AS ISO 28500:2018
ISO 28500:2017

AS ISO 28500:2018

Australian STANDARD

Information and documentation– WARC file format

STANDARDS Australia

AS ISO 28500:2018

This Australian Standard ® was prepared by IT-019, Information and Documentation, Information Technology - Learning, Education, Training and Research. It was approved on behalf of the Council of Standards Australia on 24 January 2018.

This Standard was published on 22 February 2018.

The following are represented on Committee IT-019:

Australian Computer Society
Australian Library and Information Association
Charles Darwin University
CSIRO
Institute for Metadata Management
Northern Territory Library
NSW Department of Education
University of Southern Queensland

This Standard was issued in draft form for comment as DR AS ISO 28500:2017.

Keeping Standards up-to-date

Ensure you have the latest versions of our publications and keep up-to-date about Amendments, Rulings, Withdrawals, and new projects by visiting:
www.standards.org.au

www.saiglobal.com (sales and distribution)

ISBN 978 1 76072 002 5

Australian STANDARD

Information and documentation—WARC file format

First published as AS ISO 28500:2018.

Published by SAI Global Limited under licence from Standards Australia Limited, GPO Box 476, Sydney, NSW 2001, Australia.

Preface

This Standard was prepared by the Australian members of the Joint Standards Australia/Standards New Zealand Committee IT-019, Information and Documentation, Information Technology — Learning, Education, Training and Research.

After consultation with stakeholders in both countries, Standards Australia and Standards New Zealand decided to develop this Standard as an Australian Standard rather than an Australian/New Zealand Standard.

The objective of this Standard is to specify the storage and support requirements for WARC (Web ARChive) file format.

This Standard is identical with, and has been reproduced from, ISO 28500:2017, *Information and documentation — WARC file format*.

As this document has been reproduced from an International Standard, a full point substitutes for a comma when referring to a decimal marker.

Australian or Australian/New Zealand Standards that are identical adoptions of international normative references may be used interchangeably. Refer to the online catalogue for information on specific Standards.

The terms 'normative' and 'informative' are used in Standards to define the application of the appendices or annexes to which they apply. A 'normative' appendix or annex is an integral part of a Standard, whereas an 'informative' appendix or annex is only for information and guidance.

Contents

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

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

This second edition cancels and replaces the first edition (ISO 28500:2009), which has been technically revised.

Introduction

Websites and web pages emerge and disappear from the World Wide Web every day. For the past 10 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 offers a standard way to structure, manage and store billions of resources collected from the web and elsewhere. It is used to build applications for harvesting, managing, accessing, mining and exchanging content. While it represents the unique standard format for web archives, it has been adopted beyond the web archiving community to store born-digital or digitized materials. 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.

Australian Standard®

Information and documentation—WARC file format

1 Scope

This document 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 documents are referred to in the text in such a way that some or all of their content constitutes requirements 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.

RFC1035[1)] Mockapetris, P. *Domain names - implementation and specification*, STD 13, November 1987

RFC2045[2)] Freed, N. and Borenstein, N. *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*, November 1996

RFC2540[3)] Eastlake, D. Detailed *Domain Name System (DNS) Information*, March 1999

RFC2616[4)] 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)

RFC3629[5)] Yergeau, F. *UTF-8, a transformation format of ISO 10646.* STD 63, November 2003

RFC3986[6)] Berners-Lee, T., Fielding, R., Masinter, L. *Uniform Resource Identifier (URI): Generic Syntax. STD 66*, January 2005 (TXT, HTML, XML)

RFC4027[7)] Josefsson, S. *Domain Name System Media Types*, April 2005

---

1) Available at: https://www.ietf.org/rfc/rfc1035.txt.

2) Available at: https://www.ietf.org/rfc/rfc2045.txt.

3) Available at: https://tools.ietf.org/html/rfc2540.

4) Available at: https://www.ietf.org/rfc/rfc2616.txt.

5) Available at: https://tools.ietf.org/html/rfc3629.

6) Available at: https://www.ietf.org/rfc/rfc3986.txt.

7) Available at: https://tools.ietf.org/html/rfc4027.