WebAnalyzer Data File Formats

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 Open Specification Promise or the Community Promise. 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.
<table>
<thead>
<tr>
<th>Date</th>
<th>Revision History</th>
<th>Revision Class</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/13/2009</td>
<td>0.1</td>
<td>Major</td>
<td>Initial Availability</td>
</tr>
<tr>
<td>02/19/2010</td>
<td>1.0</td>
<td>Editorial</td>
<td>Revised and edited the technical content</td>
</tr>
<tr>
<td>03/31/2010</td>
<td>1.01</td>
<td>Editorial</td>
<td>Revised and edited the technical content</td>
</tr>
<tr>
<td>04/30/2010</td>
<td>1.02</td>
<td>Editorial</td>
<td>Revised and edited the technical content</td>
</tr>
<tr>
<td>06/07/2010</td>
<td>1.03</td>
<td>Editorial</td>
<td>Revised and edited the technical content</td>
</tr>
<tr>
<td>06/29/2010</td>
<td>1.04</td>
<td>Editorial</td>
<td>Changed language and formatting in the technical content.</td>
</tr>
<tr>
<td>07/23/2010</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>09/27/2010</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>11/15/2010</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>12/17/2010</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>03/18/2011</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>06/10/2011</td>
<td>1.04</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>01/20/2012</td>
<td>1.5</td>
<td>Minor</td>
<td>Clarified the meaning of the technical content.</td>
</tr>
<tr>
<td>04/11/2012</td>
<td>1.5</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
<tr>
<td>07/16/2012</td>
<td>1.5</td>
<td>No change</td>
<td>No changes to the meaning, language, or formatting of the technical content.</td>
</tr>
</tbody>
</table>
# Table of Contents

1 **Introduction** ..........................................................5
   1.1 Glossary ..................................................................5
   1.2 References ..................................................................5
      1.2.1 Normative References ........................................5
      1.2.2 Informative References ......................................6
   1.3 Structure Overview (Synopsis) ....................................6
   1.4 Relationship to Protocols and Other Structures ..............6
   1.5 Applicability Statement ............................................6
   1.6 Versioning and Localization ........................................6
   1.7 Vendor-Extensible Fields ...........................................6

2 **Structures** .....................................................................7
   2.1 Common file structures ..............................................7
   2.2 Input Files ...................................................................9
      2.2.1 delete ..............................................................9
      2.2.2 eqrepr ............................................................9
      2.2.3 links ...............................................................9
      2.2.4 no_links ..........................................................10
      2.2.5 sitemap ...........................................................10
      2.2.6 urieq ...............................................................10
      2.2.7 urimap .............................................................11
   2.3 Initial Processing Files ..............................................11
      2.3.1 links_by_to ......................................................11
      2.3.2 links_by_to_raw ................................................12
      2.3.3 urieq_by_class ................................................12
      2.3.4 eqrepr_by_uri ..................................................12
      2.3.5 urihash ..........................................................13
   2.4 Main Processing Files ...............................................13
      2.4.1 rank_links_by_src ............................................13
      2.4.2 rank_by_uri .....................................................13
      2.4.3 linkscore_by_dst ..............................................14
      2.4.4 links_norm_with_fromrank_by_anchor ....................14
      2.4.5 anchor_freqs_by_anchor .....................................14
      2.4.6 links_with_freqs_by_to .....................................15
      2.4.7 uri_anchors_by_urihash .....................................15
      2.4.8 anchor_by_to ...................................................16
      2.4.9 rank_by_site ....................................................16
      2.4.10 siterank_by_uri ...............................................17
      2.4.11 anchor_by_uri ................................................17
      2.4.12 anchor_by_uri_with_repr ...................................18
      2.4.13 anchor_info_new .............................................18
   2.5 Database Files .......................................................18
      2.5.1 bin ..................................................................19
      2.5.2 idx ..................................................................19
      2.5.3 idx.ofs ................................................................19
   2.6 Index Update Files ...................................................19
      2.6.1 feeduris ..........................................................19
      2.6.2 pupdateuris_by_uri ..........................................20

3 **Structure Examples** ..................................................21
1 Introduction

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

Sections 1.7 and 2 of this specification are normative and can contain the terms MAY, SHOULD, MUST, MUST NOT, and SHOULD NOT as defined in RFC 2119. All other sections and examples in this specification are informative.

