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 ..... **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**

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.

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.

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

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.

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.

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

to be migrated. It could be very expensive to migrate for large scale Web archives. The complexity and dependencies between resources also add to the challenge and make it hard to validate the migration.

b) Emulation

Emulation involves recreating the functionality and behaviour of an obsolete system environment on the current system using specialist software called emulators. The emulator mimics the obsolete system and makes it possible to access out of date resources without changing them. Emulation however is hardly perfect and only an approximation. The emulator itself is dependent on a certain system environment and subject to preservation risks. Developing emulators is costly but removes the need to deal with component resources individually. In Web archiving, an emulator needs to recreate the functionalities of common browsers and media players of the time when the Web pages were harvested.

Migration and emulation should be considered as a part of preservation planning for digital archival systems.

3.3.4 Preservation metadata

Long term preservation also includes keeping safe the metadata associated with the resources in the Web Archive, which are critical for supporting collection management, access and preservation activities. There are different types of metadata, which could be embedded within the resources, generated automatically during the archiving processes or manually added by curatorial staff. The Metadata Encoding and Transmission Standard (METS) defines 5 different types of metadata and they apply to Web archives as follows:

a) Descriptive metadata

Institutions which catalogue Web archives or manually add metadata generally possess more descriptive metadata. Those performing (automated) large scale Web archiving have to rely on extracting metadata embedded within the resources or use automatic clustering or classification to obtain such metadata.

b) Structural metadata

Internet resources are often compound objects made up of structured and interlinked elements. Their structural relationships can be expressed explicitly and recorded in metadata schemes such as METS. Such metadata can be useful in the case of file migration, where hyperlinks need also to be migrated accordingly in order to maintain the navigability of the archive. Some institutions decide not to record these relationships additionally or explicitly as they are intrinsically present within the resources.

c) Provenance metadata

Provenance metadata describe why and how a resource is created and what happened to it during its lifecycle. Some descriptive metadata, for example recording the rationale of a special collection within a Web archive, can also be regarded as provenance metadata. Provenance metadata can also be found at lower levels, include files recording activities of an archiving crawler, such as configuration files, crawl reports and log files, and information describing the interactions between the Web server and the crawler, including the URL, the crawl date, and the IP address of the server.

d) Technical metadata

Technical metadata describe the technical characteristics of a digital object, specifying how it can be accessed, modified, or preserved. This is referred to as representation information in the Open Access Information System Reference model. File format, indicated by MIME types, is an example of technical metadata relevant to Web archives and is one of the core statistics for collection characterization (see 4.3.2.3 for details).

e) Rights metadata

Rights metadata define the ownership and the legally permitted usage of resources. Conditions could be applicable to a time in the future. This information needs to be preserved together with the resources to prevent unauthorised access.

Provenance metadata, technical metadata and rights metadata are collectively referred to as administrative data.

3.4 Legal basis for Web archiving

3.4.1 General

Web archiving initiatives face many legal risks. The key ones relate to intellectual property, particularly copyrights, privacy and protection of personal data. Collecting institutions could also be held responsible for republishing libellous content and for possessing and distributing illegal material. National legislation address these risks effectively by providing the collecting institutions with certain legal protection.

National legislation is the most effective framework to enable and support Web archiving at scale. Web archiving may be introduced by legislation on copyright and/or legal deposit or any act which specifically defines the missions and status of a collecting institution. This legislation is particularly relevant to public organisations, whose status and activity are often defined by law, such as national libraries or archives, specialist legal deposit institutions (e.g. those dedicated to the preservation of broadcasts or films), public archives or museums. Organisations without a legal remit may still undertake Web archiving. They either negotiate specific agreements with right holders, or restrict access to the archived resources to manage the legal risks. Some decide to take certain risks by capturing and providing access to Internet resources based on implied permissions.

Some countries still do not have a clear legal basis for Web archiving, others have only a general framework which requires secondary legislation to interpret and regulate its implementation. Collecting institutions in some countries have developed Web archiving initiatives on the basis of voluntary deposits by publishers. Even in countries where national legislation is in place, a common practice is that legislation is often open for interpretation. Collecting institutions need to define their own approach including risk evaluation and experiment at the implementation level.

3.4.2 Collecting scope and methods

Legislation for Web archiving may explicitly include or exclude certain content. They define the boundary or territoriality for a national domain and may also specify frequencies or depth of the permitted harvesting.

A key element of the legislation is whether permissions should be obtained from the rights holders prior to the resources being harvested. This has a significant impact on an institution's collecting strategy. Bulk harvesting is only feasible if no permission is required, otherwise selective harvesting is a more appropriate model. Alternative approaches to permission management include the so called "opt out" or "notice and take down" model, where resources are harvested and made available on the basis of assumed or implied permissions, and could be taken down when requested by the rights holders. "Blanket permission" from a single publisher, covering multiple resources, is another way to reduce the costs of managing multiple agreements with publishers.

Legislation may grant a collecting institution the right to obtain protected information to help improve the quality and completeness of harvesting. Examples of such information include lists of national domain names or identification codes and digital right management information for priced publications. Legislation may specifically oblige publishers to provide the information and impose penalties for non-compliance.

Legislation may also recommend or mandate a specific collecting technique. Some for instance explicitly allow or encourage "automated" harvesting of Internet resources pending mutual agreement between the institution and the publishers on the harvesting protocol.

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.

In countries where a legal mandate exists for collecting all national Internet resources, this is typically the responsibility of a single institution. This could also be a shared responsibility between several organisations, for example:

— between national or federal institutions and local or regional institutions;

— between national libraries and national archives;

— within a network or a consortium of specialized institutions.

3.4.3 Access to Web archives

Accessibility defines the conditions under which a Web archive can be used and is an important aspect of the legislation. Access conditions are in general consistent with harvesting regulations; if permissions from rights holders are required, online access may be permitted accordingly; if bulk harvesting takes place without permission, access is more likely to be restricted.

Copyright restrictions, such as print, extraction, electronic copying or downloading, may apply to Web archives.

Where Web archiving is implemented as a form of legal deposit, legislations may also require that a national bibliography is published for the Web archives. This proves to be most challenging for large scale collections. Instead of publishing descriptive metadata, Web archives normally offer search interfaces for end users.

3.4.4 Preservation of Web archives

Long term preservation is the key justification and requirement for collecting cultural heritage material. Legislation may therefore include an indication or the obligation to ensure the longevity of Web archives. It may in particular specify whether it is permitted to delete resources or whether they should be preserved permanently. Most national libraries collecting Internet resources under legal deposit would be required to preserve these for posterity. Research libraries and other institutions may not be required to preserve resources indefinitely if the purpose of collection is to provide data sets for short or mid-term research.

3.5 Additional reasons for Web archiving

3.5.1 General

Other motivations for archiving the Web are usually policy-driven. They reflect an institution's strategic vision as well as its convention and attitude towards technological and cultural innovations.

The Web hosts a diverse range of born-digital and digitised resources. The latter used to be printed (books, periodicals, Government publications, etc.) or carried on other physical media (film, music or games on disc or tape), many of which have already been through various formats migrations. The Web evolves fast and is ephemeral. Valuable resources disappear regularly. Preserving the Web is seen as natural and critical for institutions with a long standing mission of preserving cultural heritage material.

Web archiving ensures digital continuity and is a necessary action to avoid a digital black hole in the knowledge and memory of a nation. It can also help maintain access to cited research. This motivation is particularly strong for national libraries and archives.

3.5.2 Facilitating academic research

The Internet is a highly participatory and innovative space where people communicate and collaborate. It can be argued that the Web creates new social knowledge and new artefacts of research value which are pertinent to national heritage. One can already observe the emergence of new research practices and communities dedicated to the study of the live Web and, potentially, of its archives.

For Internet researchers and scientists, Web archives offer unmatched research possibilities. They cannot only consult historical versions of individual Websites, but also perform large scale data or link mining which can help extract patterns and trends and unlock embedded knowledge. Although data mining or analytics is still in its infancy and mostly explored by social scientists, it is likely to expand towards other academic areas and become useful in many disciplines.

