SLOVENSKI STANDARD
SIST-TP ISO/TR 14873:2017

01-februar-2017

Informatika in dokumentacija - Statistika in vprašanja glede kakovosti za spletno arhiviranje

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

Information et documentation -- Statistiques et indicateurs de qualité pour l'archivage du web

Ta slovenski standard je istoveten z: **ISO/TR 14873:2013**

ICS:

01.140.20	Informacijske vede	Information sciences
03.120.99	Drugi standardi v zvezi s kakovostjo	Other standards related to quality

SIST-TP ISO/TR 14873:2017 **en**

TECHNICAL REPORT

ISO/TR 14873

First edition
2013-12-01

Information and documentation — Statistics and quality issues for web archiving

Information et documentation — Statistiques et indicateurs de qualité pour l'archivage du web

Reference number
ISO/TR 14873:2013(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 46, *Information and documentation*, Subcommittee SC 8, *Quality - Statistics and performance evalutation*.

Introduction

This Technical Report was developed in response to a worldwide demand for guidelines on the management and evaluation of Web archiving activities and products.

Web archiving refers to the activities of selecting, capturing, storing, preserving and managing access to snapshots of Internet resources over time. It started at the end of the 1990s, based on the vision that an archive of Internet resources would become a vital record for research, commerce and government in the future. Internet resources are regarded as part of the cultural heritage and therefore preserved like printed heritage publications. Many institutions involved in Web archiving see this as an extension of their long standing mission of preserving their national heritage, and this is endorsed and enabled in many countries by legislative frameworks such as legal deposit.

There is a wide range of resources available on the Internet, including text, image, film, sound and other multimedia formats. In addition to interlinked Web pages, there are newsgroups, newsletters, blogs and interactive services such as games, made available using various transfer and communication protocols. Web archives bring together copies of Internet resources, collected automatically by harvesting software, usually at regular intervals. The intention is to replay the resources including the inherent relations, for example by means of hypertext links, as much as possible as they were in their original environment. The primary goal of Web archiving is to preserve a record of the Web in perpetuity, as closely as possible to its original form, for various academic, professional and private purposes.

Web archiving is a recent but expanding activity which continuously requires new approaches and tools in order to stay in sync with rapidly evolving Web technology. Determined by the strategic importance perceived by the archiving institution, means available and sometimes legal requirements, diverse approaches have been taken to archive Internet resources, ranging from capturing individual Web pages to entire top-level domains. From an organisational perspective, Web archiving is also at different levels of maturity. While it has become a business as usual activity in some organisations, others have just initiated experimental programmes to explore the challenge.

Depending on the scale and purpose of collection, a distinction can be made between two broad categories of Web archiving strategy: bulk harvesting and selective harvesting. Large scale bulk harvesting, such as national domain harvesting, is intended to capture a snapshot of an entire domain (or a subset of it). Selective harvesting is performed on a much smaller scale, is more focused and undertaken more frequently, often based on criteria such as theme, event, format (e.g. audio or video files) or agreement with content owners. A key difference between the two strategies lies in the level of quality control, the evaluation of harvested Websites to determine whether pre-defined quality standards are being attained. The scale of domain harvesting makes it impossible to carry out any manual visual comparison between the harvested and the live version of the resource, which is a common quality assurance method in selective harvesting.

This Technical Report aims to demonstrate how Web archives, as part of a wider heritage collection, can be measured and managed in a similar and compliant manner based on traditional library workflows. The report addresses collection development, characterization, description, preservation, usage and organisational structure, showing that most aspects of the traditional collection management workflow remain valid in principle for Web archiving, although adjustment is required in practice.

While this Technical Report provides an overview of the current status of Web archiving, its focus is on the definition and use of Web archive statistics and quality indicators. The production of some statistics relies on the use of harvesting, indexing or browsing software, and a different choice of software may lead to variance in the results. This Technical Report however does not endorse nor recommend any software in particular. It provides a set of indicators to help assess the performance and quality of Web archives in general.

This Technical Report should be considered as a work in progress. Some of its contents are expected to be incorporated in the future into ISO 2789 and ISO 11620.

