



中华人民共和国国家标准

GB/T 33994—2026/ISO 28500:2017

代替 GB/T 33994—2017

信息与文献 WARC 文件格式

Information and documentation—WARC file format

(ISO 28500:2017, IDT)

2026-07-02 发布

2027-02-01 实施

国家市场监督管理总局 发布
国家标准化管理委员会

目 次

前言 III

引言 IV

1 范围 1

2 规范性引用文件 1

3 术语、定义和缩略语..... 1

4 文件和记录模型 2

5 命名字段 5

6 WARC 记录类型 10

7 记录分段..... 14

8 WARC 文件名称、大小和压缩 15

附录 A（资料性） 书写 WARC 记录的使用示例 16

附录 B（资料性） WARC 记录示例 19

附录 C（资料性） WARC 文件大小及名称建议 23

附录 D（资料性） 压缩建议 24

参考文献 25

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件代替 GB/T 33994—2017《信息和文献 WARC 文件格式》，与 GB/T 33994—2017 相比，除结构调整和编辑性改动外，主要技术变化如下：

- 更改了 URI 定义规则(见第 4 章,2017 年版的第 4 章)；
- 增加了“WARC-Refers-To-Target-URI”和“WARC-Refers-To-Date”字段(见第 4 章、5.12、5.13、6.7.2、6.7.3)；
- 增加了 WARC-Date 的精度级别(见 5.4)；
- 删除了 MIME 媒体类型的详细描述(见 2017 年版的第 8 章、第 9 章)。

本文件等同采用 ISO 28500:2017《信息与文献 WARC 文件格式》。

本文件做了下列最小限度的编辑性改动：

- 将部分示例替换为国内示例(见附录 A、附录 B)。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国信息与文献标准化技术委员会(SAC/TC 4)提出并归口。

本文件起草单位：首都图书馆、中国科学院文献情报中心、军事科学院军事科学信息研究中心、国家图书馆、中国科学技术信息研究所、北京万方数据股份有限公司。

本文件主要起草人：毛雅君、贾蕾、吴振新、王雪屏、谢鹏、真溱、韩新月、胡吉颖、季士妍、张晓丹、郝春云、杨志滨、杨贺。

本文件及其所代替文件的历次版本发布情况为：

- 2017 年首次发布为 GB/T 33994—2017；
- 本次为第一次修订。

引 言

互联网上的网站和网页每天都在不断涌现和消失。近年来,记忆存储组织致力于借助网络规模工具,寻找最适宜采集并跟踪记录这类海量重要信息的方法。与此同时,记忆存储组织对保存非网络采集的数字文件(如整套电子期刊或环境感应设备生成的数据)的需求也日益迫切。当前业界普遍要求开发一种容器格式,该格式能以简单安全的方式承载构成整体数据结构的组件数据对象,用于存储、管理和交换。这些数据对象(或资源)的类型没有任何限制(包括音频、CAD、压缩文件等二进制格式),该容器格式仅需对对象本质具备最小认知即可。

WARC(Web ARChive,网络存档)文件格式提供了一种协议,该协议将多个资源记录(数据对象)连接成一个长文件,其中每个资源记录由一组简单文本标头和任意数据内容块构成。这种格式能作为结构化、管理和存储来自网络及其他渠道的海量数字资源的一种标准,用于构建收割、管理、访问、挖掘和交换内容等各种应用,其应用范围已超越传统网络存档领域,延伸至原生数字资源及数字化资料的保存。WARC文件的创建方式以及资源的存储与呈现形式,均取决于具体的软件和应用程序。

WARC文件格式是ARC文件格式的一种扩展。相较于仅记录原始内容的ARC格式,拓展后的WARC格式还能够承载相关的辅助内容,例如分配的元数据、缩减的重复数据检测记录、后期对内容的转换记录以及大型资源的分段信息等。

信息与文献 WARC 文件格式

1 范围

本文件规定了信息与文献领域 WARC 文件格式：

- 存储来自主流互联网应用层协议(如 HTTP、DNS 和 FTP)的有效载荷内容和控制信息；
- 存储与其他已存储数据(如主题分类、语言、编码)相关的任意元数据；
- 支持数据压缩,且保证数据记录的完整性；
- 存储来自收割协议的全部控制信息(如请求标头信息),而不仅仅是响应信息；
- 存储与其他已存储数据相关的数据转换结果；
- 存储与其他已存储数据相关的重复监测活动(当相同或者大体相似的资源出现时,可减少存储消耗)；
- 在不中断当前功能的情况下进行扩展；
- 支持对超长记录在所需处进行截断或分段操作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

- IETF RFC 1035 域名 实现及规范(Domain names—Implementation and specification)
- IETF RFC 2045 多用途互联网邮件扩展(MIME) 第 1 部分:互联网消息正文的格式[Multipurpose Internet Mail Extensions(MIME)—Part One:Format of Internet Message Bodies]
- IETF RFC 2540 分离域名解析系统(DNS)信息[Detached Domain Name System(DNS)Information]
- IETF RFC 2616 超文本传输协议—HTTP/1.1(Hypertext Transfer Protocol—HTTP/1.1)
- IETF RFC 3629 UTF-8—ISO 10646 的一种转换格式(UTF-8, a transformation format of ISO 10646)
- IETF RFC 3986 统一资源标识符(URI):通用语法[Uniform Resource Identifier(URI): Generic Syntax]
- IETF RFC 4027 域名解析系统媒体类型(Domain Name System Media Types)
- IETF RFC 4291 IPv6 地址架构(IP Version 6 Addressing Architecture)
- IETF RFC 5322 互联网消息格式(Internet Message Format)

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件。

3.1.1

WARC 记录 WARC record

WARC 文件的基本组成部分,WARC 文件由一序列的 WARC 记录组成。