PD ISO/TR 14873:2013

BSI Standards Publication

Information and documentation — Statistics and quality issues for web archiving

bsi.

...making excellence a habit.™

Document Sharing Hub

National foreword

This Published Document is the UK implementation of ISO/TR 14873:2013.

The UK participation in its preparation was entrusted to Technical Committee IDT/2/15, Library and publishing statistics.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2013. Published by BSI Standards Limited 2013

ISBN 978 0 580 81232 3

ICS 01.140.20

Compliance with a British Standard cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 31 December 2013.

Amendments issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: PD ISO/TR 14873:2013

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)

Documents
Document Sharing Hub

© ISO 2013

Full document details: PD ISO/TR 14873:2013

COPYRIGHT PROTECTED DOCUMENT

© ISO 2013

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub

Full document details: PD ISO/TR 14873:2013

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Terms and definitions **1**
3 Methods and purposes of Web archiving **7**
3.1 Collecting methods 8
3.2 Access and description methods 10
3.3 Preservation methods 12
3.4 Legal basis for Web archiving 14
3.5 Additional reasons for Web archiving 15
4 Statistics **16**
4.1 General 16
4.2 Statistics for collection development 16
4.3 Collection characterization 22
4.4 Collection usage 28
4.5 Web archive preservation 31
4.6 Measuring the costs of Web archiving 35
5 Quality indicators **37**
5.1 General 37
5.2 Limitations 37
5.3 Description 38
6 Usage and benefits **47**
6.1 General 47
6.2 Intended usage and readers 47
6.3 Benefits for user groups 48
6.4 Use of proposed statistics by user groups 48
6.5 Web archiving process with related performance indicators 50
Bibliography **52**

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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*.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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.

Documents
Document Sharing Hub

© ISO 2013 – All rights reserved

Full document details: PD ISO/TR 14873:2013

Documents
Document Sharing Hub

Full document details: PD ISO/TR 14873:2013

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]

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub

Full document details: PD ISO/TR 14873:2013

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.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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)

Documents
Document Sharing Hub

© ISO 2013 – All rights reserved

Full document details: PD ISO/TR 14873:2013

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.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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]

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

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.

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

3.1 Collecting methods

3.1.1 Technical baseline

Copying online resources or harvesting is the main collecting method for Internet resources. Harvesting requires the use of robots, which successively request URLs, copy and store resources, and parse the resulting resource for further URLs. The crawler's starting point, often the home page of a Website, is called the seed. The crawler behaves like an automated Web user and can follow interlinked Internet resources almost indefinitely, unless its scope is defined or limited by crawl parameters or settings. A crawler can also come to a halt inadvertently when encountering obstacles during the harvesting process.

The coverage, depth and the overall quality of Web archive collections are closely determined by a set of technical settings, expressed as harvesting rules. Resources to be collected are described by their locations (URLs), in the form of a seed list, and by their scope. The scope is mainly defined by the frequency and depth of the harvest, which impact the comprehensiveness of a Web archive.

3.1.2 Limitations

A huge amount of information is added to the Web at an astonishing speed. Organisations typically decide on the scope of Web archiving taking into account resources related to staff, computation power and storage capacity. In addition, choices made during the selection and harvesting processes allow organisations to focus on the valuable and "good" resources, as opposed to automatically generated content of little value such as junk or spam. This is a major source of noise to avoid for Web archives, requiring active management and prioritization of the crawling process.

There are many limitations which make it challenging to collect Internet resources comprehensively. Some are related to technology, others are caused by the scale and nature of the Internet. Additional limitations may be imposed by legislations.

a) Issues due to current Web architecture and crawling technologies

Archiving crawlers are capable of capturing adequately static content which can be served by requesting a URL. When URLs are not explicitly referenced in HTML but embedded in e. g. JavaScript or Flash presentations or generated dynamically based on the results of some interactions with the users, archiving crawlers often fail to capture the referenced content. Extracting and parsing arbitrary URLs is not always a simple task as the URL syntax can be stretched to address almost any type of network resource and URLs can be generated dynamically. Overly complex URL structures include numerous variables, marked by ampersands, equals signs, session or user IDs as well as referral tracking codes. In some cases, multimedia files are served or initiated by embedded Web applications which retrieve data from the server in the background, without explicitly locating the files in the HTML.