Information and documentation — Statistics and quality issues for web archiving

1 Scope

This Technical Report defines statistics, terms and quality criteria for Web archiving. It considers the needs and practices across a wide range of organisations such as libraries, archives, museums, research centres and heritage foundations. The examples mentioned are taken from the library sector, because libraries, especially national libraries, have taken up the new task of Web archiving in the context of legal deposit. This should in no way be taken to undermine the important contributions of institutions which are not libraries. Neither does it reduce the principal applicability of this Technical Report for heritage institutions and archiving professionals.

This Technical Report is intended for professionals directly involved in Web archiving, often in mixed teams consisting of library or archive curators, engineers and managerial staff. It is also useful for Web archiving institutions' funding authorities and external stakeholders. The terminology used in this Technical Report attempts to reflect the wide range of interests and expertise of the audiences, striking a balance between computer science, management and librarianship.

This Technical Report does not consider the management of academic and commercial electronic resources, such as e-journals, e-newspapers or e-books, which are usually stored and processed separately using different management systems. They are regarded as Internet resources and are not addressed in this Technical Report as distinct streams of content of Web archives. Some organisations also collect electronic documents, which may be delivered through the Web, through publisher-based electronic deposits and repository systems. These too are out of scope for this Technical Report. The principles and techniques used for this kind of collecting are indeed very different from those of Web archiving; statistics and quality indicators relevant for one kind of method are not necessarily relevant for the other.

Finally, this Technical Report essentially focuses on Web archiving principles and methods, and does not encompass alternative ways of collecting Internet resources. As a matter of fact, some Internet resources, especially those that are not distributed on the Web (e.g. newsletters distributed as e-mails) are not harvested by Web archiving techniques and are collected by other means that are not described nor analysed in this Technical Report.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
access
successful request of a library-provided online service

Note 1 to entry: An access is one cycle of user activities that typically starts when a user connects to a library-provided online service and ends by a terminating activity that is either explicit (by leaving the database through log-out or exit) or implicit (timeout due to user inactivity).

Note 2 to entry: Accesses to the library website are counted as virtual visits.

Note 3 to entry: Requests of a general entrance or gateway page are excluded.

Note 4 to entry: If possible, requests by search engines are excluded.

[SOURCE: ISO 2789:2013, definition 2.2.1]

2.2
access tool
specialist software used to find, retrieve and replay archived Internet resources

Note 1 to entry: This may be implemented by a number of separate software packages working together.

2.3
administrative metadata
information necessary to allow the proper management of the digital objects in a repository

Note 1 to entry: Administrative metadata can be divided into the following categories:

— context or provenance metadata: describe the lifecycle of a resource to a point, including the related entities and processes, e.g. configuration and log files;

— technical metadata: describe the technical characteristics of a digital object, e g. its format;

— rights metadata: define the ownership and the legally permitted usage of an object.

2.4
archive
Web archive
entire set of resources crawled from the Web over time, comprising one or more collections

2.5
bit stream
series of 0 and 1 digits that constitutes a digital file

2.6
budget (crawl)
limitation associated with a crawl or individual seeds, which can be expressed in e.g. number of files, volume of data, or the time to be spent per crawl as defined in the crawler settings

2.7
bulk crawl
bulk harvest
crawl aimed at collecting the entirety of a single or multiple top level domain(s) or a subset(s)

Note 1 to entry: In comparison with selective crawls, bulk crawls have a wider scope and are typically performed less frequently.

Note 2 to entry: Bulk crawls generally result in large scale Web archives, making it impossible to conduct detailed quality assurance. This is often done through sampling.

2.8
capture
instance
copy of a resource crawled at a certain point in time

Note 1 to entry: If a resource has been crawled three times on different dates, there will be three captures.

2.9
collection
Web archive collection
cohesive resources presented as a group

Note 1 to entry: A collection can either be selected specifically prior to harvesting (e. g. an event, a topic) or pulled together retrospectively from available resources in the archive.

Note 2 to entry: A Web archive may consist of one or more collections.

2.10
crawl
harvest
process of browsing and copying resources using a crawler

Note 1 to entry: Crawls can be categorised as bulk or selective crawls.

2.11
crawl settings
crawl parameters
definition of which resources should be collected and the frequency and depth required for each set of seeds

Note 1 to entry: Crawl settings also include crawler politeness (number of requests per second or minute sent to the server hosting the resource), compliance with robots.txt and filters to exclude crawler traps.