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
- base64 encoding
- 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

References to Microsoft Open Specifications documentation do not include a publishing year because links are to the latest version of the technical documents, which are updated frequently. References to other documents include a publishing year when one is available.

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".
1.2.2 Informative References


[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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANCHORTEXT</td>
<td>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.</td>
</tr>
<tr>
<td>BASE64</td>
<td>MUST be a sequence of bytes in base64 encoding.</td>
</tr>
<tr>
<td>CLASS</td>
<td>The main item of an equivalence class.</td>
</tr>
<tr>
<td>EQREPR</td>
<td>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.</td>
</tr>
<tr>
<td>INTRA</td>
<td>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.</td>
</tr>
<tr>
<td>MEMBER</td>
<td>A member of an equivalence class.</td>
</tr>
<tr>
<td>RANK</td>
<td>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.</td>
</tr>
<tr>
<td>SITE</td>
<td>The site(1) or site(2) of an item. If a site is not available, the value MUST be the two hexadecimal bytes 0xc782.</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>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.</td>
</tr>
<tr>
<td>TOKEN</td>
<td>Token encoded in UTF-8 and normalized as specified in [MS-FSIN].</td>
</tr>
<tr>
<td>URL</td>
<td>The URL of an item. The complete ABNF rule set for URLs is specified in [RFC3986].</td>
</tr>
<tr>
<td>URLHASH</td>
<td>Represents the document identifier(3) of an item. This MUST contain the MD5 hash value of the URL of the item in base64 encoding. The length MUST be 21 characters.</td>
</tr>
</tbody>
</table>

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-FSFDNW], 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLHASH</td>
<td>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].</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>The time that the delete operation was submitted.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLHASH</td>
<td>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].</td>
</tr>
<tr>
<td>URL</td>
<td>The item that owns the equivalence class, represented by an URL.</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>The time that the equivalence class was submitted to the system.</td>
</tr>
<tr>
<td>EQREPR</td>
<td>The equivalence class of the item.</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>The source URL of the hyperlink, represented as a hash value.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination URL of the hyperlink, represented as a hash value.</td>
</tr>
<tr>
<td>INTRA</td>
<td>Specified in section 2.1.</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>The time that the link was submitted to the system.</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text of the hyperlink.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>The item that has no hyperlinks.</td>
</tr>
<tr>
<td>URLHASH</td>
<td>The hash value of the item that has no hyperlinks.</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>The time that the item was submitted to the system.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLHASH</td>
<td>The hash value of an item.</td>
</tr>
<tr>
<td>SITE</td>
<td>The site(2) of the item.</td>
</tr>
</tbody>
</table>

### 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.

\[ \text{ROW} = \text{CLASS SP MEMBER SP TIMESTAMP} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS</td>
<td>Represents an item. This item identifies the equivalence class.</td>
</tr>
<tr>
<td>MEMBER</td>
<td>A member of the equivalence class. Can be equal to the CLASS column.</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>The time that the mapping was submitted to the system.</td>
</tr>
</tbody>
</table>

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.

\[ \text{ROW} = \text{URL SP URLHASH} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>The URL of an item.</td>
</tr>
<tr>
<td>URLHASH</td>
<td>The hash value of the URL.</td>
</tr>
</tbody>
</table>

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.