At an institutional level, Web archiving can be a way to promote or highlight specific digital resources. This is particularly applicable to institutions such as universities, which archive the publications produced by their own scholars and students. Web archiving is a valuable effort to collect online resources which contain and are of research value. This is a compelling motivation for many Web archiving institutions.

3.5.3 Supporting various types of usage by the general public

The Internet hosts contributions from all walks of life. Unlike printed media, anyone can publish on the Web. Although the value of each individual contribution may vary, the aggregation forms a unique set of resources which reflect individual as well as community memories and interactions.

A variety of professional or private services can be built over Web archives. They could serve as evidence in a copyright dispute or be used for personal or family research and digital genealogy. Allowing free, long-term access to Internet resources, especially to content creators themselves now and their descendants in the future is a strong argument to offer Web archives as a public service.

4 Statistics

4.1 General

Statistics are objective data which provide the basis for further analysis and interpretation. Quality indicators denote a degree of value judgement: not meeting the criteria is an indication of negative assessment. In this Technical Report statistics are generally measured in absolute numbers and quality indicators in relative numbers and percentages.

Statistics and quality indicators need to be reliable, informative and comparable and the methods for obtaining them need to be practical and flexible. The current state of the art in Web archiving means that the production of some statistics has to rely on the use of harvesting, indexing or browsing software, and a different choice of software may lead to variance in the results. It is therefore recommended to use the same software to generate statistics if benchmarking is the purpose of measuring. The large scale of Web archives in general also requires that it is practical and cost-effective to obtain the statistics and measure the quality. The quality indicators and statistics presented in this Report are based on common practices of Web archiving, which over time will still offer a reliable overview of Web archives and allow comparison between them.

This report proposes generic statistics and quality indicators. Not all will be applicable to different types of Web archives. In addition, as technology progresses and Web archiving practice advances, some of them will need to be updated.

This chapter of the Technical Report proposes and describes a number of relevant statistics in each section. However a smaller set of core statistics are regarded as essential. These are presented in the table at the end of each section, with examples.

4.2 Statistics for collection development

4.2.1 General

The following statistics measure the growth of a Web Archive by keeping track of its quantitative outputs. These help plan and monitor collection development and enable detailed cost analysis.

Unlike analogue documents, Web archives contain nonlinear, interlinked resources. Some are intended to be replayed for human users, others are inseparable files and metadata which are part of resources but

not visible to the users. Statistics measuring the volume of a Web archive are therefore not comparable with those for physical materials.

Most of these statistics are specific to archived Web resources and are not applicable to live Web content.

4.2.2 Measuring collecting goals: counting targets and captures

4.2.2.1 Purpose

Collecting institutions should be able to express and evaluate the objectives of Web archiving in light of their policy so that the outcome of the resulting Web archive collection can be assessed against the goal, indicating the level of achievement and the effectiveness of the collecting process.

There is no straightforward or unified method to express goals or targets for collecting. This is in practice defined by individual institutions based on institutional policy and objectives. A generic framework is proposed which suggests the use of the concept of "target" and "target capture", which can help measure the activity of staff involved in selecting and managing the targets. This is an indicator of the selection effort, which is the time spent to determine which websites to include in a Web archive.

4.2.2.2 Method

A target consists of one or more seeds and each of the seeds also has a set of crawl settings which define its scope. It is a meaningful set of resources to be collected and its scope can range from interlinked resources hosted on the same domain, presented as a Website, to a single resource which is identified by a URL (e. g. a PDF, a video) or an entire Top Level Domain. A target can be crawled more than once. Each crawl is a capture of a target.

EXAMPLE The daily crawl of the homepage of the New York Times Website.

— the seeds may be http://www.nytimes.com/ and http://global.nytimes.com/;

— the scope may be "crawl the homepage and all resources at one click from the homepage";

— the frequency may be "every day";

— the target is all of the above;

— the single set of crawled and stored resources of a target is a target capture.

This generic method allows institutions to set operational goals and measure the outcome. In the above example, the goal could be collecting 365 target captures per year. One can then compare the actual number of target captures on an annual basis to assess if the goals have been achieved.

4.2.2.3 Limitations

Comparing the number of targets and target captures between institutions is only meaningful if the institutions share common selection policies and practices.

4.2.3 Measuring the size of a Web archive: counting URLs

4.2.3.1 Purpose

Counting URLs is one way of measuring the size of a Web archive. URLs correspond to the location of resources to be crawled and are used by archiving crawlers to identify and request resources from a Web server. The Web server returns a range of standard responses identified by status codes, indicating the status of the requested resource. It may be a response confirming that the requested resource has been successfully delivered, or a response indicating that the requested resource has moved to another location (redirection). It may also respond with an error message, indicating that the requested resource is not available. Some responses provide both metadata and content, some only metadata, some just an error code.

It is important to realize that not all URLs correspond to meaningful, human-readable resources, the equivalent of physical "documents" or "items" as in traditional printed collections of a library. Even when a Web server fails to satisfy the requests, it still delivers a response which indicates the delivery status of a request, such as redirects and errors. Such information provides an audit trail of the harvesting process and provenance information to the Web archive collection, and can be very useful for access or preservation purposes. This Technical Report therefore recommends that all responses are kept and regarded as part of the Web archive.

URLs are used as identifiers for resources on the Web and in the HTTP message system. URLs also represent the smallest unit of self-contained content in a Web archive and are commonly used in storage and access systems for archived Web resources. We therefore propose that URLs are used for identifying resources and the corresponding responses returned by the Web server. The different types of status codes can then be used to sort or group resources in the Web archive.

4.2.3.2 Method

Table 1 contains the various status codes, which are 3-digit numbers of which the first digit defines the class of the responses. Each status code also has a reason phrase, intended for the human user and gives a short textual description of the status code. More details on the status code can be found in the RFC 2616 (see bibliography).

Table 1 — List of http status codes

Status code	Reason phrase
1xx	Informational
100	Continue
101	Switching Protocols
2xx	Successful.
200	OK
201	Content was created
202	Accepted but not acted upon now
203	Non-authoritative information
204	No content
205	Reset content already sent
206	Partial content
3xx	Redirection
300	Multiple Choices
301	Moved Permanently
302	Found
303	See other
304	Not modified
305	Use proxy
307	Temporary Redirect
4xx	Client error
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden

Table 1 *(continued)*

Status code	Reason phrase
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Time-out
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Large
415	Unsupported Media Type
416	Requested range not satisfiable
417	Expectation Failed
5xx	Server error
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Time-out
505	HTTP Version not supported

This Technical Report recommends that all URLs are included when calculating the total number of (harvested) resources in a Web archive, regardless of the corresponding status codes.

It is however important to understand the nature and meaning of the status codes as they can be used to group or filter resources for the purpose of analysing particular segments of the Web archive. The 2xx series status codes for example indicate successful delivery of requested resources and responses with the 3xx series usually only return metadata without the requested resource. The 5xx series are intended for technical use. The 2xx series are of particular interest if an institution wishes to maintain analogy between its physical collections and the Web archive. The recommendation is to count specifically the URLs with status codes 200, 201, 203, 205 and 206.

The number of URLs may be calculated before and after de-duplication, if a process de-duplication occurred. De-duplication occurs, during a crawl, when a robot recognizes that an URL about to be crawled has already been collected and is already available in the archive; so the robot does not crawl it again. The robot may generate information indicating this de-duplication process; it is called a “revisit” record in WARC. Both numbers are useful:

— the number of URLs after de-duplication represents the amount of resources in the archive. It is the reference number used for storage provision and for long-term preservation;

— the number of URLs before de-duplication is intended for human users and significant from an intellectual or content perspective. It is only interesting if information on the de-duplication process is still available (e.g. thanks to revisit records in WARC files). This information indicates indeed that the de-duplicated URL was still online at a certain date, even though the robot did not actually crawled it.