2.12
crawler
harvester
archiving crawler
DEPRECATED: spider
software that will successively request URLs and parse the resulting resource for further URLs

Note 1 to entry: Resources may be stored and URLs discarded in accordance with a predefined set of rules [see *crawl settings* (2.11) and *scope (crawl)* (2.40)].

2.13
crawler trap
Web page (or series thereof) which will cause a crawler to either crash or endlessly follow references to other resources deemed to be of little or no value

Note 1 to entry: Crawler traps could be put in place intentionally to prevent crawlers from harvesting resources. This could also occur inadvertently for example when a crawler follows dates of a calendar endlessly.

2.14
curator tool
application that runs on top of a Web crawler and supports the harvesting processes

Note 1 to entry: A core function is the management of targets and the associated descriptive and administrative metadata. It may also include components for scheduling and quality control.

2.15
data mining
computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing it, and summarizing potential relationships and impacts

[SOURCE: ISO 16439:—, definition 3.13]

2.16
deep Web
DEPRECATED: hidden Web
DEPRECATED: invisible Web
part of the Web which cannot be crawled and indexed by search engines, notably consisting of resources which are dynamically generated or password protected

2.17
descriptive metadata
information describing the intellectual content of a digital object

2.18
domain name
identification string that defines a realm of administrative autonomy, authority, or control on the Internet, defined by the rules and procedures of the domain name system (DNS)

2.19
domain name system
DNS
hierarchical, distributed global naming system used to identify entities connected to the Internet

Note 1 to entry: The Top Level Domains (TLDs) are the highest in the hierarchy.

2.20
emulation
recreation of the functionality and behaviour of an obsolete system, using software (called emulator) on current computer systems

Note 1 to entry: Emulation is a key digital preservation strategy.

2.21
host
portion of a URI that names the network source of the content

Note 1 to entry: A host is typically a domain name such as www.archive.org, or a subdomain such as web.archive.org.

2.22
HTML
Hypertext Markup Language
the main mark-up language for Web pages, consisting of elements which are used to add structural and semantic information to raw text

2.23
HTTP
Hypertext Transfer Protocol
client/server communication protocol used to transfer information on the Web

2.24
hyperlink
link
relationship structure used to link information on the Internet

2.25
junk
spam
unsolicited contents which are deemed to be of no relevance or long-term value

Note 1 to entry: Intentional spam is commonly used to manipulate search engine indexes. Junk can also be generated inadvertently when a crawler falls in a crawler trap.

Note 2 to entry: Collecting institutions in general try to avoid collecting junk and spam so that resources can be used to harvest “good” resources. Some, however, keep a small sample of this as a part of the record of the Web.

2.26
link mining
processing and analysis that focus on extracting patterns and heuristics from hyperlinks, e. g. to draw network graphs

2.27
live Web leakage
common problem in rendering archived resources, which occurs when links in an archived resource resolve to the current copy on the live site, instead of to the archival version within a Web archive

Note 1 to entry: Live Web leakage also occurs when scripts on archived Web pages continue to reference, and successfully request, live Web resources within the archival rendering. This may cause live Web social media feeds or streaming videos, for example, to appear in the archived webpage.

2.28
log file
file automatically created by a server that maintains a record of its activities

2.29
metadata
data describing context, content and structure of digital object and their management through time

[SOURCE: ISO 15489-1:2001, definition 2.12]

Note 1 to entry: Metadata can be categorised as descriptive, structural and administrative metadata.

2.30
migration
conversion of older or obsolete file formats to newer or current ones for the purpose of maintaining the accessibility of a digital object

Note 1 to entry: Migration is a key preservation strategy.

[SOURCE: ISO 15489-1:2001, definition 3.13]

2.31
MIME type
Internet media type
content type
two-part identifier for file formats on the Internet

Note 1 to entry: MIME (Multipurpose Internet Mail Extensions) uses the content-type header, consisting of a type and a subtype, to indicate the format of a resource, e.g. image/jpeg.

2.32
nomination
candidate resource to be considered for inclusion in a Web archive

2.33
page
Web page
structured resource, which in addition to any human-readable content, contains zero or more relationships with other resources and is identified by a URL

2.34
permission
authorization to crawl a live website and/or to publicly display its content on a Web archive

Note 1 to entry: Permission can be expressed by a formal licence from the rights holder or exempted by the virtue of legal deposit.

