INTERNATIONAL STANDARD

ISO 28500

First edition
2009-05-15

Information and documentation — WARC file format

Information et documentation — Format de fichier WARC

Reference number
ISO 28500:2009(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 28500 was prepared by Technical Committee ISO/TC 46, *Information and documentation*, Subcommittee SC 4, *Technical interoperability*.

Introduction

Websites and web pages emerge and disappear from the World Wide Web every day. For the past ten years, memory storage organizations have tried to find the most appropriate ways to collect and keep track of this vast quantity of important material using web-scale tools such as web crawlers. A web crawler is a program that browses the web in an automated manner according to a set of policies; starting with a list of URLs, it saves each page identified by a URL, finds all the hyperlinks in the page (e.g. links to other pages, images, videos, scripting or style instructions, etc.), and adds them to the list of URLs to visit recursively. Storing and managing the billions of saved web page objects itself presents a challenge.

At the same time, those same organizations have a rising need to archive large numbers of digital files not necessarily captured from the web (e.g. entire series of electronic journals, or data generated by environmental sensing equipment). A general requirement that appears to be emerging is for a container format that permits one file simply and safely to carry a very large number of constituent data objects for the purpose of storage, management, and exchange. Those data objects (or resources) need to be of unrestricted type (including many binary types for audio, CAD, compressed files, etc.), but fortunately the container needs only minimal knowledge of the nature of the objects.

The WARC (Web ARChive) file format offers a convention for concatenating multiple resource records (data objects), each consisting of a set of simple text headers and an arbitrary data block into one long file. The WARC format is an extension of the ARC file format (ARC) that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web. Each capture in an ARC file is preceded by a one-line header that very briefly describes the harvested content and its length. This is directly followed by the retrieval protocol response messages and content. The original ARC format file has been used by the Internet Archive (IA) since 1996 for managing billions of objects, and by several national libraries.

The motivation to extend the ARC format arose from the discussion and experiences of the International Internet Preservation Consortium (IIPC), whose members include the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA), and the Internet Archive (IA). The California Digital Library and the Los Alamos National Laboratory also provided input on extending and generalizing the format.

The WARC format is expected to be a standard way to structure, manage and store billions of resources collected from the web and elsewhere. It will be used to build applications for harvesting (such as the open source Heritrix web crawler), managing, accessing, and exchanging content. The way WARC files will be created and resources stored and rendered will depend on software and applications implementations.

Besides the primary content recorded in ARCs, the extended WARC format accommodates related secondary content, such as assigned metadata, abbreviated duplicate detection events, later-date transformations, and segmentation of large resources. The extension may also be useful for more general applications than web archiving. To aid the development of tools that are backwards compatible, WARC content is clearly distinguishable from pre-revision ARC content.

The WARC file format is made sufficiently different from the legacy ARC format files so that software tools can unambiguously detect and correctly process both WARC and ARC records; given the large amount of existing archival data in the previous ARC format, it is important that access and use of this legacy not be interrupted when transitioning to the WARC format.

After the Internet Engineering Steering Group (IESG: http://www.ietf.org/iesg.html) approval, IANA (Internet Assigned Numbers Authority: http://www.iana.org/) is expected to register the WARC type "application/warc" using the application provided in this International Standard and following procedures defined in [RFC2048].

Information and documentation — WARC file format

1 Scope

This International Standard specifies the WARC file format:

— to store both the payload content and control information from mainstream Internet application layer protocols, such as the HTTP, DNS, and FTP;

— to store arbitrary metadata linked to other stored data (e.g. subject classifier, discovered language, encoding);

— to support data compression and maintain data record integrity;

— to store all control information from the harvesting protocol (e.g. request headers), not just response information;

— to store the results of data transformations linked to other stored data;

— to store a duplicate detection event linked to other stored data (to reduce storage in the presence of identical or substantially similar resources);

— to be extended without disruption to existing functionality;

— to support handling of overly long records by truncation or segmentation, where desired.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 8601, *Data elements and interchange formats — Information interchange — Representation of dates and times*

[RFC1035] Mockapetris, P. *Domain names — Implementation and specification*. STD 13, November 1987. Available at: http://www.faqs.org/rfcs/rfc1035.html

[RFC1884] Hinden, R. and Deering, S. *IP Version 6 Addressing Architecture*. December 1995. Available at: http://www.faqs.org/rfcs/rfc1884.html

[RFC2045] Freed, N. and Borenstein, N. *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*. November 1996. Available at: http://www.faqs.org/rfcs/rfc2045

[RFC2540] Eastlake, D. *Detached Domain Name System (DNS) Information*. March 1999. Available at: http://www.faqs.org/rfcs/rfc2540.html

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T. *Hypertext Transfer Protocol — HTTP/1.1*. June 1999 (TXT, PS, PDF, HTML, XML). Available at: http://www.faqs.org/rfcs/rfc2616.html