The current harvesting technology is still to a large extent not adequate to deal with the Web in full, leaving certain types of content on the Web out of reach. The current recursive URL-based crawling method falls short of collecting an increasingly bigger portion of the Web, including content behind Web forms and query interfaces, commonly known as the "deep or hidden Web", streaming media, content delivered over non-HTTP protocols and social media.

The most fundamental challenge facing the community however is the rapidly changing Web with new formats, protocols and platforms, requiring the archiving organisations to respond to its continuous development and improve the capability to archive new content types as they emerge.

b) Issues due to Web resources update frequency

Another commonly cited technical issue related to crawling is temporal incoherence. If a website is been updated while being crawled, this could result in a distorted snapshot with the co-existence of Web pages with different life spans.

Legislation may impose further restrictions to the way in which Web archiving is done. A key decision influenced by legislation is whether robots.txt exclusion standards should be respected or ignored. It can make a significant difference to whether certain content is captured or missed.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

The above-mentioned limitations bring up many challenges to the process of evaluating Web archiving activities using comparable measures. The general approach followed in this Technical Report is to acknowledge the limitations and to focus on what is known and comparable.

3.1.3 Collecting strategies

There are two broad categories of collecting strategy, which vary in the level of automation involved as well as the scope and scale of the resulting Web archive.

— Bulk harvesting, such as national domain harvesting, is intended to capture a snapshot of the state of an entire domain (or a subset such as a national domain) at a given point in time, resulting in large scale Web archive collections. The best known bulk archive is the Internet Archive's Wayback Machine, which was established with the goal of preserving the global Web. Bulk harvesting is a fairly automated process but limited by the scale of operation. It tends to take place infrequently, often just once or twice a year. Quality assurance, if undertaken, often relies on automatic checking of missing content by examining HTTP status codes.

— Selective archiving is performed on a much smaller scale, more focused and undertaken more frequently. A selection process takes place to identify relevant Websites based on criteria such as theme, event, format (e.g. audio or video files) or agreement with content owners. Quality assurance is a common element of selective archiving, which currently heavily relies on visual comparison, review of previous harvests and crawl logs. A selective Web archive also tends to have more descriptive metadata, often added by curators during the selection or after the harvesting process, which can be used to build richer search and browsing functions in the user interface of a Web archive.

— A number of institutions use a combination of the above. Some Websites update frequently and the changes will not be captured by just relying on infrequent bulk or domain harvests. It is not uncommon for a single archiving organization to define a strategy where high-priority Websites are captured more frequently while lower priority Websites are only captured through bulk or domain harvesting.

3.1.4 Selection criteria

Selection criteria are usually set out by legislation and institutional collection development strategy, in alignment with the core mission of a collecting institution. Operational considerations or limitations such as staffing, resource and expertise often impact the implementation of the strategy. The selection criteria define the scope of a Web archive and can be expressed in various ways:

— By domain names used to host the resources, for example by national or top level domains such as .fr or .de or second-level domains reserved to certain publishers such as .gov for government publications. Domain names however cannot strictly identify or define national content as the Internet is a global system and resources are distributed across physical or geographical boundaries.

— By characteristics of the resource, for example by themes or topics of the websites' content, by popularity with the audience or language used, by communication protocols used to deliver the resource such as HTTP or by formats such as text or video;

— By access condition or copyright status of the resource, for example whether it is freely available, or available for purchase or subscription;

— By what an organization can afford to archive. An organization may only have the financial resources to support limited harvesting frequency or afford a highly selective sampling approach;

— By explicit restrictions or exceptions related to the content. For example a selection criterion could be to exclude resources containing personal, sensitive data or illegal content.

It is not always clear-cut whether to include or exclude certain types of resources. While one organization decides to archive social networks, blogs and similar interactive platforms, others may decide that they are out of scope. This is essentially a policy decision which equally applies to resources such as online advertising, pornography and resources containing or affected by computer viruses. Sampling is a

© ISO 2013 – All rights reserved

Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

common way of archiving such resources, which might contain value to some researchers but whose future use is difficult to foresee at present.