The calculation method (i.e. before or after de-duplication) should always be indicated, especially in case of comparison between archives.

4.2.3.3 Limitations

The Web in the early days mainly consisted of static HTML pages with explicitly referenced resources. It however has evolved rapidly and there is an increasing amount of interactive and dynamically generated content on the Web, which requires us to think beyond the traditional model of the Web as a collection of self-contained HTML "documents" or "publications". The number of URLs in a Web archive does not equal the number of "documents" or "publications" in what these terms conventionally represent in the context of a library. When calculating statistics for Web archives, it is important to take into account the nature of the Web and think of it as networked and interlinked online resources. The statistics should not only include the resources intended for the human user but also the associated metadata and programs which are integral parts of Web archives.

It is also important to realize that not all status codes returned by Web servers are reliable or trustworthy. Below are some examples.

a) Missing 404

Many Web servers do not return a proper 404 status code when the requested resource is unavailable. The server may send a "200 OK" response instead, with a content block explaining that the requested resource does not exist. There is no way for the archiving crawler to know that this should be regarded as "404 Not Found", so it will be counted as a "good" or successful response.

b) Duplicates with different session ID

Many Web servers generate URLs dynamically, leading to duplicates of resources in the Web archive. A Web server sometimes appends a unique identifier to each URL when returning resources to the user-agent, In order to keep track of a session, for example:

User-agent 1 gets a URL that looks like http://www.example.com/id=12345/picture.jpg

User-agent 2 gets a URL with a different ID: http://www.example.com/id=67890/picture.jpg

The two URLs serve the same resource, in this case a jpeg image, but use different IDs to identify the users. The archiving crawler could be collecting the same resource many times but with different URLs and these duplicates would be counted as unique resources.

c) Absence of status codes

In the early days of the Web, it was common for servers to return just the requested resources without any status code and metadata. This is sometimes referred to as HTTP 0.9. Some servers may still use the dated protocol. The lack of status code and metadata may particularly be an issue for institutions holding long standing historical collections of the Web.

4.2.4 Measuring the size of a Web archive: counting domains or hosts

4.2.4.1 Purpose

Counting domains or hosts is another indicative way of measuring the size of a Web archive. This is commonly used by collecting institutions and in practice often as a substitute to mean the number of Websites. A Website on the contrary is a conceptual intellectual unit which consists of a set of interconnected Web pages, representing a person, a group or an organization. It is however not something that can be defined technically nor allow for practical or systematic measurement.

Domains and hosts are measurable systematically but should not be seen as equivalent of Websites themselves as they are merely used to name and locate Websites. These statistics may be useful for detailed collection characterization or technical analysis, to determine the type of resources in a Web archive (e. g. .com or .org) or whether the intended scope has been followed by the crawler.

4.2.4.2 Method

Number of domains or hosts can be calculated automatically from the crawl reports or by other automated means of analysing the stored files.

4.2.4.3 Limitations

Counting domains or hosts has limitations. As with counting URLs, this tends to generate higher numbers than the amount of actual visible and human readable resources in a Web archive. Not all domains resolve to active or meaningful resource and there are also aliases or duplicates.

a) Inactive domains

A domain name can be purchased but inactive, which does not resolve to any resource. There are also parked domains which do resolve to resources but often consist of a single Web page offering domain names for sale. The former is identifiable through the status code 204. The latter is strictly speaking not inactive and will be included when counting domains. However such resources are not regarded as containing any significant value from the curatorial perspective. In the context of selective harvesting, they can be actively excluded during the selection process and not harvested. However there is no easy way to automatically identify and filter them from bulk harvests, apart from manually checking resources of very small size.

Where possible, it is suggested to keep track of inactive domains in a Web archive through sampling, to help characterize the collection and evaluate the effectiveness of quality assurance.

b) Aliases

An alias is an alternative domain name. Domain aliasing allows one to host a Website on one domain and point other domain names to it. There are many aliases on the Internet because domain owners may want to use several domain names in order to be as visible as possible to users. Aliases are mainly implemented through redirection.

In reports generated by archiving crawlers, aliases are unique domains despite pointing to the same resources. Detection of aliases requires visual comparison between pages coming from the same server or comparison of checksums. Aliases are easier to detect or less relevant in selective archiving and more likely to be included in bulk harvests and will result in duplicate resources in a Web archive.

Where possible, it is suggested to keep track of detected aliases in a Web archive to help characterize the collection and de-duplicate resources.

4.2.5 Measuring the size of a Web archive: counting bytes

4.2.5.1 Purpose

The size of a Web archive can also be measured in bytes. This is a useful statistic which can help plan storage and other resources. It is comparable to the linear metre or miles for the management of stacks in a library.

4.2.5.2 Method

Size of an archive in bytes can be generated automatically by adding up the size of crawled resources from crawl reports or by other automated means which examine the disk occupancy of the archive.

Web archives are often of large scale. Their sizes range from small collections of a few hundred gigabytes to national collections of a few hundred terabytes. For those collecting the global Web and holding a long-standing archive, the size of the archive could even reach petabytes. A common practice to store Web archives is to compress the data. Annex D of the ISO 28500 WARC file format specification, the standard archival format for Web archives, for example explains how to use GZIP compression for WARC.

The size of a Web archive can be measured both compressed and uncompressed. When benchmarking archives, it is however important to use the same criteria, i. e., one should not compare the compressed size of one Web archive with another which is uncompressed.

— the compressed size represents the disk occupancy of the resources. It is the reference size used for storage provision and for long-term preservation;

— the uncompressed size represents the volume of resources as they were on the live Web. This is intended for human users and significant from an intellectual or content perspective.

The size of a Web archive can also be measured both before and after de-duplication, for reasons explained in 4.2.3.2. The calculation method should again been clearly indicated.

It is also a common practice to store files in container files such as ARC or WARC files. Containerisation aggregates files and makes it easier to store and handle a few larger files rather than numerous small files. Container files generally allow the storage of metadata along with the harvested resources. The number of container files is also a useful statistic for Web archiving as they are often used as basic management units for storage, data exchange, and sometimes for long-term preservation purposes.

4.2.6 Core statistics for collection development

Table 2 — Core statistics for collection development

Statistics	Purpose	Example
Number of targets	Collecting goal/quantitative outputs	8 000 targets
Number of target captures	Collecting goal/quantitative outputs	14 000 target captures
Number of URLs (before and after de-duplication)	Quantitative outputs	14 billion URLs harvested, 10 billion after de-duplication
Distribution of URLs by status codes	Number of resources by type	2 million successfully crawled resources (code 200)
Number of domains or hosts	Quantitative outputs	3 million domain names
Size in bytes (uncompressed and compressed, before and after de-duplication)	Quantitative outputs	200 terabytes uncompressed before de-duplication, 160 terabytes compressed and after de-duplication
Number of WARC or any other container files.	Qualitative outputs	18 000 WARC files

4.3 Collection characterization

4.3.1 General

Statistics proposed in this section describe the characteristics of Web archives and are useful to help scope Web archives and make informed curatorial decisions. While some statistics are specific to selective or bulk harvesting, others are generic and applicable to Web archives established using both strategies.

The scale of Web archives generally precludes manual counting. Some statistics can only be gathered through sampling, especially those specifically related to bulk harvesting. Manual collection of statistics may be possible for selective archives but should only be performed without spending unwarranted resources.

4.3.2 Common statistics

4.3.2.1 Distribution by top and second level domains

4.3.2.1.1 Purpose

The top level domains (TLDs) indicate the geographical distribution of the resources in a Web archive. This statistic is of particular interest to national libraries and archives who have the remit to preserve an entire country's intellectual output. Second level domains, intended for specific categories of organisations for areas of interest, are also useful as they reveal the broad nature of resources in an archive. Resources under a gov.uk domain name for example are published by a UK government organization.