\[ \text{ROW} = \text{FROM SP TO SP ANCHORTEXT} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>The source URL of the hyperlink.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination URL of the hyperlink.</td>
</tr>
<tr>
<td>Column name</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text of the hyperlink.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>The source URL of the hyperlink.</td>
</tr>
<tr>
<td>SITE</td>
<td>The site of the source URL in the FROM column.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination URL of the hyperlink.</td>
</tr>
<tr>
<td>INTRA</td>
<td>Specified in section 2.1.</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text of the hyperlink.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>The URL of the item.</td>
</tr>
<tr>
<td>EQREPR</td>
<td>The equivalence class of the item.</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLHASH</td>
<td>The hash value of the URL of an item.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>The source of the hyperlink.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination of the hyperlink.</td>
</tr>
<tr>
<td>COUNT</td>
<td>The number of hyperlinks from the source.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLHASH</td>
<td>The hash value of the URL of an item.</td>
</tr>
<tr>
<td>RANK</td>
<td>The quality score specified for an item.</td>
</tr>
</tbody>
</table>
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.

\[ \text{ROW} = \text{FROM} \text{ SP } \text{TO} \text{ SP } \text{RANK} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>The source of the hyperlink.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination of the hyperlink.</td>
</tr>
<tr>
<td>RANK</td>
<td>The quality score associated with the destination item.</td>
</tr>
</tbody>
</table>

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.

\[ \text{ROW} = \text{RANK} \text{ SP } \text{FROM} \text{ SP } \text{TO} \text{ SP } \text{ANCHORTEXT} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANK</td>
<td>The quality score associated with the source item.</td>
</tr>
<tr>
<td>FROM</td>
<td>The source of the hyperlink.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination of the hyperlink.</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text of the hyperlink.</td>
</tr>
</tbody>
</table>

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.

\[ \text{ROW} = \text{AFREQ} \text{ SP } \text{ARANK} \text{ SP } \text{ANCHORTEXT} \]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFREQ</td>
<td>The frequency count of the anchor text in hyperlinks in the system.</td>
</tr>
<tr>
<td>ARANK</td>
<td>The calculated quality score of the anchor text in the system.</td>
</tr>
</tbody>
</table>
## 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANK</td>
<td>The quality score of the source of the hyperlink.</td>
</tr>
<tr>
<td>AFREQ</td>
<td>The frequency count of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>ARANK</td>
<td>The calculated quality score of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>TO</td>
<td>The destination of the hyperlink.</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text from the hyperlink.</td>
</tr>
</tbody>
</table>

## 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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TO-RANK</td>
<td>The quality score of the destination of the hyperlinks.</td>
</tr>
<tr>
<td>LAFREQ</td>
<td>The frequency count of the anchor text across the hyperlinks pointing to the destination, the TO column.</td>
</tr>
<tr>
<td>LARANK</td>
<td>The calculated quality score of the anchor text across all hyperlinks that point to the destination, the TO column.</td>
</tr>
<tr>
<td>AFREQ</td>
<td>The frequency count of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>ARANK</td>
<td>The calculated quality score of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>TO</td>
<td>The common destination of the hyperlinks.</td>
</tr>
</tbody>
</table>
### Column name | Description
--- | ---
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.

\[
\text{ROW} = \text{TO-RANK SP LAFREQ SP LARANK SP AFREQ SP ARANK SP URL SP SITE SP ANCHORTEXT}
\]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TO-RANK</td>
<td>The quality score of the destination of the hyperlinks.</td>
</tr>
<tr>
<td>LAFREQ</td>
<td>The frequency count of the anchor text across all hyperlinks that point to the destination, the URL column.</td>
</tr>
<tr>
<td>LARANK</td>
<td>The calculated quality score of the anchor text across all hyperlinks that point to the destination, the URL column.</td>
</tr>
<tr>
<td>AFREQ</td>
<td>The frequency count of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>ARANK</td>
<td>The calculated quality score of the anchor text across all hyperlinks in the system.</td>
</tr>
<tr>
<td>URL</td>
<td>The common destination of the hyperlinks.</td>
</tr>
<tr>
<td>SITE</td>
<td>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.</td>
</tr>
<tr>
<td>ANCHORTEXT</td>
<td>The anchor text from the hyperlinks.</td>
</tr>
</tbody>
</table>