3.2 Access and description methods

3.2.1 Technical baseline: description methods

3.2.1.1 General

It is a common practice to bring together archived resources and provide access through stand-alone Web archives. They work in the same way as the live Web, often with dedicated user interfaces which allow the users to search and to navigate within the boundary of the archive. A particular requirement for designing the user interface is to take into account the temporal dimension, allowing the users to find different versions of the same resource captured at different times and to navigate easily between these to see the evolution over time. The most common way to browse a Web archive is by URL and this could be combined with the date of capture.

3.2.1.2 Indexing by URL (mandatory)

Indexes provide entry points into the Web archive. They speed up searching and sorting operations and offer better user experience. The most basic index is the URL or a modified variant of the original URL to point to the Web server hosting the archive. The date the resource was crawled can be combined into the URL in order to differentiate between versions of the same resource. An alternative approach is to implement a persistent identifier to each resource which may take the form of a URL, but the key element is that the collecting institution guarantees that the identifier will provide a reference and access method to the resource indefinitely.

3.2.1.3 Other kinds of indexing (optional)

Full-text search is an access method increasingly adopted by Web archives. This requires a full text index and a search engine. While being a more scalable access solution, it is technically challenging to implement. Keyword and metadata can also be extracted automatically from the archived resources and utilized to provide access. There has been a growing need expressed among research communities for data and link mining in Web archives. Some new developments already demonstrate a shift of focus in Web archiving, from the level of individual resource or Website to the entire Web archive. Using visualization and data analytic techniques, there are opportunities to provide access to different views of a Web archive, unlocking embedded patterns and trends, relationships and contexts. Before the above-mentioned developments become widely adopted practices, making available the previous states of individual Internet resources remains the primary access mechanism for Web archives. This is also the focus of this Technical Report.

3.2.1.4 Cataloguing (optional)

The traditional bibliographical management methods can be applied to Web archives by cataloguing resources just as printed books and journal articles. This is a useful way to integrate Web archives with a library's existing collections so that they become discoverable through catalogue search. However this approach is resource intensive and difficult to scale up to apply to Web archives, because of the large amount of objects they contain as well as the challenge in defining the resource to be catalogued. Cataloguing may take place at a higher level of granularity, for example at the level of special collections rather than that of individual Websites.

3.2.1.5 Resource discovery tools using metadata (optional)

Access may be provided to resources through the addition of metadata associated with the resource. Websites may be classified by curators or through automatic means into subject hierarchies or into collections around events or thematic based collections. The usage of tags (user added keywords) whether added by curators or by the public can also be built into the user interface.

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013

3.2.2 Technical baseline: access methods

3.2.2.1 General

Access to archived Web material is provided through the use of specialist software to find, retrieve and replay archived Web material. This may be implemented by a number of separate software packages working together. The whole software system is commonly called an Access Tool.

Regardless how an Access Tool is designed and implemented, it has a set of common attributes, some mandatory, some optional:

3.2.2.2 Rendering (mandatory)

Access tool software has to be able to uniquely identify resources (even if the same resource has been harvested numerous times) and retrieve that object from the archive repository.

3.2.2.3 URL re-writing (mandatory)

The HTML pages returned by the access software have to be modified from their original manifestation. Embedded links (absolute or relative) should point to the location of the resource within the digital archive, not to the location of the original resource. This can be achieved in a number of ways:

— This can occur at harvesting time (a curatorial decision may be taken to re-write the URL within the content immediately and ingest this modified content into the archive repository).

— A preservation action upon the archived resources which achieves the same goal as indicated above may be undertaken at a later stage. The resource therefore implicitly points links to their new location.

— Runtime URL re-writing can be achieved by code executed on the server upon request of a resource or by supplying a copy of the original resource to the client along with code to be executed by the client to dynamically re-write the URL.

3.2.3 Limitations

The process of harvesting and processing archived Internet resources involves transformations which may impact the appearance, behaviour and user experience of the original resources when they are replayed at the point of access. Archived copies of resources should be regarded as snapshots frozen at points of time, which loses the interactivity of the live versions. Examples of these include message boards, discussion forums, Web forms and search. It is also possible that a resource is harvested properly but limited by the capability of the rendering software and consequently becomes inaccessible to the end users.