4.3.2.1.2 Method

Distributions of TLDs and second level domains can be calculated automatically from reports generated by archiving crawlers or other automated means of analysing the domains. They can be measured in absolute numbers or in percentage. It could also be useful to list the top 5 or 10 of the most occurring TLDs in a Web archive.

— The number or percentage of national TLD collected: 70 % of URLs within the National Library of France's latest domain crawl are hosted on .fr domain names. 3 % of URLs within the National Library of France's latest domain crawl are hosted on .de domain names.

— The number or percentage of second level domains collected: 1,5 % of URLs within the National Library of France's latest domain crawl are hosted on .gouv.fr domain names, which are resources published by French government organisations.

A Web archive containing a greater proportion of national TLDs than other domains can be considered to have a national scope.

4.3.2.1.3 Limitations

Some institutions consider resources hosted outside their national TLDs as in scope. For example, www.lego.com (Danish company) is considered a Danish Website although it uses a non .dk domain name. This shows that TLDs are not always sufficient to define the scope or boundary of a national domain.

4.3.2.2 Distribution by volume of resources per domain (and / or host)

4.3.2.2.1 Purpose

Analysing and reporting on the size of resources hosted under each domain and/or host and how this is distributed across a Web archive not only provides insight into the characteristics of the collection but also helps manage the crawling process.

Distribution of the size by domains and/or host across an archive can reveal the characteristics of resources hosted under certain types of domains. It is also an indication of the archive's capability of collecting resources of various sizes, especially large Websites which are technically challenging to crawl.

Grouping resources of different sizes by domain and/or host also helps configure and organize the crawling process. A common practice is to group and crawl domains of similar size as a separate process or "job" as these require similar settings and take a similar amount of time to complete. This makes best use of machine resources and eases the monitoring and management tasks.

4.3.2.2.2 Method

The volume of resources per domain and/or host can be measured in MB/GB/TB:

— < 10 MB

— < 100 MB

— 101-999 MB

— 1 GB

— > 1 GB

Alternatively one can also count the number of URLs per domain:

— < 10 000 URL

— 50 000-100 000 URL

— > 100 000 URL

The only way to ascertain the volume of resources per domain on the live Web is using figures provided by search engines. For a Web archive, the above statistics can be calculated automatically from the crawl reports or by other automated means of analysing the stored files.

4.3.2.2.3 Limitations

As with some of the other statistics proposed in the previous section, there is a level of approximation attached to this statistic. They are more useful when measured over time so that comparison can be made.

4.3.2.3 Distribution by format type

4.3.2.3.1 Purpose

Analysing and reporting on the distribution of file formats in a Web archive is a key digital preservation activity, as well as an element of archive characterization.

In order to monitor and manage the preservation risks related to format obsolescence, it is essential to gain knowledge of the types of files present in the archive.

Format information is comparable to the high level classification traditionally used by libraries to categories of publications, such as image, film and sound recording. Collecting this information over time can also reveal trends in technology and help us understand the evolution of the Web.

4.3.2.3.2 Method

Format statistics can be calculated automatically from the crawl reports or by other automated means of analysing the stored files.

Distribution of format types can be calculated and organized in different ways, for example:

— by resource type: 70 % of files are text (e.g. html), 15 % are image (e.g. jpeg and gif), 3 % are audio (e.g. mpeg);

— by most common file formats (top 50 or 100): e.g. html is the most common file format;

— by tracking certain formats, starting from its first appearance in the archive and showing increase or decrease over time;

— by least used format type: e.g. if video formats in a given Web archive are less significant than on the live Web, it may be an indication of their under-representation in the archive.

4.3.2.3.3 Limitations

Analysing the format profile of a Web archive normally generates a long list of multiple file formats, containing thousands of items. It is suggested to focus on the 50 or 100 commonly used formats unless one is interested in a particular non-common format.

Formats of crawled resources (MIME types) reported by Web servers and recorded in crawl logs are not always reliable. Web servers may return faulty MIME types. Some institutions use additional format identification tools to obtain more accurate information.

4.3.2.4 Language characterization

4.3.2.4.1 Purpose

Analysing the languages used in archived resources provides insight into the linguistic patterns of a Web archive and can be useful in understanding characteristics of a national collection including its diversity or cultural proximity to other countries. Language characterization is most relevant for nations where unique languages are spoken as it can be used to help identify resources on the Web relevant for those countries.

Resources on a national domain also use foreign languages, such as Basque on French domains and Arabic on Danish domains. Language characterization can help analyse a variety of social and cultural issues and their reflection on the Web.

4.3.2.4.2 Method

There are no standard methods or technologies for automatic language detection of archived Web resources. Structures specific to a natural language can be used to help analyse and identify resources using that language. If available, one can also make use of the language-related elements in HTML and HTTP headers. There are also language detection or natural language processing tools which can be used for this purpose.

The distribution of natural languages in a Web archive can be presented either as absolute numbers, e. g. the total number of pages using a particular language, or as percentage, e.g. the percentage of pages using a particular language in the archive. It could also be useful to list the top 5 or 10 most represented languages.

4.3.2.4.3 Limitations

Many resources lack language-related metadata which makes it hard to identify automatically the languages being used to construct the resource. In addition, automatic translating programmes that allow users to view Web pages in multiple languages could distort this statistic.

4.3.2.5 Chronological coverage

4.3.2.5.1 Purpose

Web archives can be characterized by their chronological coverage, which refers to the time span during which the resources are collected. A general assumption is the longer the time span, the more likely it is that the original resources of which copies exist in the Web archive would have disappeared from the live Web. This adds value to the archive as it could hold unique historical records of certain time periods or historical events. Chronological coverage of the archive is also essential information required for preservation planning. Archives covering longer time spans are more prone to risks of obsolescence.

4.3.2.5.2 Method

The chronological coverage of a Web archive can simply be measured by counting the number of subsequent years after the date when the very first resource was captured. This can be used in combination with other statistics, such as the size of the archive, or distribution of file formats, to show trend or development of the archive over time.

4.3.3 Statistics for selective archives

The following statistics are only relevant and applicable to archives collected by means of selective harvesting.

4.3.3.1 Permissions

4.3.3.1.1 Purpose

When permissions from publishers or holders of rights are required prior to their Websites being harvested, statistics about permissions become important indicators of the efficiency of this particular workflow. When compared with the costs of maintaining a permission management system, permission statistics can for example reveal the unit cost of each successfully obtained permission.

The number of permissions also reflects the publishers' interest and their general awareness of Web archiving.

4.3.3.1.2 Method

It is useful to count the number of granted permissions, as well as the number of requests for permissions sent to the holders of rights. This can be done manually or using automated functions available in a permission management system or a curator tool.

4.3.3.2 Nominations

4.3.3.2.1 Purpose

Nominations are proposed candidate resources for inclusion in a Web archive. Nominations can come from a wide range of sources, including enthusiasts, users and supporters of a Web archive and the general public. Some Web archives proactively solicit nominations by using an editorial board, social networks, or online nomination pages. The selection process determines whether a nomination is in scope for a Web archive, judged against its collection development policy. Selection is usually performed by subject librarians, digital archivists or curators.

Counting the number of nominations reveals the impact and awareness of Web archive activities among the stakeholders and can be used to guide engagement activities. It also measures productivity and efficiency related to the selection process, and can be useful to assess performance and effort of individual digital curators (e. g. the number of nominations per curator). In case an institution employs a mixed strategy of bulk and selective harvesting, comparison can be made between the two, highlighting the costs related to manual selection versus the more automated approach of bulk harvesting. Further statistics can be generated to examine how a Web archive is used, whether the curator-selected content is used more and thus perceived to contain more value than automatically selected content.

4.3.3.2.2 Method

The number of nominations can be collected manually or using automated functions of a curator tool.

4.3.3.2.3 Limitations

Not all nominations lead to the actual targeting or capture of contents. Many factors impact the outcome. The permission required to archive the Website may not be obtained successfully; the crawler may not be able to capture the content due to technical reasons; an institution may not be able to afford harvesting all selected resources due to financial constraints. Even in cases where a resource is harvested successfully, issues related to access could make it impossible to be replayed from the Web archive.

