Counting the uncountable: statistics for web archives

Clement Oury, Roswitha Poll

Counting the uncountable:
Statistics for web archives

Author 1
Clément Oury, Head of Digital Legal Deposit at the Bibliothèquenationale de France, is the convenor of the ISO TC 46 SC 8 WG 9, "Statistics and Quality Indicators for web archives". He is also convenor of the ISO TC 46 SC 4 WG 12, "WARC file format" and Treasurer of the International Internet Preservation Consortium (IIPC).

Author 2
Roswitha Poll, former chief librarian of Münster University and Regional Library, Germany, is chair of ISO TC 46 SC 8 "Quality – Statistics and performance evaluation" and convenor of three ISO working groups.

Keywords
Libraries, web archives, statistics, quality measures, ISO/TR 14873

Abstract
Formore than a decade, libraries have started to "collect the web". National libraries in particular select, collect and store publications and websites from their national domain, seeing this as a task similar to traditional legal deposit. The collection policies and collecting methods vary, so that it is difficult to compare the quantity and quality of the respective web archives.
In order to harmonize the evaluation of web archives, ISO TC 46 SC 8 has produced a Technical Report that standardizes the terminology and statistics and offers tested indicators for assessing the quality of web archiving.
The paper describes the aimsand contents of the ISO Report.

1. General

1.1. Standards forassessingthequantityandqualityoflibraryservices

In the last decades, ISO has developed a set of standards for assessing the quantity and quality of library services. Thiswork was and is done within the Technical Committee 46 Information and documentation, Subcommittee 8 Quality – Statistics and performance evaluation.

- **ISO 2789**International library statistics
  Thisis the central standard for all statistical procedures in libraries. The 5th edition is being published.
- **ISO 11620** Library performance indicators
  This standard describes quality indicators for library services. The 3rd edition is planned for the end of 2013.
- **ISO/TR 28118** Performance indicators for national libraries
  The Technical Report (published 2009) describesquality indicators for the specific tasks of national libraries. As a number of these indicators have been taken over into ISO 11620, no new edition of the TR is planned at the moment.
- **ISO/DIS 16439**Methods and procedures for assessing the impact of libraries

This new standard describes the ways to identify and provide proof ofa positive impact of libraries on individuals and society. Publication is planned for the end of 2013.

The standards, especially ISO 2789, are intended to provide definitions and statisticalprocedures for all types of library collections and library services. New activities are continuously integrated into the new versions, e. g. digitisation of analogue material or services for mobile devices.

There is one library activity that has developed recently and has not yet found its place in the existing standards for library evaluation, namely web archiving.Evidently, this huge new task did not seem to fit in with the usual role of libraries. Archiving the web means selecting and capturingInternet resources, storingthem in web archives, preserving them and managing sustainable access to the archives.The collecting processes are managed automatically at regular intervals by harvesting software.

Web archiving startedtowards the end of the nineties, mainly in national libraries. Most national libraries are responsible under legal deposit law for collecting and preserving the printed cultural heritage of their country. This task has now been – legally or voluntarily – extended to Internet resources. The reason for the collection is the same as for print material: the danger of items – in this case websites – disappearing and being lost for future generations.

The libraries started their archiving activities from different approaches, but cooperation soon proved to be crucial in this quickly changing area. Therefore the technical and legal problems were taken up by IIPC,the International Internet Preservation Consortium, founded in 2003[1].

2.2. A first step towards standardisation in web archives: the WARC format

The first challenges which heritage institutions needed to face were of a technical nature: how to harvest, store, access and preserve the immense set of data available on the web. This is the reason why the first achievements towards standardisation were in the technical domain. When pioneer institutions started web archiving, in the early 2000s, there was at least a *de facto* standard: the ARC file, a container file designed by Internet Archive in 1996 to concatenate, store and handle the thousands of files harvested on the web[2].

The ARC file was indeed the reference format for all tools developed and used by Internet Archive and its partners in the framework of the IIPC: the Heritrix crawler used to harvest websites, the Wayback Machine used to access web archives… However in the middle of the 2000s emerged a strong need for the adoption of a standard format:

- first, the ARC format needed to evolve to better take into account new requirements for collection description and long term preservation: new characteristics, new functionalities and record types were expected;
- as the ARC format specification was short, simple and highly adaptable, the creation and usage rules of ARC files were mainly dictated by the way tools developed by Internet Archive were working; other institutions needed a better overview of the evolutions of the format;

[1] http://www.netpreserve.org/about/index.php
[2] http://archive.org/web/researcher/ArcFileFormat.php

- finally, as web archiving was a new activity, being able to build collections and tools based on an internationally standardised format was a way to give strong institutional confidence on the long term maintenance of web archives and web archiving activities.

