

ICS 01.140.20

CCS A 14



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

GB/Z 43768-2024

**Information and documentation - Statistics and quality issues
for web archiving**

Issued on: March 15, 2024

Implemented on: October 1, 2024

**Issued by: State Administration for Market Regulation, China National
Standardization Administration**

Table of Contents

Preface	
Introduction	
1 Scope	1
2 Normative references	1
3 Terms and Definitions	1
4 Methods and Purpose of Web Archiving	7
4.1 Collection methods	7
4.2 Access and Description Methods	9
4.3 Saving Method	11

Foreword

This document is in accordance with the provisions of GB/T 1.1-2020 "Guidelines for standardization work Part 1. Structure and drafting rules for standardization documents"

Drafting.

This document is equivalent to ISO /T R14873.2013 "Statistics and quality issues for information and documentation network archiving", and the file type is

IEC 's technical reports have been adjusted to become my country's national standardization guiding technical documents.

This document adds a chapter on "Normative References".

Introduction

This document is developed to guide the management and evaluation of network archiving and network archiving products in my country.

Web archiving refers to the selection, capture, storage, and preservation of snapshots of Internet resources over time.

In the late.1990s, it was foreseen that archiving Internet resources would become an important part of future research, business, and government.

To record, start implementing web archiving. Internet resources are considered part of cultural heritage and can be preserved like printed books. Many participants

The network archives see this as an extension of their long-standing mission to protect the

nation's cultural heritage and are subject to many national laws and regulations such as statutory fees.

Recognition and support of the existing system.

The Internet provides a variety of resources, including text, pictures, movies, audio and other multimedia formats.

In addition to the web pages that are linked to, there are newsgroups, newsletters, blogs, and interactive services (such as games) that are provided using various transmission and communication protocols.

The goal of a web archive is to automatically collect copies of Internet resources (usually on a regular basis) using collection software.

Realize the playback of resources, including internal connections, such as through hypertext links, to present the same effect as the original environment as much as possible.

Network

The main goal of the archive is to preserve web records permanently in as original a state as possible for a variety of academic, professional, and private purposes.

Web archiving is an emerging but expanding activity that requires the continued introduction of new methods and tools to keep pace with the rapid development of web technologies.

Different archiving institutions have different perceptions of strategic importance, available approaches, and legal requirements, leading to the emergence of a variety of Internet archives.

Resource archiving methods, ranging from single web page crawling to full top-level domain crawling. Web archiving maturity of different organizations, etc.

For some organizations, web archiving is already a regular part of their business, while for others, this challenge is just beginning.

Experiment plan.

Based on the scale and purpose of the collection, network archiving strategies can be divided into two categories. bulk collection and selective collection.

A collection, such as a national domain collection, aims to capture a snapshot of an entire domain (or a subset of it). A selective collection is much smaller, more focused, and more

Frequent, often based on a rule, such as subject matter, event, format (such as audio or video files), or agreement with content owners.

The key difference between these two strategies is the degree of quality control, that is, the evaluation of the collected websites to determine whether they meet the predefined quality

standards.

The scale of the domain collection (so large) that it is impossible to manually verify the collected resources and the live version of the resources.

This method is a common quality assurance method in selective collection.

This paper aims to demonstrate that web archives, as part of a broader collection of cultural heritage resources, can be used in a similar way to traditional library workflows.

This document describes the collection construction, characterization, description, conservation, use and organization structure, and

While some adjustments may be needed in practice, most aspects of the traditional collection management workflow are still applicable to the Web in principle.

Archive.

This document provides an overview of the current state of web archiving, with an emphasis on the definition and use of web archiving statistics and quality indicators.

The generation of statistical data depends on the acquisition, indexing or browsing software used. Choosing different software may lead to different results.

It does not present specific or recommended software, but rather provides a set of metrics to help assess the overall performance and quality of web archives.

Statistics and quality issues in online archiving of information and documents

1 Scope

This document defines statistics, terminology and quality standards for web archiving. This document takes into account libraries, archives, museums, research

The needs and practices of a wide range of institutional organizations, including the Centre for International Cooperation and Development and the Cultural Heritage Foundation.

This document is intended for experts directly involved in web archiving, typically web archiving agency leadership, engineers, and preservation managers.

It is also useful to funding agencies and stakeholders in web archiving. The terminology used in this document attempts to

Ability to represent a wide range of interests and expertise held by the audience, with a balance between computer science, management, and librarianship.

This document is not intended for the management of academic and commercial electronic resources, such as electronic journals, electronic newspapers, or electronic books, which are usually managed using

Different management systems store and process them separately. Although they are considered Internet resources, they are not included in this document as specific web archives.

Some organizations also collect electronic documents that are distributed over networks, such as through publishers' electronic repositories and repository systems.

The principles and techniques used in this type of collection are very different from those used in network archiving, so this article will not discuss them in detail.

Statistics and quality indicators of the documents may not be applicable.

This document focuses on the principles and methods of web archiving and does not cover other ways of collecting Internet resources.

Sources, especially those that are not distributed over the Internet (such as communications distributed in the form of e-mail), are not collected through network archiving technology.

The data are not collected by other methods, but are collected by other methods, which are not within the scope of application of this document.

2 Normative references

This document has no normative references.

3 Terms and definitions

The following terms and definitions apply to this document.

3.1

Access

A successful request for an online service provided by the library (3.36).

Note 1. A visit is a period of user activity that usually begins when the user connects to an online service provided by the library and ends explicitly (by logging off or

The termination activity ends either by exiting (leaving the database) or implicitly (timeout due to user inactivity).

Note 2. Visits to the library website (3.52) are considered virtual visits.

NOTE 3 Requests (3.36) to general entry or gateway pages (3.33) are not included.

Note 4. Requests initiated by search engines (3.36) are excluded as far as possible.

[Source. ISO 2789.2022, 3.2.1]

3.2