This above statistic measures the selection effort, rather than the outcome.

4.3.3.3 Subjects

4.3.3.3.1 Purpose

Subject coverage of a Web archive characterizes its content and is considered extremely valuable for collection development as it helps profile and balance the collection and reveal possible gaps in the content. It could also be used to understand whether and how a Web archive meets the requirements of researchers of various disciplines.

4.3.3.3.2 Method

There are various ways of extracting subject information. Resources in some Web archives have been manually assigned subject terms by curators, which can then be grouped, filtered and analysed. Standard classification systems such as the Dewey Decimal Classification (DDC) or the Library of Congress Subject Headings (LCSH) are commonly used to describe archived resources.

Most Web archives do not have manually added subject terms. Subject information, if available, is then obtained by extracting metadata embedded in the resources, such as values of the "keyword" meta tag in the HTTP header or from the Dublin Core "subject" field. Automatic clustering or classification can also be used to semantically analyse resources in a Web archive and automatically classify them based on a subject scheme. There is currently not an agreed or common practice of doing this in the Web archiving community.

The percentage of the resources per subject can be listed to help understand how the content of a Web archive is distributed over various subjects. Consequent effort could be put in place to address under-presented subjects. Subject information can also be used in combination with other statistics or information, for example with provenance metadata to understand the relationship between subject areas and the type of publishers.

4.3.3.3.3 Limitations

Manually assigning subject terms more accurately describes or classifies archived resources but is costly to implement. Automatic extraction of embedded metadata relies on the metadata being available. Automatic clustering or classification is not yet a developed area for Web archiving and is technically challenging, especially for large scale Web archives.

4.3.4 Core statistics for collection characterization

Table 3 — Core statistics for collection characterization

Statistic	Purpose	Example
Distribution by top level or second level domains	Geographic distribution	10 billion URLs under the .fr TLD in the archive
Distribution by volume of resources per domain	Domain analysis	2 million domain names have less than 10 URLs, whereas 150 000 domain names host more than 10 000 URLs
Distribution by format types	Format characterization	500 million URLs of the last bulk crawl are in html
Chronological coverage	Temporal analysis	The archive contains resources collected from 1996 to date
Number of granted permissions	Productivity	20 000 permissions requests result in permissions granted by the publishers
Number of nominations	Productivity	1 000 new nominations were added in a year

4.4 Collection usage

4.4.1 General

As described in 3.2. *Access and Description Methods*, the conditions of description and consultation of archived Internet resources differ according to national legislations and institutional policies, the distinction between white, grey and dark archives being the most useful to choose the appropriate methods for measuring usage.

In the case of an online archive, statistics for Web archive usage will use similar methods and standards as those existing for measuring usage on the live Web. In the case of a grey archive, statistics for Web archive usage will use similar methods and standards as those existing for assessing the use of electronic resources in libraries. As a result, this Technical Report mostly provides references to existing methods and standards while providing further technical information and definitions to facilitate a better understanding of the specific issues related to access to and usage of Web archive collections.

4.4.2 Definitions and methods for measuring usage

a) Physical visitors

Legislation or institutional policy may dictate that Web archives are restricted to specific access controlled geographic locations such as reading rooms at national libraries or archives. In such cases, primary access data can be collected about the users of such systems through registration and computer system login data. This provides the highest quality usage information as it is possible to ask the user their intent directly.

b) Virtual visitors

Web archives which are published as publicly accessible websites can gather usage statistics through the use of Web Analytics, as standardised by the Web Analytics Association. Web Analytics is the growing field of analysing usage patterns of websites through the following methods.

Virtual visits are defined in ISO 2789 as one continuous cycle of user activities on the library website by users from outside the library's IP address space (usually from outside the library premises), regardless of the number of pages or elements viewed.

NOTE 1 A Website visitor is either a unique and identified Web browser program or an identified IP address that has accessed pages from the library's website.

NOTE 2 The interval between two consecutive requests generally is no longer than a time-out period or 30 minutes if they are to be counted as part of the same virtual visit. A longer interval initiates a new visit.

NOTE 3 Web servers providing services whose statistics are reported at another site are to be excluded from the statistics of the library website.

c) Robot visitors

A type of virtual visitor to Internet resources is non-human agents designed for crawling purposes. They are usually associated with search engines, but archiving crawlers obviously fall into this category as well. Generally robot visitors are excluded from Web analytic statistics gathering by filtering on IP address, User Agent string or more complex methods such as identifying repetitive usage patterns (following every link) that a human visitor would not normally display.

It may be useful to analyse robot behaviour particularly if the access tool has explicit robot exclusions (robots.txt) enabled to stop archive material being listed in search engines (and therefore potentially competing with the live site in search listings).

d) Involuntary visitors

Statistics may be gathered about involuntary visitors. However, the user did not intend to visit the site and therefore acts as false positives. Examples may be search engines forward page scanning, or

virus check plug-ins to browsers that automatically request and scan links before the user has actually requested them. Certain browsers also pre-empt users' browsing choices by forward caching links and generating thumbnails. Automatic redirection to archives from live sites may also fall into this category depending on the subsequent user activity.

e) Log file analysis

Every HTTP transaction can be logged by the Web server with information supplied in the HTTP request header and also HTTP response header generated by the server itself. Information such as Date/time of the request, IP address of the requester (even resolved domain name), requested resource (URL), the referring resource and the response code can be logged for every request/response transaction.

In the early 1990s analysis counting of each log entry provided a count of the 'Hits' a Website received. As HTML became more complex with additions of embedded objects and inclusions of resources such as style sheets or JavaScript, 'Hits' as a metric devalued and today it is only useful as a measure of Web server load.

Log analyser tools however can provide insights into the data contained within the log files with respect to aggregate statistics.

The wide use of Web proxies and caching disrupts the accuracy of log statistics as many users can be hidden behind the same IP address and subsequent requests can be served by the local Web cache, instead of by the Web server running the Access tool of the Web archive. Web caches can also be used to speed up the performance of the Access tool itself at the cost of accurate usage statistics as the access tool Web cache will attempt to service requests before the actual access tool itself.

f) Page tagging

Page tagging is the process by which embedded HTML objects make dynamic call backs when they are requested by a browser when loading an html document. The first application of this method was the use of image page counters in the early 1990s where an image would be requested from a CGI script usually containing the numeric value of the count of the number of times that script had been executed and therefore the number of visitors to that page. This has advanced to using JavaScript and cookies to supply information about the user and the page. As the call back from the page tag can be to any location (not just the hosting Web server) this method has enabled the growth of the outsourced Web Analytics industry.

g) Application level logging

As the access tool is a Web application itself it is possible to build logging metrics of usage directly into the application itself.

h) Privacy

Collecting information regarding user activities by any of the methods described above requires an appropriate privacy policy to be in place and accessible to the users of the system.

The widespread use of page tagging (see clause f) above) can present challenges to publicly accessible Web archives, as the archived sites themselves may contain the embedded tags. Therefore viewing the archived page can result in cookies being set on the client browser and call backs to analytic aggregators (live Web leakage).

4.4.3 Basic statistics measuring archive usage

Table 4 contains statistics predominantly identified by the Web Analytics Association. It is divided into simple count statistics and more complex aggregate or dimension statistics. Some advertising specific metrics have been dropped as irrelevant to the scope of this Technical Report and a number of additions of useful metrics have been placed at the end of the table. In the right column, letters indicate the importance of each proposed statistic: H for High, M for Medium, L for Low. We mostly recommend applying the high importance statistics to Web archives, which would be included as part of standard reporting by most analytics programs.

