6 c5ec 4091 c4bc e48c fca2 e268 h..fEl@.D<d.|"bh
000000d0  4620 5b83 15a4 1ec4 d205 9306 5848 1620 F [..$.DR...XH.
000000e0  0936 0ec2 6503 92a5 4599 a985 c5d1 0c20 .6.Be..%E.).EQ.
000000f0  0100 dc8c 2367 0000 0000 0000 0000 0000 ..\.#g..........

  Addr     0 1  2 3  4 5  6 7  8 9  A B  C D  E F 0 2 4 6 8 A C E
--------  ---- ---- ---- ---- ---- ---- ---- ---- ----------------
00000100  9c00 0000 5d8c cb0a 0231 0c45 2bf8 5cf8 ....].K..1.E+x\x
00000110  156e 5cb5 d332 8ee0 afcc 6aa8 8516 9d54 .n\5S2.`/Lj(...T
00000120  9b0c 45fc 7993 2ecd 229c 73b9 dc2f 1e94 ..E|y..M".s9\/..
00000130  523e 0305 a074 c713 5b24 7add 8ca9 b56a R>.. tG.[$z].)5j
00000140  9f63 ae09 42f9 30ce 0632 041d 697e e29e .c..By0N.2..i~b.
00000150  7b98 2894 091e 7864 b1da f617 3bb4 bbe2 {.(...xd1Zv.;4;b
00000160  8e93 097c cc05 c715 f379 23fd 353f 9940 ...|L.G.sy#}5?.@
00000170  896c 53ab bbfe 5f05 daae 80d3 cee1 9661 .lS+;~_.Z..SNa.a
00000180  2929 bc71 6452 dd0f 6f62 2a89 0000 0000 ))<qdR].ob*.....
00000190  0000 0000 0000 0000 0000 0000 0000 0000 ................
000001a0  9c00 0000 6d8d cb0e 0221 0c45 31f1 b9f0 ....m.K..!.E1q9p
000001b0  3b74 c523 3338 89bf 322b 3292 401c 60a4 ;tE#38.?2+2.@.`$
000001c0  3568 fc79 0b89 3bbb 3af7 e4b6 fdc0 8131 5h|y..;;:wd6}@.1
000001d0  36a5 8836 a2bf c189 9243 5cae 4294 52f8 6%.6"?A..C\.B.Rx
000001e0  945c 2a3e dafc 260c c2be 4c58 66db 7387 .\*>Z|&.B>LXf[s.
000001f0  6186 3d75 c1a3 cd26 dee1 4841 71d5 6b75 a.=uA#M&^aHAqUku

  Addr     0 1  2 3  4 5  6 7  8 9  A B  C D  E F 0 2 4 6 8 A C E
--------  ---- ---- ---- ---- ---- ---- ---- ---- ----------------
00000200  6933 c08e 8c89 934b 19c6 15f1 7953 fb75 i3@....K.F.qyS{u
00000210  e977 06aa ee5a 5172 a9a4 fa67 d6c4 edc3 iw.*nZQr)$zgVDmC
00000220  b649 3568 ddf0 99bd 7dc0 c8aa fd02 d3f4 6I5h]p.=}@H*}.St
00000230  2f0b 0000 0000 0000 0000 0000 0000 0000 /...............
00000240  9c00 0000 6d8c cd0e c220 1084 31a9 7f07 ....m.M.B ..1)..
00000250  9f43 4f50 9262 135f a527 8224 100b b4ec .COP.b._%'.$..4l
00000260  1a34 bebc 0b89 37f7 b099 99fd 763e 7064 .4><..7w0..}v>pd
00000270  8c99 14d1 46f4 7738 9373 88cb 4d88 520a ...QFtw8.s.KM.R.
00000280  37c9 a5e2 a3cd 6f92 41d8 970e cb6c 1577 7I%b#Mo.AX..Kl.w
00000290  1866 3810 0b1e 6dd6 f101 2732 92cb 41c9 .f8...mVq.'2.KAI
000002a0  6b9b 11f6 94e8 685c ca30 6d48 5fb6 95af k..v.hh\J0mH_6./
000002b0  4fbf 1aa8 b184 1ded 9ecb 51fd 093a 92ad O?.(1..m.KQ}.:.-
000002c0  bf6b fdfd d08e cfec ed0a 13ab dc17 1ca4 ?k}}P.Olm..+\..$
000002d0  2e50 0000 0000 0000 0000 0000 0000 0000 .P..............
```

3.4.2 idx

Example for the file format described in section 2.5.2 is as follows.

```
  Addr     0 1  2 3  4 5  6 7  8 9  A B  C D  E F 0 2 4 6 8 A C E
--------  ---- ---- ---- ---- ---- ---- ---- ---- ----------------
00000000  513d 8e0d 9a4a cf10 38b9 1646 27a3 5167 Q=...JO.89.F'#Qg
```

3.4.3 idx.ofs

Example for the file format described in section 2.5.3 is as follows.

```
  Addr     0 1  2 3  4 5  6 7  8 9  A B  C D  E F 0 2 4 6 8 A C E
--------  ---- ---- ---- ---- ---- ---- ---- ---- ----------------
00000000  0000 0000 0400 0000 0900 0000 0e00 0000 ................
```

The value in the first 4 bytes, 0x000000, specifies the offset to the first record entry in the example in section 3.4.1. Adding the size of the file header, 124 bytes, and the record size, 4 bytes, results in an offset 128 bytes from the beginning of the file. The following 4 bytes, 0x04000000, is an offset of 4. This value is multiplied by 32, which results in an offset of 128. The length of the header, 128 bytes, is added to this offset. Thus, the second record begins at 256 bytes from the beginning of the file.

3.5 Index Update Files

3.5.1 pupdateuris_by_uri

Example for the file format described in section 2.6.2 is as follows.

```
http://www.cohowinery.com/example1.html
http://www.cohowinery.com/example2.html
http://www.cohowinery.com/example3.html
http://www.cohowinery.com/example4.html
http://www.cohowinery.com/example5.html
```

4 Security Considerations

None.

5 Appendix A: Product Behavior

The information in this specification is applicable to the following Microsoft products or supplemental software. References to product versions include released service packs:

- Microsoft® FAST™ Search Server 2010

Exceptions, if any, are noted below. If a service pack or Quick Fix Engineering (QFE) number appears with the product version, behavior changed in that service pack or QFE. The new behavior also applies to subsequent service packs of the product unless otherwise specified. If a product edition appears with the product version, behavior is different in that product edition.

Unless otherwise specified, any statement of optional behavior in this specification that is prescribed using the terms SHOULD or SHOULD NOT implies product behavior in accordance with the SHOULD or SHOULD NOT prescription. Unless otherwise specified, the term MAY implies that the product does not follow the prescription.

6 Change Tracking

No table of changes is available. The document is either new or has had no changes since its last release.

7 Index

G

I

L

M

N

O

P

R

S

T

U

V