ICS 01.140.20
CCS A 14

中华人民共和国国家标准化指导性技术文件

GB/Z 43768—2024/ISO/TR 14873:2013

信息与文献　网络存档的统计和质量问题

Information and documentation—Statistics and quality issues for web archiving

（ISO/TR 14873:2013,IDT）

2024-03-15 发布　　2024-10-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

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。

本文件等同采用ISO/TR 14873:2013《信息与文献　网络存档的统计和质量问题》,文件类型由IEC的技术报告调整为我国的国家标准化指导性技术文件。

本文件增加了“规范性引用文件”一章。

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。

本文件做了下列最小限度的编辑性改动:

——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;

——针对我国没有明确网络信息存档的法定呈缴机构的现状,修改第1章相关表述。

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:中国科学院文献情报中心、国家图书馆、中国科学院档案馆、北京大学图书馆。

本文件主要起草人:吴振新、张冬荣、潘亚男、敦文杰、朱佳丽、曲云鹏、孙超、谢靖、付鸿鹄、单嵩岩、薛杰、吴欣雨、孔贝贝、胡吉颖、陈子俊、张静。

引　言

本文件是为了指导我国网络存档以及网络存档产品的管理和评估而制定。

网络存档指随着时间推移，对互联网资源的快照进行选择、抓取、存储(Storing)、保存(Preserving)和访问管理的活动。20 世纪 90 年代末，人们预见到互联网资源存档将成为未来研究、商业和政府的重要记录，开始实施网络存档。互联网资源被视为文化遗产的一部分，能像印本那样得到保存。许多参与网络存档的机构将此视为保护国家文化遗产这一长期使命的延伸，且受到许多国家法律法规如法定缴存制度的认可和支持。

互联网上提供多种类型的资源，包括文本、图片、电影、音频及其他多媒体格式的资源。除了相互链接的网络页面外，还有通过使用各种传输与通信协议提供的新闻组、时事通讯、博客和交互式服务(如游戏)。网络存档通过采集软件对互联网资源副本进行自动采集(通常是定期执行)。网络存档的目标是实现资源的回放，包括内在关联，例如通过超文本链接，尽可能呈现出与原始环境中一样的效果。网络存档的主要目标是尽可能地按原始状态永久保存网络记录，以满足各种学术、专业和私人用途。

网络存档是一项新兴但不断扩展的活动，需要持续引入新方法和工具以与快速发展的网络技术保持同步。由于存档机构对战略重要性的认识、可采用方式以及法律要求的不同，导致出现了多种互联网资源的存档方法，存档范围涉及单个网络页面抓取到全部顶级域抓取。不同组织的网络存档成熟度等级也不同，对于某些组织来说，网络存档已成为其常规业务活动，而有些组织则针对这项挑战刚刚启动试验计划。

根据采集的规模和目的，网络存档策略分为两大类：批量采集和选择性采集。大规模的批量采集，如国家域采集，旨在抓取整个域(或其子集)的快照。选择性采集的规模则小得多，采集更集中且更频繁，经常是依据某项规则执行，例如，主题、事件、格式(如音频或视频文件)或与内容所有者之间的协议。这两种策略的关键区别在于质量控制程度，即对所采集网站进行评估以确定是否达到预定义的质量标准。域采集的规模(如此之大)使得无法通过人工对所采集的资源和该资源的实时版本进行任何人工比对，而该方式在选择性采集中则是一种常用的质量保证方法。

本文件旨在证明网络存档作为广义文化遗产资源集合的一部分，基于传统的图书馆工作流，用类似的和兼容的方式进行评估和管理。本文件阐述了资源集合建设、表征、描述、保存、使用和组织结构，同时表明，尽管在实践中需要做出调整，但传统资源集合管理工作流的大多数方面原则上仍然适用于网络存档。

本文件概述了网络存档的现状，重点给出了网络存档统计数据和质量指标的定义和使用。一些统计数据的产生依赖于所使用的采集、索引或浏览软件，选择不同的软件可能会导致结果的差异。本文件并不给出特定或推荐的软件，而是提供一组指标来帮助评估网络存档的总体性能和质量。

信息与文献　网络存档的统计和质量问题

1　范围

本文件为网络存档定义了统计数据、术语和质量标准。本文件考虑了图书馆、档案馆、博物馆、研究中心和文化遗产基金会等众多机构组织的需求和实践。

本文件面向直接参与网络存档的专家，通常是由网络存档机构的领导决策人员、工程师和保存管理人员组成的团队。对网络存档机构的资助机构和利益相关方也同样有用。本文件使用的专业术语试图能够表达受众所拥有的广泛兴趣和专业知识，并在计算机科学、管理和图书馆学之间达到平衡。

本文件不适用于学术和商业电子资源的管理，如电子期刊、电子报纸或电子书，这些资源通常使用不同的管理系统单独存储和处理。它们虽然被视为互联网资源，但在本文件中不作为网络存档的特定内容流进行阐述。一些组织还采集通过网络分发的电子文档，如通过出版商的电子存储库和仓储系统，这些内容也不在本文件的阐述范围。这类采集使用的原理和技术与网络存档有很大不同，因此本文件的统计数据和质量指标不一定适用。

本文件专注于网络存档的原理和方法，不包括其他采集互联网资源的方式。事实上，一些互联网资源，尤其是那些不在网络上传播的资源（如以电子邮件形式传播的通信），不是通过网络存档技术采集的，而是通过其他方式采集的，而这些方式也不属于本文件的适用范围。

2　规范性引用文件

本文件没有规范性引用文件。

3　术语和定义

下列术语和定义适用于本文件。

3.1

访问　access

图书馆提供的在线服务的成功请求（3.36）。

注 1：一次访问是用户活动的一个周期，通常从用户连接到图书馆提供的在线服务时开始，并以显式（通过注销或退出离开数据库）或隐式（由于用户不活动而导致超时）的终止活动结束。

注 2：对图书馆网站（3.52）的访问被视为虚拟访问。

注 3：不包括通用入口或网关页面（3.33）的请求（3.36）。

注 4：尽可能地不包括搜索引擎发起的请求（3.36）。

[来源：ISO 2789:2022，3.2.1]

3.2

访问工具　access tool

用于查找、检索和回放存档互联网资源的专业软件。

注：该工具通过组合运行多个独立软件包实现。

3.3

管理元数据　administrative metadata

妥善管理存储库中数字对象所必需的信息。