2.35
registered user
person or organization registered with a library in order to use its collection and/or services within or away from the library

Note 1 to entry: Users can be registered upon their request or automatically when enrolling in the institution.

Note 2 to entry: The registration is monitored at regular intervals, at least every three years, so that inactive users can be removed from the register.

[SOURCE: ISO 2789:2013, definition 2.2.28]

2.36
request
HTTP-formatted message sent by a requesting system (e.g. a browser or a crawler) to a remote server for a particular resource identified by a URL

2.37
response
answer by a remote server to an HTTP request for a resource, containing either the requested resource, a redirection to another URL or a negative (error) response, indicating why the requested resource could not be returned

2.38
response code
status code
three-digit number indicating to the requesting server the status of the requested resource

Note 1 to entry: Codes starting with a 4 (4xx), for example, indicate that the requested resource is not available.

2.39
robots.txt
robots exclusion standard
protocol used to prevent Web crawlers from accessing all or part of a website

Note 1 to entry: robots.txt is not legally binding.

Note 2 to entry: It may also be used to request a minimum delay between consecutive requests or even to provide a link to a site map to facilitate better crawling of the site.

2.40
scope (crawl)
set of parameters which defines the extent of a crawl, e.g. the maximum number of hops or the maximum path depth the crawler should follow

Note 1 to entry: The scope of a crawl can be as broad as a whole top level domain (e. g. .de) or as narrow as a single file.

2.41
scope (Web archive)
extent of a Web archive or collection, as determined by the institutional legal mandate or collection policy

2.42
second level domain
subdivisions within the top level domains for specific categories of organisations or areas of interest (e. g. .gov.uk for governmental websites, .asso.fr for associations' websites)

2.43
seed
targeted URL
URL corresponding to the location of a particular resource to be crawled, used as a starting point by a Web crawler

2.44
selection
curatorial decision-making process which determines whether a meaningful set of resources is in scope for a Web archive, judged against its collection development policy

2.45
selective crawl
selective harvest
crawl aimed at collecting resources selected according to certain criteria

Note 1 to entry: In comparison with bulk crawls, selective crawls have a narrower scope and are typically performed more frequently.

Note 2 to entry: Selective continuous crawls are crawls aimed at collecting resources selected according to certain criteria, such as scholarly importance, relevance to a subject or continuous update frequency of the resource.

Note 3 to entry: Selective event crawls are time-bound crawls, which end at a certain date, aimed at collecting resources related to unique events, such as elections, sport events and disasters.

2.46
structural metadata
information that describes how compound objects are constructed together to make up logical units

2.47
target
meaningful set of resources to be collected as defined by one or more seeds and the associated crawl settings

2.48
top level domain
TLD
highest level of domains in the Domain Name System (DNS), including country-code top-level domains (e. g. .fr, .de), which are based on the two-character territory codes of ISO 3166 country abbreviation, and generic top-level domains (e. g. .com, .net, .org, .paris.)

Note 1 to entry: Unless specifically stated, this term is used to mean country-code TLDs in the report.

2.49
Uniform Resource Identifier
URI
extensible string of characters used to identify or name a resource on the Internet

2.50
Uniform Resource Locator
URL
subset of the Uniform Resource Identifier (URI) that specifies the location of a resource and the protocol for retrieving it

2.51
WARC format
file format that specifies a method for combining multiple digital resources into an aggregate archival file together with related information

Note 1 to entry: The WARC (Web ARChive) format has been an ISO standard since 2009 (ISO 28500:2009).

2.52
website
set of legally and/or editorially interconnected Web pages

Note 1 to entry: Usually websites represent official institutions, organizations, private firms and private homepages.

2.53
Web
main publishing application of the Internet, enabled by three key standards: URI, HTTP and HTML

3 Methods and purposes of Web archiving

The form and content of Web archives are determined by institutional policies as well as technical possibilities. While high level policies are primarily set by national legislation, institutions employ a variety of collecting strategies, driven by respective business objectives and selection criteria. In-scope resources, however, sometimes cannot be added to Web archives due to technical limitations. Capturing and replaying multimedia and interactive resources for example pose significant challenges for the Web archiving community and often require expensive, customised solutions.