A draft standard for an evolution of the ARC format came out of the original discussions of the members of the IIPC, called WARC (Web ARChive format)[3]. After a few years of standardisation process, the first version of the "ISO 28500 WARC file format" was published on 15 May 2009. This standard has been adopted by most heritage institutions and is now even used for other kind of collections: storage of e-journals, of digitized material…[4]

2. The new Technical Report: ISO/TR 14873

2.1 Why introduce standardisation for web archives?

Web archiving institutions are not only facing technical issues, and other standardisation needs emerged when they started gathering a large amount of resources. It became rapidly critical to identify quantitative and qualitative measures for the evaluation of this new library service. Collecting the web is a complex and expensive activity, and funding institutions ask for evidence of the cost-effectiveness of the service and of its value for society. Definitions and collecting methods for statistical data and quality indicators were needed.

At first, standardisation was needed in order to provide a necessary clarification of terms and measures used to define and assess the different parts of a web archiving process. Although the majority of web archiving institutions use similar crawling technologies, and even the same software, they do not necessarily describe their activity the same way. For example, some of them were crawling or harvesting, whereas others were collecting or archiving – institutions were sometimes using different words for the same kind of activity; or the same words for different kind of activities. It is even more complex when it comes to the result of the crawl: should institutions call the copy of a website performed by a crawler a "capture", a "version" or an "archive" of this website?

However, this was only a matter of definitions for new tools and new procedures. The traditional library paradigms were more severely challenged by the fact that institutions had to deal with totally different kinds of documents. The web offers several levels of granularity at which institutions may identify "documents" that could be assimilated to "classical" library holdings. The website appears at first sight as a possible documentary unit; but some institutions may only choose to harvest parts of websites, or only individual pages – this is especially the case for very large websites: institutional websites or publication platforms. How to count those parts or those pages along with websites? Finally, each individual component of a webpage – i.e. each web file – may also be viewed as a separate document: a PDF, a JPEG or PNG image, etc.

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717; see also http://bibnum.bnf.fr/warc/.

[4] See for example David S. H. Rosenthal, "LOCKSS: Lots Of Copies Keep Stuff Safe", in *Proceedings of the US Workshop on Roadmap for Digital Preservation Interoperability Framework*, Gaithersburg (USA), 29-31 March 2010, online: http://ddp.nist.gov/workshop/papers/03_06_Dave_Rosenthal_NIST2010.pdf; and EldZierau, "Package Formats for Preserved Digital Material", in *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)*, Toronto (Canada), 1-5 October 2012, p. 54-62, online: https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf.

These two issues – the lack of maturity of the terminology, and the complex granularity of documents on the web – may first be seen as problematic within institutions. How to define a content policy if the borders of the content itself are fuzzy? How to explain to readers what they can find in the collection if librarians are not even able to precisely describe it? They become even more complex when a dialogue or a comparison must be established between institutions.

Standardisation was therefore identified as one of the ways to face this kind of issue. The clarification of definitions, and the proposal of agreed statistics and indicators were intended to facilitate:

- international measures and comparisons between institutions;
- therefore, the identification of best evaluation practices within institutions
- and a better understanding and advocacy of web archiving initiatives in a wider environment, either national, regional or international.

Indeed, having a standard for statistics and quality indicators was in itself a way to obtain international recognition of web archives as heritage and research library material, and as an activity in the durability and reliability of which library managers could have confidence. This is the kind of role that the standardisation of WARC, the storage and preservation format for web archives, played few years ago.

The standardisation process

To this end, ISO TC 46 SC 8 decided to address the topic of web archiving. A working group was launched in December 2009. In order to ensure that the needs of the widest range of libraries were taken into account, members of this working group included both long-time practitioners and beginners in the field of web archiving. However, they were all working in national libraries (France, Germany, Spain, Sweden, United Kingdom; with the specific case of the state library of Bavaria). The draft documents were therefore distributed to the other members of the IIPC, in order to ensure that the chosen terminology, statistics and indicators were consistent with the current practices in university libraries, or in other research and heritage institutions (national archives, not for profit foundations, etc.). The working group notably got detailed and valuable comments from the Library of Congress and the University of North Texas libraries, which had worked on metrics for web archives in the framework of the End of Term Project[5].

SC 8 decided that instead of incorporating web archiving directly into the existing library standards, the topic should be first handled in a Technical Report. This type of ISO publication has less formal regulations than a standard and can best serve the purpose of quickly proposing an instrument for the evaluation of web archiving activities. After a phase of testing, the definitions and procedures may be integrated into new editions of ISO 2789 and ISO 11620.