A common problem in replaying archived resources is the so-called “live Web leakage”, which occurs when links in an archived resource resolve to the current copy on the live site, instead of pointing to the archival version within the Web archive. This is usually caused by incorrect URL-rewrite, often a result of links embedded in JavaScript not being detected by the access tool.

3.2.4 Access strategies

Internet resources, despite many being freely available, are generally copyrighted. Depending on the relevant legislations (see 4.4) and what is permitted legally, collecting institutions employ a range of access strategies:

— Dark archive: collections cannot be accessed by anyone (except sometimes by staff for curatorial purposes).

— Grey archive: collections can be viewed only by authorized end users (e.g. researchers), and/or restricted to on-site premises (e.g. library reading rooms).

— Online archive: access for all users, usually accessible from a collecting institution’s Website.

Documents
Document Sharing Hub

© ISO 2013 – All rights reserved

Full document details: PD ISO/TR 14873:2013

Alternatively an archive could employ a mixed model, with parts of the archive adopting one of the above strategies separately. It is also worthwhile to note that access to resources harvested on the so called "opt out" basis, where permission is only assumed or implied, not expressly given, could be taken down when requested by the rights holders.

3.3 Preservation methods

3.3.1 Technical baseline

Preservation of analogue material such as books or records focuses on preserving the original items while digital preservation deals with very different issues. Digital resources at the lowest level consist of 1's and 0s (the bit stream), which are independent of the data carrier or media on which they are stored. It is possible to copy the bits to another carrier without losing information and create copy identical to the source or the original. Because data carriers deteriorate or become obsolete, there is a need to move the bits to newer carriers to keep them safe. If the copying is done regularly, it is reasonable to assume that the bit stream is permanently preserved without any loss.

In addition to keeping the bits safe, the real challenge for digital preservation is to keep the bits usable. The bit stream is not understandable for human beings until rendered using the originally intended software and hardware environment. With rapidly evolving technology, new systems are often not compatible with older systems and newer rendering software might not be able to interpret older file formats. Even if it is possible to render older software in current systems, current users may still not be able to use it as users could be expected to interact with it in completely different ways.

Digital preservation should be considered in every step of the Web archiving workflow. In comparison with other digital resources, the particular challenge in preserving Web archives is the large amount of data and the diversity of file formats and media types. Web pages can contain pictures, videos, music, games, databases and many types of applications. A key characteristic of the Web is the linkage between Web pages, which can pose digital preservation challenges due to the dependencies introduced by the links.

3.3.2 Limitations

Web archives contain still recent resources and there is a lack of proven strategies with convincing results or confidence to demonstrate the community's long term preservation capacity for archived Internet resources. This Technical Report is not intended to offer practical solutions but highlights the current practice, standards and issues.

3.3.3 Preservation strategies

The aim of digital preservation at the minimum level is to prevent data loss by maintaining the integrity of the original bit stream. The main strategy for bit stream or physical preservation is mainly duplication and backup, including actions such as parallel data storage at separate physical locations, regular backups and checks for read errors. Data also need to be kept secure to avoid unauthorised access. Bit stream preservation is a minimum requirement which applies to all digital resources. The scale of Web archives however needs to be taken into account when implementing bit stream preservation.

Migration and emulation are more sophisticated preservation strategies intended to preserve the functionality, behaviour and user experience of resources. They are referred to as "logical preservation" and require regular data, format and risk analysis to implement. Logical preservation is extremely challenging for Web archives due to their large scale and multiple file formats.

a) Migration

File format migration involves converting a file to a new format before it is no longer usable within the current technical environment. Every conversion to a new file format changes the content and may cause damage. A risk analysis is therefore required in advance to assess the probability and impact of information loss. Migration can be performed when a file format runs the risk of becoming obsolete or at the point of access (on-the-fly migration). The cost of migration is directly related to the number of files

© ISO 2013 – All rights reserved
Documents
Document Sharing Hub
Full document details: PD ISO/TR 14873:2013