Table 4 — Basic statistics measuring archive usage

Name	Type	Calculation	Importance
Page View	Count	The number of times a page was viewed.	H – indication of the raw usage of the archive.
Visit (Sessions)	Count	A visit is an interaction by an individual with a Website consisting of one or more requests for a page. If an individual has not taken another action (typically additional page views) on the site within a specified time period, the visit will terminate by timing out.	H – basic count of users of the archive.
Unique Visitors	Count	The number of inferred individual people (filtered for spiders and robots) within a designated reporting timeframe, with activity consisting of one or more visits to a site. Each individual is counted only once in the unique visitor measure for the reporting period.	M
Event	Dimension and/or Count	Any logged or recorded action that has a specific date and time assigned to it by either the browser or server.	L

4.4.4 Aggregated statistics for advanced characterization of archive usage

Table 5 — Aggregated statistics for advanced characterization of archive usage

Name	Type	Calculation	Importance
Entry Page	Dimension	The first page of a visit.	M
Landing Page	Dimension	A page view intended to identify the beginning of the user experience resulting from a defined marketing effort.	L
Exit Page	Dimension	The last page on a site accessed during a visit, signifying the end of a visit/session.	L
Visit Duration	Count	The length of time in a session. Calculation is typically the timestamp of the last activity in the session minus the timestamp of the first activity of the session.	H
Referrer	Dimension	Referrer is a generic term that describes the source of traffic to a page or visit.	M
Page Referrer	Dimension	Page referrer describes the source of traffic to a page.	M
New Visitor	Count	The number of Unique Visitors with activity including a first-ever visit to a site during a reporting period. Note that "first-ever" is with respect to when data began being properly collected by using the current tool.	M
Returning Visitor	Count	The number of Unique Visitors with activity consisting of a visit to a site during a reporting period and where the Unique Visitor also visited the site prior to the reporting period.	M
Repeat Visitor	Count	The number of Unique Visitors with activity consisting of two or more visits to a site during a reporting period.	M
Single Page Visit (Bounce)	Dimension or Count	A visit that consists of one page view.	M
Visitors by Geographic Location	Count	Geo-IP reporting of requestor IP address.	L

Table 5 *(continued)*

Name	Type	Calculation	Importance
Search terms to find the archive	Count	The terms used in search engines to find the Access tool website.	M
Search terms used within archive	Count	The search terms used within the Access tool to find archived captures.	H

4.4.5 Core statistics for collection usage

Table 6 — Core statistics for collection usage

Statistic	Purpose	Example
Number of pages viewed	Extent of use	48 318 pages in the UK Web Archive were viewed between 1st and 30th June 2012
Number of Visits	Extent of use	There were 11 415 visits to the UK Web Archive between 1st and 30th of June 2012
Number of Unique Visitors	Extent of use	There were 9 434 unique visitors to the UK Web Archive between 1st and 30th of June 2012
Visit duration	Users' interest in the archive	On average each visit to the UK Web Archive between 1st and 30th June 2012 lasted 3 min and 25 s
Search terms used within archive	Users' behaviour	The top keyword used to search the UK Web Archive between 1st and 30th June 2012 was "goji berry"

4.5 Web archive preservation

4.5.1 General

The long term preservation of Web archives should not be isolated from wider frameworks applicable to the preservation of all digital resources. Ideally collecting institutions should put in place a dedicated digital preservation system for its digital holdings, compliant with standards such as ISO 14721 (Open Archival Information Systems). The OAIS reference model specifies an archival system dedicated to preserving and maintaining access to digital information over time.

This Technical Report does not discuss the OAIS model in detail but uses its basic concepts and definitions to describe preservation issues related to Web archives. Some of the proposed statistics are generic to other types of digital resources; others are specific to Web archives.

As described in *3.3 Preservation methods*, digital preservation can be done at two levels, at the basic level to keep the bits safe, and at a more sophisticated level using strategies such as migration and emulation, to preserve the appearance, function, behaviour and even the user experience of digital resources. This former is as referred to as "bit stream or physical preservation" and the latter "logical preservation". Statistics described in 4.5.2 are intended to measure the efficiency of bit stream preservation activities. A template is proposed in 4.5.3 to help institutions report metadata which are expected to be preserved in a Web archive. Statistics in 4.5.4 are those relevant to logical preservation.

4.5.2 Statistics for bit-stream preservation

4.5.2.1 Volume of lost or deteriorated resources

4.5.2.1.1 Purpose

Many heritage institutions have experienced data loss due to physical media failures: hard-drives crash unexpectedly, media become obsolete. Some data may be deleted by accident. Deteriorated resources are resources that are not totally lost but whose integrity has been damaged, resulting in impossibility to completely access or render the resource. Information about data loss is not commonly available but it is important to monitor the amount of lost or deteriorated resources, as this is an important indicator of the integrity of Web archives.

4.5.2.1.2 Method

The amount of lost and deteriorated data can be measured in bytes or in number of URLs, obtained by regular comparison of checksums.

4.5.2.2 Volume of replicated and distributed resources

Resources not backed up or replicated are at risk of permanent loss and cannot be recovered. It is a common practice to replicate resources on multiple locations to avoid creating a single point of failure. The volume of replicated resources therefore is an indication of collection safety.

However, replication of assets is not in itself enough to guarantee collection safety. Various other parameters need to be set in order for replication of data to be efficient. The main issues here are diversity and integrity. This includes:

— number of varying soft- and hardware configurations for the different instances;

— physical distance between instances;

— rate and method for data integrity check between data instances;

Resources whose integrity is ensured by such measures are considered replicated and distributed.

Costs are involved in replicating and distributing data. Decisions about the extent to which a Web archive should be replicated should be made by balancing the impact of risk, the costs and the complexity of managing replicated resources.

4.5.3 Statistics related to metadata preservation

The importance of preserving resources together with associated metadata are explained in 3.3.4. It is recommended that collecting institutions report regularly on the nature and volume of metadata within a Web archive, using Table 7.

Table 7 — Statistics related to metadata preservation

Metadata type	Description	Standard used (if any)	Percentage of resources containing the metadata	Comments
One of the metadata types described in 3.3.4, e.g. Descriptive	Description of the metadata		Percentage of resources containing the metadata	Any useful or relevant comments
Examples				
Descriptive	DCMI metadata Elements set, Term name: Subject. Topic of the resource	Dublin Core Metadata Initiative (DCMI); LCSH	30 %	Subject term manually assigned by curators and stored in the Web Curator Tool
Provenance	Configuration files		90 %	Configurations files of the 2004 crawls were discarded
Technical	File formats (MIME types)	Multipurpose Internet Mail Extension (MIME) Part Two: Media Types	100 %	All harvested files have MIME type information, but this could be unreliable
Rights	Permission to archive and provide online access		100 %	Required for open access targets only

4.5.4 Logical preservation statistics

While bit-stream preservation keeps the bytes safe on physical media, logical preservation ensure that resources remain usable over time. This Technical Report proposes three main indicators for logical preservation activities.

4.5.4.1 Distribution by (identified) file formats

4.5.4.1.1 Purpose

The distribution of Web archive resources by file formats is a statistic described in detail in 4.3.2.3, as an element of archive characterization. Knowledge of file formats is also crucial to digital preservation. It is impossible to define a preservation strategy for any resource without format information. Web archives typically contain resources in a wide range of formats. MIME types returned by Web servers are the only readily available information about file formats, which are known to be unreliable. In order to obtain more accurate format information for a Web archive for preservation purposes, the use of format identification tools is required.

In addition to the formulation of a digital preservation strategy, the format information can also be used to identify preservation risks and prioritize preservation actions.

4.5.4.1.2 Method

See "*Method*" described in 4.3.2.3. For the purpose of preservation, more accurate format information is required and this could be achieved by using format identification tools. Then outputs can be calculated and organized in the same ways suggested in 4.3.2.3.

Examples of file format identification tools include DROID and Jhove.

4.5.4.1.3 Limitations

Format identification tools cannot always identify all formats. A good example is new formats which the tools do not yet recognize.

