ICS 35.240.30
A 14

中华人民共和国国家标准

GB/T 33994—2017/ISO 28500:2009

信息和文献　WARC文件格式

Information and documentation—WARC file format

(ISO 28500:2009,IDT)

2017-07-12 发布　　2018-02-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准使用翻译法等同采用 ISO 28500:2009《信息和文献　WARC 文件格式》。

与本标准中规范性引用的国际文件有一致性对应关系的我国文件如下：

——GB/T 7408—2005　数据元和交换格式　信息交换　日期和时间表示法(ISO 8601:2000，IDT)。

本标准做了下列编辑性修改：

——增加了缩略语：LWS、MIME、US-ASCII(见 3.2)；

——为了增强易读性，在保留国际标准中示例的基础上，将部分示例替换为国内示例(见附录 B)。

本标准由全国信息与文献标准化技术委员会(SAC/TC 4)提出并归口。

本标准起草单位：国家图书馆、中国科学院文献情报中心、中国国防科技信息中心、中国科技信息研究所、北京万方数据股份有限公司。

本标准主要起草人：毛雅君、李春明、吴振新、真溱、曲云鹏、张晓丹、张兰、杨贺、敦文杰、张彪。

引　言

每天，网站和网页从互联网上产生或消失。十多年来，记忆存储组织尝试用网络规模工具（如网络爬虫）寻找最适宜采集并跟踪记录海量的重要信息的方法。与此同时，记忆存储组织对保存非网络抓取的数字化资源的需求也与日俱增（如，整套电子期刊或环境感应设备生成的数据）。出现了一种需求，即希望能有一种文件格式，通过一个文件简单并安全地承载大量组成文件的数据对象，以便进行存储、管理和交换。

WARC（Web ARChive，网络存档）文件格式提供了一个由多个资源记录（数据对象）连接成一个长文件的协议，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。WARC 格式是 ARC 文件格式的扩展。WARC 格式将作为组织、管理和储存采集来自网络和其他数以亿计的数字资源的一种标准，可用于构建收割（如 Heritrix 网络爬虫，一种开源软件）、管理、访问和交换内容等各种应用。

除了用 ARC 记录的原始内容外，扩展的 WARC 格式还容纳相关的二次级内容，如分配的元数据、缩减的重复检测活动、后期转换及大型资源的切分等。

信息和文献　WARC文件格式

1　范围

本标准规定了WARC文件格式：

——存储来自于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，可以减少存储消耗)；

——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

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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。

ISO 8601　数据元和交换格式　信息交换　日期和时间表示法(Data elements and interchange formats—Information interchange—Representation of dates and times)

RFC 1035　域名　实现及标准(Domain names—Implementation and specification)

RFC 1884　IPV6地址架构(IP Version 6 Addressing Architecture)

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]

RFC 2540　分离域名解析系统(DNS)信息[Detached Domain Name System(DNS) Information]

RFC 2616　超文本传输协议—HTTP/1.1(Hypertext Transfer Protocol—HTTP/1.1)

RFC 2822　互联网消息格式(Internet Message Format)

RFC 3629　UTF-8——ISO 10646的一种转换格式(UTF-8, a transformation format of ISO 10646)

RFC 3986　统一资源标识符(URI)：通用语法[Uniform Resource Identifier(URI): Generic Syntax]

RFC 4027　域名解析系统媒体类型(Domain Name System Media Types)

W3CDTF　日期和时间格式：提交到W3C的注释(Date and Time Formats: note submitted to the W3C)

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。