2.3 A report addressing all components of web archiving activities

ISO/TR 14873“Statistics and quality issues for web archiving”, which will be released in 2013,proposes statistics and indicators not only forall individualsdirectly involved in web archiving, but also for the management of the collecting institutions and for funding

[5]See http://research.library.unt.edu/eotcd/wiki/Main_Pageandespeciallyhttp://research.library.unt.edu/eotcd/wiki/Category:Metrics

authorities. This wide potential audience is taken into consideration in the terminology and descriptions used, and more general information is given about the background, practicesand problems of web archiving than would be needed for experts in this topic. After defining the terminology, the Report describes the following:

a. *The different collecting strategies: bulk harvesting, selective harvesting, or a combinationof both*
   Bulk harvesting means that the crawl (the process of browsing and copying web resources) is intended to capture entirely a top level domain or a subset (e. g. .fr, .de). In selective harvesting web resources are collected according to certain criteria, such as relevance to specific subjects or events (e.g. economiccrisis, Inuit culture) or scholarly importance of the resources.

b. *Access and description strategies*
   Indexes in web archives basicallyconsist of the URL of the requested resource, plus the date of its harvest, in order to differentiate between capturesof the same resource. Other methods of describing web archives include:
   - cataloguing: a usefulwaytointegratewebarchiveswith alibrary'sexistingcollections, so thattheybecomediscoverablethroughcatalogueSearch, but resource intensive and not suitable for bulk harvest;
   - full-text search: a powerful way to discover resources but technicallychallengingtoimplement;
   - automated keyword or metadata extraction
   - data mining techniques (text mining, link analysis…): a growingneedexpressedamongresearchcommunitiesthatofferopportunitiestoprovide accessto different viewsof a webarchive, unlockingembeddedpatternsandtrends, relationshipsandcontexts.

   Special software is needed for finding and replaying resources in the web archive. The software must be able to identify and retrieve unique resources. But as the archived resources have been "frozen" at a specified point of time, they cannot show the same interactivity as live versions (e.g.message boards, discussion forums).

c. *Preservationmethods*
   The value of web archives generally reveals itself after a certain period of time, when the original websites are no longer available online. Therefore, maintaining a long term access to these collections is a critical issue. Compared to other kinds of digital documents, the specific difficulties of preserving web archives are related to the huge amount of data involved and the diversity of file formats and media types.

d. *The legal basis for web archiving*
   Web archiving is mainly regulated by national legislation on copyright and/or legal deposit. This legislation can include or exclude certain resources and can restrict web archiving to permission-based collecting which requires permissions from the rights holders prior to the harvesting.
   In countries where no specific legislation is available, institutions may choose a permission-based approach where authorization to harvest resources should be obtained from the rights holders. 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 canbe taken down when requested by the rights holders.

3. Measuring quantity in web archiving

As is the case with the statistics for all other library services, the data collected for web archiving should be appropriate for the evaluation of the service, for comparison over time and between libraries, for the support of internal library management, and for promotingthe value of the service. The Technical Report recommends the continuous collection of the following statistics:

3.1. Collection: statistics onthe size and growth of the web archive

The first question about web archives is usually: How big is such an archive, and how much is added every day, month or year?
Traditional library materials are counted as volumes, audio-visual documents or manuscripts, the electronic resources as eBooks, e-journals or databases. Quantifying the size of a web archive demands other measures.
Indeed, as mentioned above, it isnot straightforward to define what the discrete "document" ison the web. Each file may be considered a separate document: a jpeg may be considered an image, a pdf may be considered a monograph… On the other hand, html pages groups together several, sometimes tens of different files; and generally individual files are meaningless if they are not considered in relation to the other files to which they are connected. Libraries do not count the number of pages in the books they hold in their stacks… except perhaps for the calculation of digitization budgets.
It is thus tempting to refer to the notion of website. However a website is merely an intellectual entity: it is a set of interconnected pages produced by a single publisher (a person, a group or an organisation), generally (but not necessarily) hosted on the same domain and/or on the same host. There is no real technical way to count the number of websites within a web archive.
Therefore, the following measures are recommended:

- number of targets (i.e. number of intellectual entities selected by librarians; which generally correspond to a website); and number of captures of these targets (the number of captures depends on the crawling frequency);
- number of domains or hosts (again used as a substitute for the number of websites, this measure is also applicable to bulk crawls, contrary to the number of targets);
- number of URLs, i.e. number of responses to the http request sent by the crawler (even 404 errors should be included);
- number of bytes (compressed and uncompressed): this measure is useful whenplanning storage, and is comparable to the linear metrefor the management of stacks in a library.

3.2. Collection: statistics onthe contents

The contents of a web archive can also be described and counted accordingto the following criteria: geography, format,language, and chronology.

- *geographical distribution, indicated by top level or second level domains*
Resources may be hosted on generic top level domains (TLDs), such as .com or .org; or on country code TLDs, such as .dk, .de… The country code TLDs indicate the

geographical distribution of the resources in a Web archive. This measure is especially useful for national libraries which are entitled to collect the entire intellectual output of their country. For a more thorough characterization, it is also possible to look at the secondlevel domains, which are subdivisions within the top level domains for specific categories of organisations (e.g. .gov.uk for governmental websites).