[RFC2822] Resnick, P. (ed.) *Internet Message Format*. April 2001. Available at: http://www.faqs.org/rfcs/rfc2822

[RFC3629] Yergeau, F. *UTF-8, a transformation format of ISO 10646*. STD 63, November 2003. Available at: http://www.faqs.org/rfcs/rfc3629.html

[RFC3986] Berners-Lee, T., Fielding, R., Masinter, L. *Uniform Resource Identifier (URI): Generic Syntax*. STD 66, January 2005 (TXT, HTML, XML). Available at: http://www.faqs.org/rfcs/rfc3986.html

[RFC4027] Josefsson, S. *Domain Name System Media Types*. April 2005. Available at: http://www.faqs.org/rfcs/rfc4027.html

[W3CDTF] *Date and Time Formats: note submitted to the W3C.* 15 September 1997 (W3C profile of ISO 8601). Available at: http://www.w3.org/TR/NOTE-datetime

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1.1
WARC record
basic constituent of a WARC file, consisting of a sequence of WARC records

3.1.2
WARC record content block
part (zero or more octets) of a WARC record that follows the header and that forms the main body of a WARC record

3.1.3
WARC record payload
data object referred to, or contained by a WARC record as a meaningful subset of the content block

3.1.4
WARC record header
beginning of a WARC record, consisting of one first line declaring the record to be in the WARC format with a given version number, followed by lines of named fields up to a blank line

3.1.5
WARC named fields
set of elements consisting of a name, a colon, and a value, with long values continued on indented lines

3.1.6
WARC logical record
in the context of segmentation, a logical record may be composed of multiple segments, each represented by a WARC record

3.2 Abbreviated terms

ABNF augmented Backus-Naur form

ARC archive

CRLF carriage return line feed

DNS	domain name system
FTP	file transfer protocol
HTTP	hypertext transport protocol
IANA	Internet Assigned Numbers Authority
IESG	Internet Engineering Steering Group
RFC	request for comments
UR (I/L/N)	uniform resource (identifier/locator/name)
WARC	web archive

4 File and record model

A WARC format file is the simple concatenation of one or more WARC records. The first record usually describes the records to follow. In general, record content is either the direct result of a retrieval attempt (web pages, inline images, URL redirection information, DNS hostname lookup results, stand-alone files, etc.) or is synthesized material (e.g. metadata, transformed content) that provides additional information about archived content.

A WARC record shall consist of a record header followed by a record content block and two new lines. The WARC record header shall consist of one first line declaring the record to be in the WARC format with a given version number, then a variable number of line-oriented named fields terminated by a blank line. The WARC record header format shall follow the general rules of HTTP/1.1 [RFC2616] and [RFC2822] headers with one major exception: it shall also allow UTF-8 characters, as specified in [RFC3629].

The top-level view of a WARC file can be expressed in an ABNF grammar, reusing the augmented constructs defined in section 2.1 of HTTP/1.1 [RFC2616]. (In particular, note that to avoid the risk of confusion, where any WARC rule has the same name as an [RFC2616] rule, the definition here has been made the same, except in the case of the CHAR rule, which in WARC includes multibyte UTF-8 characters.)

```
warc-file     = 1*warc-record
warc-record   = header CRLF
                block CRLF CRLF
header        = version warc-fields
version       = "WARC/1.0" CRLF
warc-fields   = *named-field CRLF
block         = *OCTET
```

The record version shall appear first in every record and hence shall also begin the WARC file itself.

The WARC record relies heavily on named fields. Each named field consists of a name followed by a colon (":") and the field value. Field names are not case-sensitive. The field value may be preceded by any amount of linear white space (LWS), though a single space is preferred. Header fields can be extended over multiple lines by preceding each extra line with at least one space or tab character.

Named fields may appear in any order and field values may contain any UTF-8 character. Both defined-fields and extension-fields follow the generic named-field format. Extension-fields may be used in extensions of the core format.

```
named-field     = field-name ":" [ field-value ]
field-name      = token
field-value     = *( field-content | LWS )      ; further qualified
                                                ; by field
                                                ; definitions
field-content   = <the OCTETs making up the field-value
                  and consisting of either *TEXT or combinations
                  of token, separators, and quoted-string>
OCTET           = <any 8-bit sequence of data>
token           = 1*<any US-ASCII character>
                  except CTLs or separators>
separators      = "(" | ")" | "<" | ">" | "@"
                     | "," | ";" | ":" | "\" | <">
                     | "/" | "[" | "]" | "?" | "="
                     | "{" | "}" | SP | HT
TEXT            = <any OCTET except CTLs,
                  but including LWS>
CHAR            = <UTF-8 characters; RFC3629>  ; (0-191, 194-244)
DIGIT           = <any US-ASCII digit "0".."9">
CTL             = <any US-ASCII control character
                  (octets 0 - 31) and DEL (127)>
CR              = <ASCII CR, carriage return>  ; (13)
LF              = <ASCII LF, linefeed>         ; (10)
SP              = <ASCII SP, space>            ; (32)
HT              = <ASCII HT, horizontal-tab>   ; (9)
CRLF            = CR LF
LWS             = [CRLF] 1*( SP | HT )         ; semantics same as
                                               ; single SP
quoted-string   = ( <"> *(qdtext | quoted-pair ) <"> )
qdtext          = <any TEXT except <">>
quoted-pair     = "\" CHAR                     ; single-character quoting
uri             = "<" <'URI' per RFC3986> ">"
```


Although UTF-8 characters are allowed, the 'encoded-word' mechanism of [RFC2047] may also be used when writing WARC fields and shall also be understood by WARC reading software.

