IS 16214 : 2014
ISO 28500 : 2009

Indian Standard

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