

ICS 35.240.30
CCS A 14



**NATIONAL STANDARD OF THE
PEOPLE'S REPUBLIC OF CHINA**

GB/T 33994-2026

Replacing GB/T 33994-2017

Information and documentation - WARC file format

信息与文献 WARC文件格式

(English Translation)

Issued on: July 2, 2026

Implemented on: February 1, 2027

Issued by: State Administration for Market Regulation

Table of Contents

Preface
Introduction
1 Scope
2 Normative References
3 Terms, Definitions, and Abbreviations
4 Document and Record Model
5 Named Fields
6 WARC record type
8 WARC file name, size, and compression
Appendix A (Informative) Example 16 of using WARC records
Appendix B (Informative) Example 19 of WARC Records
Appendix C (Informative) Suggestions for WARC File Size and Name
Appendix D (Informative) Compression Recommendations
References

Foreword

This document complies with the provisions of GB/T 1.1-2020 "Standardization Work Guidelines Part 1: Structure and Drafting Rules of Standardization Documents".

Drafting.

This document replaces GB/T 33994-2017 "Information and Documentation WARC File Format". Compared with GB/T 33994-2017, except for...

Aside from structural adjustments and editorial changes, the main technical changes are as follows.

---The URI definition rules have been changed (see Chapter 4, Chapter 4 in the 2017 edition);

---Added the "WARC-Refers-To-Target-URI" and "WARC-Refers-To-Date" fields (see Chapter 4, 5.12).

5.13, 6.7.2, 6.7.3);

---Added precision levels to WARC-Date (see 5.4);

---Detailed descriptions of MIME media types have been removed (see Chapters 8 and 9 in

the 2017 edition).

This document adopts the ISO 28500:2017 standard, "Information and Documentation - WARC file format".

The following minimal editorial changes have been made to this document.

---Some examples have been replaced with domestic examples (see Appendix A and Appendix B).

This document was proposed and is under the jurisdiction of the National Technical Committee on Information and Documentation Standardization (SAC/TC4).

Introduction

Websites and web pages on the internet are constantly emerging and disappearing every day. In recent years, memory storage organizations have been committed to leveraging network-scale engineering...

The goal is to find the most suitable methods for collecting and tracking this massive amount of important information. Meanwhile, memory storage organizations are working to preserve data collected non-networked.

The demand for digital files (such as entire electronic journals or data generated by environmental sensing devices) is also becoming increasingly urgent. Currently, the industry generally requires the development of...

A container format that can hold component data objects that make up the overall data structure in a simple and secure way for storage, management, and application.

Exchange. There are no restrictions on the type of these data objects (or resources) (including binary formats such as audio, CAD, and compressed files), the container...

The format only requires a minimal understanding of the nature of the object.

The WARC (WebARChive) file format provides a protocol for linking multiple resource records (data objects).

These are concatenated into a single long file, where each resource record consists of a simple text header and arbitrary blocks of data content. This format can be used as...

A standard for structuring, managing, and storing massive amounts of digital resources acquired from the internet and other channels, used to build harvesting, management, access, and mining capabilities.

Its applications extend beyond traditional online archiving, encompassing various uses such as content exchange, and extend to the preservation of native digital resources and digitized data.

WARC is an extension of the ARC file format. Compared to the ARC format, which only records the original content, the extended WARC format...

The WARC format can also carry related auxiliary content, such as allocated metadata, reduced deduplication records, and post-processing of the content.

Conversion records and segmentation information for large resources, etc.

Information and Documentation WARC File Format

1 Scope

This document specifies the format for WARC files in the field of information and documentation.

---Stores payload content and control information from mainstream Internet application layer protocols (such as HTTP, DNS, and FTP);

---Stores any metadata related to other stored data (such as subject classification, language, encoding);

---Supports data compression while ensuring the integrity of data records;

---Store all control information from the harvesting protocol (such as request header information), not just response information;

---Store the data transformation results related to other stored data;

---Storing duplicate monitoring activities related to other stored data (when the same or broadly similar resources appear, storage can be reduced).

Consumption);

---Expand without interrupting current functionality;

---Supports truncating or segmenting of excessively long records where necessary.

2 Normative references

The contents of the following documents, through normative references within the text, constitute essential provisions of this document. Dated citations are not included.

For references to documents, only the version corresponding to that date applies to this document; for undated references, the latest version (including all amendments) applies.

This document.

ISO 10646

Syntax

3.1 Terms and Definitions

The following terms and definitions apply to this document.

3.1.1

WARC Record

The basic components of a WARC file are a series of WARC records.