#### 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.

\[
\text{ROW} = \text{SITE SP RANK SP URL}
\]

The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE</td>
<td>The site part of the URL of the item.</td>
</tr>
<tr>
<td>RANK</td>
<td>The calculated quality score for the item.</td>
</tr>
<tr>
<td>Column name</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>URL</td>
<td>The item.</td>
</tr>
</tbody>
</table>

### 2.4.10 siterrank_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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE-OR-TO-URL</td>
<td>A site, or the URL of an item.</td>
</tr>
<tr>
<td>SITE-RANK</td>
<td>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.</td>
</tr>
</tbody>
</table>

### 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.

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

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQREPR</td>
<td>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.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE-OR-TO-URL-HASH</td>
<td>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.</td>
</tr>
<tr>
<td>ANCHORINFO</td>
<td>This column contains a dictionary, composed of key-value pairs. The dictionary MUST contain the keys &quot;anchors&quot;, &quot;queries&quot;, &quot;contentid&quot;, &quot;rank&quot;, &quot;siterank&quot;, and &quot;urieqs&quot;. The key fields contain values as specified in [MS-FSFDMW]. The dictionary is serialized as specified in [MS-FSWCU], and then encoded using base64 encoding.</td>
</tr>
</tbody>
</table>

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 a dictionary described in the following table.

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>offset_step</td>
<td>32</td>
</tr>
<tr>
<td>len_field_type</td>
<td>1</td>
</tr>
<tr>
<td>serializer</td>
<td>pyfastmarshal</td>
</tr>
<tr>
<td>compression_type</td>
<td>gzip</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>The URL of an item to update in the index.</td>
</tr>
<tr>
<td>COLLECTION</td>
<td>The content collection of the item.</td>
</tr>
</tbody>
</table>