4.5.4.2 Number of file formats with defined preservation strategy

4.5.4.2.1 Purpose

File format identification is a useful starting point. A digital preservation strategy needs to be defined for the major formats used in the Web archive to ensure longevity. Institutions may apply migration and emulation to different formats when required, depending on the format itself, the purpose of use and the resources required for implementation. For example, an institution may decide that emulation is a more appropriate strategy for files such as Shockwave Flash, in order to recreate the interactive user experience, while migration works better for Microsoft office documents. It is important to realize that, based on risk assessment, "do nothing" may also be a valid strategy.

Preservation strategy for key formats in a Web archive indicates the institutional commitment to long term preservation.

4.5.4.2.2 Method

List the file formats for which a preservation strategy has been defined. This can be combined with statistics on resource distribution per file format, measured in bytes or in the number of URLs.

4.5.4.3 Volume of resources per format with activated preservation strategy

4.5.4.3.1 Purpose

Preservation strategies may be defined for formats that are in use and not yet obsolete, in which case institutions may prefer to test the strategies on a sample of resources, instead of activating them for the entire collection. Resource limitations may also prevent them from applying the strategy. Calculating the volume of resources with activated or implemented strategy indicates to what extent at risk resources benefited from a digital preservation strategy. It also demonstrates a level of institutional commitment to digital preservation. In the case of a "do-nothing strategy", each review of the strategy with the decision to continue with it can be regarded as an activation, which ensures that informed decisions are taking place.

There is no "definitive" preservation strategy. As technology evolves, resources may become obsolete again and new strategies for the same format may need to be defined and activated multiple times.

4.5.4.3.2 Method

List the file formats for which the defined preservation strategy has been activated, which can be calculated from the digital preservation system. This statistic is then combined with resource distribution per file format, measured in bytes or in URLs.

4.5.5 Core statistics for collection preservation

Table 8 — Core statistics for collection preservation

Statistic	Purpose	Example
Volume of replicated resources	Safety and resilience	150 terabytes of the Web archive are replicated
Distribution by (identified) file formats	Preservation capability	60 % of the archive are in HTML
The number of formats for which a preservation strategy has been defined	Preservation capability and commitment	5 formats have a defined preservation strategy: HTML, JPEG, GIF, PNG and PDF

4.6 Measuring the costs of Web archiving

The costs of Web archiving can be assessed in similar ways as the costs of building and preserving other digital collections (e.g. digitisation). The only point to bear in mind is that Web archiving activities are still recent and some aspects of the effectiveness and costs can only be measured over time. This is particularly true for costs related to the preservation of Web archives.

4.6.1 Outsourcing

A collecting institution may outsource all or some of its Web archiving activities to a vendor or a third party. Such service may include data harvesting, indexing, access or storage. It may also cover the acquisition of retrospective or historical collections and specific software development activities.

In the case of outsourcing, it is more straightforward to calculate the outsourcing costs of Web archiving as this would constitute the total amount of money charged to the institution by the service provider. There may be additional costs related to selecting content, managing the contract and others, usually funded by the institution internally. These should be added to outsourcing costs to calculate the total costs for Web archiving.

4.6.2 Web archiving in-house

Measuring the costs of in-house Web archiving activities is no more challenging than evaluating more familiar processes such as cataloguing in libraries. Four main categories of expenses should be taken into account: hardware, computing, software, and labour.

4.6.2.1 Hardware costs

Hardware costs include the acquisition and maintenance of the infrastructure necessary to harvest, index, ingest, store and provide access to the data and other digital resources.

4.6.2.2 Computing costs

Computing costs include costs related to power and network (bandwidth).

4.6.2.3 Software costs

Depending on the choice of software, licence fee may be applicable for Web archiving. Many collecting institutions currently use free, open source software developed and maintained by international organisations or not-for-profit foundations. Many use solutions developed through international collaborations, such as the International Internet Preservation Consortium (IIPC). This reduces costs for major in-house developments or paying licence fees to commercial companies.

The implementation of open source software and integration with in-house systems requires technical expertise. Developer resources have to be available to update the local implementation for each new release. Moreover, each institution would have specific requirements, which implies software

developments for customisation. Staff resources for technical operations and software development should be included as labour costs.

4.6.2.4 Labour costs

Labour costs can be measured as usual in FTE (Full Time Equivalent) or person-days. There are in general three broad categories of staff involved in web archiving: curatorial, technical and managerial. Collaboratively they carry out a wide range of activities including collection development, technical operations, planning and performance management. There are typical divisions of labour based on staff members' expertise: for example, a curator is more likely to scope and describe a collection and an engineer to operate the crawler or develop software. However, web archiving often requires hybrid skills, and tasks are in practice often distributed across expertise areas. Staff must acquire a basic understanding and broad knowledge of web archiving beyond their own professional expertise as this improves collaboration and helps deliver better performance.

The tasks can be distributed across various departments, and many staff members may contribute. The number of staff employed in Web archiving is calculated by adding the time spent by all permanent and temporary staff, including project-based staff, on that service area. Several methods are possible:

a) Estimate: Calculate the number of full-time-equivalent positions directly assigned to Web archiving. Estimate the average time spent by those employees on other services and deduct the time from the number. Estimate the average time spent by employees of other service areas on Web archiving and add the time to the number of FTE positions.

   EXAMPLE 3,5 FTE staff are directly assigned to Web archiving. During the reporting period, they spent 10 % of their time on other tasks. Staff members from other services (8 FTE) spent 20 % of their time in Web archiving. Total FTE for Web archiving was then 3,5 – 0,35 +1,6 = 4,75.

b) Time logging: Choose a sampling period (normally one or two weeks) during which Web archiving experiences average workload. Record the time – by work diaries – that staff members, including members of other areas, spend on Web archiving. The sample counts are grossed up to FTE numbers for the reporting period.

c) The costs can also be differentiated as to curating, technical work and management.

4.6.2.5 Other costs

Other costs may include:

— acquisition of metadata, e.g. purchase domain names lists from domain registrars;

— legal expertise: legal advice, provisions for legal action or compensation following a court decision (e. g. a publisher sues a collecting institution for reputation or financial damage);

— international cooperation: Web archiving is performed and supported by a global community. Involvement in international cooperation may generate membership fees and expenses for travel and subsistence.

Table 9 — Core statistics for collection costs

Statistic	Purpose	Example
Hardware costs	Costs of acquiring and maintaining hardware	Cost of replacing the storage infrastructure was € 50 000
Computing costs	Costs related to power and network	Bandwidth costs € 10 000 per year
Software costs	Costs of acquiring, integrating, developing, or improving software	The development of a new curator tool has been outsourced for a cost of € 80 000
Labour costs	Costs for human resources (e.g. curators, engineers…) in FTE or amount of money	The web archiving team includes three full-time engineers and four full-time curators

5 Quality indicators

5.1 General

Quality is defined as "the degree to which a set of inherent characteristics fulfils requirements" (ISO 9000:2005). This clause includes the indicators that allow assessing the degree to which a set of inherent characteristics of a Web archiving programme fulfils the requirements set out by its management and stakeholders.

The quality indicators proposed in this section are intended to help collecting institutions answer fundamental questions such as:

— **Do we know what to collect?**

  — If not, there is a need for a clear policy which defines the scope of the archive.

— **Are we collecting what we want to collect?**

  — If not, there is a need to ensure consistency between the collected and target resources.

— **Are we making the best use of our resources?**

  — If not, there is a need to improve procedures and workflow to increase efficiency.

— **How accessible and searchable is the archive?**

  — It is important to continuously improve the usability of the archive.

— **Can we guarantee that the Web archive will remain accessible over time?**

  — If not, there is a need to put in place reliable preservation procedures.

The proposed indicators are most suitable to assess the quality of services provided by the same organization over time. It is recommended that quality be measured and evaluated regularly. Benchmarking between institutions is possible, if the indicators are applied and interpreted in the same way. However, such comparisons should always be made with caution, taking into account the differences in institutional remit, resources and procedures.

5.2 Limitations

Quality indicators proposed in this Technical Report reflect the current state of the art in Web archiving, both from a technological and a curatorial perspective. They should be revisited and updated in alignment of the developments in Web archiving.