The rest of the WARC record grammar concerns defined-field parameters such as record identifier, record type, creation time, content length, and content type.

```
defined-field  = WARC-Type
               | WARC-Record-ID
               | WARC-Date
               | Content-Length
               | Content-Type
               | WARC-Concurrent-To
               | WARC-Block-Digest
               | WARC-Payload-Digest
               | WARC-IP-Address
               | WARC-Refers-To
               | WARC-Target-URI
               | WARC-Truncated
               | WARC-Warcinfo-ID
               | WARC-Filename                ; warcinfo only
               | WARC-Profile                 ; revisit only
               | WARC-Identified-Payload-Type
               | WARC-Segment-Origin-ID       ; continuation only
               | WARC-Segment-Number
               | WARC-Segment-Total-Length    ; continuation only
```

Every WARC record shall have a type, reported in the WARC-Type field. Eight WARC record types are defined in this International Standard as follows:

- 'warcinfo',
- 'response',
- 'resource',
- 'request',
- 'metadata',
- 'revisit',
- 'conversion',
- 'continuation'.

Other types of WARC records may be defined in extensions of the core format. The relevant fields for each record type are described in detail in Clause 6. Each field's meaning and legal value format are described in Clause 5.

The record block shall contain octet content, interpreted based on the record type and other header values. All records shall include a Content-Length field to specify the length of the block.

Some record types (and possibly future record types) also define a payload, such as a meaningful subset of the block or content from a predecessor record. Some headers pertain to the payload of a record rather than the block directly.

For example, in a 'response' record with a content block consisting of HTTP headers and a data object, the payload would be the data object. All 'response', 'resource', 'request', 'conversion' and 'continuation' records may have a payload. All 'warcinfo', 'metadata' and 'revisit' records shall not have a payload.

Content matching the warc-file rule shall have the MIME content-type "application/warc", as specified in 8.2.

Content matching only the warc-fields rule is useful as a simple descriptive format, and has MIME content-type "application/warc-fields", as specified in 8.3.

5 Named fields

5.1 General

Named fields within a WARC record provide information about the current record. WARC both reuses appropriate headers from other standards and defines new headers, all beginning "WARC-", for WARC-specific purposes.

WARC named fields of the same type shall not be repeated in the same WARC record (for example, a WARC record shall not have several WARC-Date or several WARC-Target-URI), except as noted (e.g. WARC-Concurrent-To).

Because new fields may be defined in extensions to the core WARC format, WARC processing software shall ignore fields with unrecognized names.

5.2 WARC-Record-ID (mandatory)

A WARC-Record-ID is an identifier assigned to the current record that is globally unique for its period of intended use. No identifier scheme is mandated by this specification, but each WARC-Record-ID shall be a legal URI and clearly indicate a documented and registered scheme to which it conforms (e.g. via a URI scheme prefix such as "http:" or "urn:"). Care should be taken to ensure that this value is written with no internal white space.

```
WARC-Record-ID    = "WARC-Record-ID" ":" uri
```

All records shall have a WARC-Record-ID field.

5.3 Content-Length (mandatory)

The Content-Length is the number of octets in the block, similar to [RFC2616]. If no block is present, a value of "0" (zero) shall be used.

```
Content-Length    = "Content-Length" ":" 1*DIGIT
```

All records shall have a Content-Length field.

5.4 WARC-Date (mandatory)

The WARC-Date is a 14-digit UTC time-stamp formatted as YYYY-MM-DDThh:mm:ssZ, and shall conform to the W3C profile of ISO 8601, i.e. [W3CDTF]. The time-stamp shall represent the instant that data capture for record creation began. Multiple records written as part of a single capture event (see 5.7) shall use the same WARC-Date, even though the times of their writing will not be exactly synchronized.


```
WARC-Date    = "WARC-Date" ":" w3c-iso8601
w3c-iso8601 = <YYYY-MM-DDThh:mm:ssZ>
```

All records shall have a WARC-Date field.

See Annex A for examples on usage of WARC-Date fields.

5.5 WARC-Type (mandatory)

WARC-Type is the type of WARC record. Record types defined in this International Standard are:

— 'warcinfo',

— 'response',

— 'resource',

— 'request',

— 'metadata',

— 'revisit',

— 'conversion', and

— 'continuation'.