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-FSFDWM]. The following table specifies the columns.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>The URL of an item to update in the index.</td>
</tr>
</tbody>
</table>
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/Wx8KQiui3h 0 1261403658 example4
UnX+GSPfPgCh/QgP65+pF 5ClAoACOAY2/Wx8KQiui3h 0 1261403672 example4
ccfq2he9YK11h9szWQ0i1 5ClAoACOAY2/Wx8KQiui3h 0 1261403681 example4
hjLsdjTaTfuvbOf6kFimU E9O5O8O1DdH18l81d8Vak 0 1261403690 example5
5Roj9xC51ItjtuLA5FzkH e+t0ML6KvvqputJZXm3V 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 mt94KfEMAAzxJi1BvzVx
http://www.cohowinery.com/example3.html k8EyJ1bIdfG9y/mM65YDZ
http://www.cohowinery.com/example4.html Rha5OAU0xVUB5HuG5EvL
http://www.cohowinery.com/example5.html Z1GjJ824myhwBL0F+q2HA
http://www.cohowinery.com/none.html EM9KmJKeqTHYQsDC6a485
```

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 EM9KmJKeqTHYQsDC6a485 none
4nv9pjGzJ9r8bUn0ethTL Rha5OAU0xVUB5HuG5EvL example4
k8EyJ1bIdfG9y/mM65YDZ Rha5OAU0xVUB5HuG5EvL example4
mt94KfEMAAzxJi1BvzVx Rha5OAU0xVUB5HuG5EvL example4
Rha5OAU0xVUB5HuG5EvL 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 Rha5OAU0xVUB5HuG5EvL 1
Rha5OAU0xVUB5HuG5EvL Z1GjJ824myhwBL0F+q2HA 1
Z1GjJ824myhwBL0F+q2HA EM9KmJKeqTHYQsDC6a485 1
k8EyJ1bIdfG9y/mM65YDZ Rha5OAU0xVUB5HuG5EvL 1
mt94KfEMAAzxJi1BvzVx Rha5OAU0xVUB5HuG5EvL 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 Rha5OAU0XwYUB5HuG5EvL example4
1.04 1 0.1755 1 0.1755 Z1GjJ8B4myhWL0F+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 example4
1.04 1 0.1755 1 0.1755 http://www.cohowinery.com/example5.html example5
2.22 1 1.04 1 1.04 http://www.cohowinery.com/none.html none

The following is the same information shown in hexadecimal format.

00000000 302e 3137 3535 2033 2030 3a30 3130 3134 0.1755 3 0.01014
00000010 2033 2030 3a30 3130 3134 2068 7474 703a 3 0.01014 http:
00000020 2f2f 7777 772e 636f 686f 7769 6e65 7279 //www.cohowinery
00000030 2e63 6f6d 2f6f 6666 6666 6666 6666 6666 .com/example4.ht
00000040 6d6c 20c7 2065 78 616d 706c 6535 206e 322e 3232 ml /example5
00000050 312e 3034 2031 2030 2e31 3735 3120 1.04 1 1.04 http://www.cohowinery
00000060 7765 706c 734e 6f6e 652e 6874 ml none.html
00000070 772e 636f 686f 7779 6e65 7279 2e63 6f6d 322e 3232 .com/example5
00000080 2f65 6c65 206d 6f6e 652e 6874 6d6c 2035 6f6e 652e 6874 /example5
00000090 6774 703a 2f2f 7777 772e 636f 686f 7769 7279 2e63 6f6d /www.cohowinery
000000a0 2031 2030 2e30 34 2031 2031 2e30 34 2068 7474 703a 1 0.0 1 1.04 http:
000000b0 2f2f 7777 772e 636f 686f 7777 772e 636f 686f 7769 ttp://www.cohowinery
000000c0 2e6e 652e 6874 2e6d 6f6e 652e 6874 nery.com
000000d0 6d6c 20c7 6e6f 6e65 ml none

3.3.5 anchor_by_uri_with_repr

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

1.4516666667 0.1755 3 0.01014 3 0.01014 http://www.cohowinery.com/example4.html example4
1.4516666667 1.04 1 0.1755 1 0.1755 http://www.cohowinery.com/example5.html example5
1.4516666667 2.22 1 1.04 1 1.04 http://www.cohowinery.com/none.html none
1.4516666667 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 3720 . 1.1451666667
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  0020 312e 3034 2068 7474 703a 2f2f 7777 1.04 http://www.co
00000060  772e 636f 686f 7769 6e65 7279 2e63 6f6d w.cohowinery.co
00000070  2f6e 6f6e 652e 6874 6d6c 206e 6f6e 650a /none.html none.
00000080  0020 312e 3134 3531 3636 3636 3720 312e 1.1451666667 1.
00000090  3636 3720 322e 3232 2031 2031 2e30 3420 667 2.22 1 1.04
000000a0  3120 312e 3034 2068 7474 703a 2f2f 7777 1 1.04 http://ww