The interpretation of the results of applying quality indicators should be performed with caution. Sampling and measuring errors may occur, and lead to inaccuracy.

5.3 Description

5.3.1 General

Quality indicators are listed below based on their relevance to the key aspects of Web archiving programmes: policy, harvesting, access and preservation.

When calculation may be done in URLs or in bytes, it is recommended to use calculation in URLs. As a matter of fact, these indicators intend to calculate the amount of resources and URLs are more similar than bytes to library resources.

Management

1) Cost per collected URL
2) Percentage of staff involved in Web archiving

Quality of the collecting process

3) Percentage of resources disappeared from the live Web during a given period of time
4) Achieved percentage of mandated scope
5) Percentage of requests for agreements or permissions granted by rights holders

Accessibility and Usage

6) Percentage of resources accessible to end users
7) Percentage of full-text indexed resources
8) Percentage of catalogued resources
9) Annual percentage of accessed resources
10) Percentage of library visits including a visit to the Web archive
11) Number of pages viewed per visit

Preservation

12) Percentage of resources with at least one replication
13) Percentage of lost or deteriorated resources
14) Percentage of resources with identified file format
15) Percentage of resources whose format has a defined preservation strategy
16) Percentage of virus-checked resources

5.3.2 Management

Indicator number	**1**
Name	**Cost per collected URL**
Objective	To assess the efficiency of Web archiving processes.
Prerequisites	- The total cost of Web archiving as described in 4.6; - The total number of crawled URLs.
Method	Cost per collected URL is: A/B where A is the total cost of Web archiving in a specified time period B is the number of crawled URLs within the same period of time Round off to one decimal place.
Comments	Low cost per collected URL in general demonstrates high efficiency of Web archiving processes; higher costs per collected URL could also indicate a high level of curation. This indicator is best used to compare collections of similar sizes and purposes.

Indicator number	**2**
Name	**Percentage of staff involved in Web archiving**
Objective	To indicate the institutional commitment to Web archiving.
Prerequisites	- The number of staff (FTE) involved in Web archiving; - The total number of library staff (FTE).
Method	Percentage of staff involved in Web archiving is: A/B * 100 where A is the number of library staff (FTE) involved in Web archiving: selection, harvesting, providing access, preservation B is the total number of library staff (FTE), including permanent and temporary staff, also project-related employees Round off to the nearest integer. For full-time staff involved in Web archiving only part-time, their time spent on Web archiving may be calculated by self-declaration or by time logging.
Comments	

5.3.3 Quality of the collecting process

Indicator number	**3**
Name	**Percentage of resources disappeared from the live Web during a given period of time**
Objective	To assess the value of the Web archive.
Prerequisites	- The number of targets in the archive; - The number of targets in the archive that have disappeared from the live Web. The second figure may be gathered automatically by the absence of DNS response or 404 responses or manually by checking the live Web.
Method	Percentage of resources disappeared from the live Web during a given period of time is: A/B * 100 where A is the number of disappeared targets B is the number of targets in the archive Round off to one decimal place.
Comments	The disappearance of a target is hard to determine. A target of which the domain (internet address) has changed does not necessarily disappear as it can have a new, different domain. In some cases, only parts of a target disappear – this scenario is excluded for practical reasons. In the sense of this indicator, a target disappears when there is no DNS response or when its Internet address generates a 404 response. If feasible, the most reliable method is checking the live Web manually.

Indicator number	**4**
Name	**Achieved percentage of mandated scope**
Objective	To assess whether the Web archiving results correspond to the mandate.
Prerequisites	- The number of targets harvested by the institution per year; - The number of targets as defined by or derived from mandated scope.
Method	Achieved percentage of mandated scope is: A/B * 100 where A is the number of targets harvested by the library per year B is the number of in-scope targets defined by or derived from the Web archiving mandate (e.g. legal deposit) Round off to one decimal place.
Comments	The mandate for Web archiving is often national as well as institutional. Where national legal deposit legislation exists, it usually defines territoriality of a national domain. For Web resources hosted using the national TLDs, this determination is straightforward. However, resources in scope for a particular national remit are not always hosted under domain names with clear geo references. Only a third of French Websites, for example, are hosted on .fr TLDs according to AFNIC, the .fr registry. There is a need for other means to determine territoriality than national TLDs. This may include obtaining information from domain name registrars.

Indicator number	**5**
Name	**Percentage of requests for agreements or permissions granted by rights holders**
Objective	To assess the effectiveness of permission requests
Prerequisites	- The number of requests for agreements or permissions sent to rights holders; - The number of agreements or permissions granted by rights holders.
Method	Percentage of requests for agreements or permissions granted by right holders is: A/B * 100 where A is the number of agreements or permissions granted by to rights holders B is the number of requests for agreements or permissions sent to rights holders Round off to one decimal place.
Comments	A high rate indicates the successfulness of the permission requesting activity. It is also recommended to record the number of explicit refusals and the number of non-responses. A better communication and advocacy towards website producers may be a solution to improve successfulness. A communication plan may be elaborated to identify the key arguments to convince producers, and to decide on the best channels for distributing information.

5.3.4 Accessibility and usage

Indicator number	**6**
Name	**Percentage of resources accessible to end users**
Objective	To assess the availability of the Web archive.
Prerequisites	- the total number of resources in the Web archive; - the number of resources available online (online archive); - the number of resources available onsite (grey archive). The above can be measured in URLs or in bytes.
Method	Percentage of resources accessible to end users is: (A+A')/B * 100 where A is the number of resources in the Web archive that are available online; A' is the number of resources in the Web archive that are available only onsite; B is the total number of resources in the Web archive. Round off to one decimal place.
Comments	A high rate indicates high visibility or accessibility of the Web archive. The indicator may also be calculated singly for the resources available online in order to assess the direct availability of resources for end users. The measuring unit used for calculation should be reported, i.e. URLs or bytes.

Indicator number	**7**
Name	**Percentage of full-text indexed resources**
Objective	To assess the searchability of the Web archive.
Prerequisites	- the total number of resources in the Web archive; - the number of resources that have been full-text indexed. The calculation may be done in URLs or in bytes.
Method	Percentage of full-text indexed resources is: A/B * 100 where A is the number of resources that have been full-text indexed B is the total number of resources in the Web archive Round off to one decimal place.
Comments	Full-text searches greatly enhance the accessibility and usability of the Web archive. The measuring unit used for calculation should be reported, i.e. URLs or bytes.

Indicator number	**8**
Name	**Percentage of catalogued resources**
Objective	To assess the searchability and the level of curation of the Web archive.
Prerequisites	- The number of targets in the Web archive; - The number of targets that have a catalogue record.
Method	Percentage of catalogued resources is: A/B * 100 where A is the number of targets that have a catalogue record B is the total number of targets in the Web archive Round off to one decimal place.
Comments	Cataloguing resources in Web archives enhances their accessibility and usability. It also helps integrate Web archive resources with other resources held by the library. Cataloguing however is costly and may not be possible for large scale Web archives collected through bulk crawls. When calculating this indicator it is recommended to also report the harvesting strategy used to collect the resources.

Indicator number	**9**
Name	**Percentage of accessed resources**
Objective	To assess the breadth of actual usage of the Web archive.
Prerequisites	- The total number of domain names in the Web archive; - The number of domain names on which at least a page was viewed in a year.
Method	Percentage of accessed resources is: A/B * 100 where A is the number of domain names on which at least a page was viewed in a year B is the total number of domain names in the Web archive Round off to one decimal place.
Comments	A high percentage indicates wide usage of the Web archive. The reason to use domain names instead of URLs is that possibly only the resources under a limited number of domains are actively used. Calculating usage per domain indicates the breadth of usage and the comprehensiveness of the archive. A better communication towards researchers as well as general public is a way to improve this percentage. As stated for indicator 5, a communication plan may be elaborated to identify the more convincing examples of valuable resources, and to decide on the best channels for distributing information.