000000b0  772e 636f 686f 7769 6e65 7279 2e63 6f6d .cohowinery.co
000000c0  2f6e 6f6e 652e 6874 6d6c 206e 6f6e 650a /none.html none.
000000d0  0020 312e 3134 3531 3636 3636 3720 . 1.1451666667
000000e0  2031 2030 2e31 3735 3520 3120 302e 317 1 0.1755 1 0.17
000000f0  3535 2068 7474 703a 2f2f 7777 772e 6368 http://www.co
00000100  6f7769 6e65 7279 2e63 6f6d 2f65 7861 howinery.com/example
00000110  6d70 6c65 352e 6874 6d6c 2065 7861 6d70 5.html example
00000120  6l65 350a 0020 312e 3134 3531 3636 65le5.. 1.14516666
00000130  3636 3720 322e 3232 2031 2031 2e30 3420 667 2.22 1 1.04
00000140  3120 312e 3034 2068 7474 703a 2f2f 7777 1 1.04 http://ww
00000150  772e 636f 686f 7769 6e65 7279 2e63 6f6d .cohowinery.co
00000160  2f6e 6f6e 652e 6874 6d6c 206e 6f6e 650a /none.html none.

3.3.6 anchor_info_new

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

931639469199963915837568340241444652
e3M3AAAAY29udGVudGlkcycAAABodHRwOi8vd3d3LmNvaG93aW51cnkuY29tL2V4YW1wbGU0Lmh0bWxzCAAAAH
NpdGVYYW5rcwQAAAkJEOuNTET2NJy2NYj3cwQAAAAbhmOb3JzWVEAAAOQAAAHMIAAAAZhhbXBsZTZhRzAQAAADzBwA
AADuMDEwMTRzAQAAADNzBwAAADAuMDEwMTRzBAAAAAHJhbmtzAQAAADAuMTc1NXMGAAAAdXJpZXFzZwAAAAw

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
--------  ----- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ----

23 / 31

[MS-FSWADF] — v20120630
WebAnalyzer Data Files Formats

Copyright © 2012 Microsoft Corporation.

Release: July 16, 2012
Example for the file format described in section 2.5.2 is as follows.
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
-------- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- -----
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

A
anchor_by_to 16
example 22
anchor_by_uri 17
anchor_by_uri_with_repr 18
example 22
anchor_freqs_by_anchor 14
example 22
anchor_info_new 18
example 23
Applicability 6

B
bin 19
example 23

C
Change tracking 28
Common data types and fields (section 2 7, section 2 7)
Common file structures 7

D
Data types and fields
common 7
Data types and fields - common 7
Database files 18
bin 19
feeduris 19
idx 19
idx.ofs 19
pupdateuris_by_uri 20
Delete 9
Details
anchor_by_to 16
anchor_by_uri 17
anchor_by_uri_with_repr 18
anchor_freqs_by_anchor 14
anchor_info_new 18
bin 19
common data types and fields (section 2 7, section 2 7)
common file structures 7
database files 18
delete file 9
eqrepr file 9
eqrepr_by_uri file 12
feeduris 19
idx 19
idx.ofs 19
input files 9
links file 9
links_by_to file 11
links_by_to_raw file 12
links_norm_with_fromrank_by_anchor 14
link_freqs_by_to file 15
linkscore_by_dst file 14
main_processing_files 13
no_links file 10
pupdateuris_by_uri 20
rank_by_site 16
rank_by_uri file 13
rank_links_by_src file 13
sitemap file 10
siterank_by_uri 17
uri_anchors_by_urihashfile 15
urieq file 10
urieq_by_class_file 12
urihash file 13
urimap file 11

E
Eqrepr 9
eqrepr_by_uri 12
Examples 21
database file
bin 23
idx 24
idx.ofs 25
index update file
pupdateuris_by_uri 25
initial processing file
anchor_by_to 22
anchor_by_uri_with_repr 22
anchor_freqs_by_anchor 22
anchor_info_new 22
links_by_to 21
rank_links_by_src 21
uri_anchors_by_urihash 22
input file
links 21
urimap 21

F
feeduris 19
Fields - vendor-extensible 6
Files
anchor_by_to 16
anchor_by_uri 17
anchor_by_uri_with_repr 18
anchor_freqs_by_anchor 14
anchor_info_new 18
bin 19
database 18
delete 9
eqrepr 9
eqrepr_by_uri 12
feeduris 19
idx 19
idx.ofs 19
input 9
links 9
WebAnalyzer Data File Formats

Copyright © 2012 Microsoft Corporation.

Release: July 16, 2012
Vendor-extensible fields 6
Versioning 6