- *differentiation by format type*
This differentiation can use resource types (e.g. text, image, audio) or file formats(e.g. html, jpeg).It is also especially useful for preservation issues.

- *differentiation by language*
Identification of languages within a web archive and within one top level domain helps the understanding of the cultural diversity in a country or the proximity to other countries; it is however technically challenging. Characterisation by language has always been used for library collections.

- *chronological distribution*
Chronology here does not mean the date of origin of a resource, but the point of time at which it was archived. This statistic is useful for assessing the uniqueness of resources: If they have been archived years ago, they are the more likely to have disappeared on the live web. The data can also be used for preservation issues: older resources are more likely to be in an obsolete format.

3.3. Statistics on the usageof the archive

The conditions for using the archived resources depend on the national legislations and on the policies of the collecting institution that can restrict direct access to specific locations (e.g. the reading-room of a national library) or to certain parts of the archive. Web archives are called white, grey or dark according to their degree of accessibility (online access / onsite access only / access only for library professionals).

If online access to the archive is possible, usage data can be collected via web analytics (e.g. page views or visits). If access is restricted to a location, unique visitors can be counted and can also be surveyed as to purpose and success of their visit. Usage statistics are crucial for web archives as they can show a direct benefit of the collection.

3.4. Preservation data

As with other kind of digital documents, preservation of web archives should be done at two levels: at the basic level,the integrity of the “bitstream”should be ensured, and at a more sophisticated level, the appearance, function, behaviour and even the user experience of digital resources should be preserved, using strategies such as migration and emulation. Statistics are proposed for both kinds of preservation levels:
- for bitstream or physical preservation, statistics help in calculating if the data are kept in a safe manner, indicating for example what proportion of the archive gets a replicated storage;

- statistics for logical preservation deal with the nature and the reliability of metadata available for the resources. The most critical information is related to the format of resources; it is necessary to calculate what proportion of the archive is in a format for which a preservation strategy has been defined.

3.5. Cost data

The costs of web archiving can be assessed much in the same way as the costs of other library services, namely staff costs, costs of hardware and software and maintenanceof buildings. But web archiving activities are still recent, and some cost factors, especially for long-term preservation, will become more visible over time.

4. Measuring quality in web archiving

The quality - effectiveness and efficiency - of library activities and services should be measured against the background of a library's mission and goals. The goals of web archiving as defined by legislation and/or libraries' mission statements can be summarised as following:

- to collect and preserve the contents of the web as part of cultural heritage
- to organise permanent access to the archived material for research and general information

The Technical Report offers a number of quality indicators or performance indicators that have been tested within the ISO working group. The indicators have been selected according to the criteria for library performance indicators as described in ISO 11620: such indicators should be reliable if used repeatedly, informative for the library and its authorities, yielding results that are comparable between libraries, and practical, using data that the library can collect without high levels of effort (e.g. automated statistics from harvesting software). For measuring the quality of web archiving, another criterion can be added: flexibility. The indicators should be adaptable to quick changes that are to be expected in the development of the web.

Web archiving activities show high quality if they comply with the following requirements:

a. **The library has a clear statement of the intended coverage of the collection and succeeds in following it.**
   This criterion is difficult to measure even for the print legal deposit collection.[6]
   For the web archive, the indicator is
   - Achieved percentage of mandated scope
   For permission-based collecting, the indicator measures:
   - Percentage of requests for agreements or permissions granted by rights holders

b. **Care is given to the structuring and indexing of the collection in consideration of searching possibilities.**
   The indicators are:

[6] ISO/TR **28118: 2009**Performance indicators for national libraries, Indicator A.1

- Percentage of full-text indexed resources
- Percentage of catalogued resources

c. The library has implemented long-term preservation procedures.
There are several indicators for preservation:
- Percentage of the resources with at least one replication
- Percentage of lost or deteriorated resources
- Percentage of resources with identified file format
- Percentage of resources whose format has a defined preservation strategy
- Percentage of virus-checked resources

d. As far as possible, there is free (online)access to the collection.
The indicators measure accessibility and accesses.
- Percentage of resources accessible to end users
- Annual percentage of accessed resources
- Percentage of library visits including a visit to the Web archive

e. The archiving activity is performed in a cost-effective way.
The quality indicators measure:
- The costs per collected URL
- The percentage of total library staff that is involved in web archiving

There is one indicator that underlines thevalue of web archiving for future generations by assessing what part of the web archive does not exist anymore on the live web:

- Percentage of resources disappeared from the live web during a given period of time

5. Use of the measures

Not all of the proposed statistics and quality measures should be used for every web archive; some are more adapted for selective harvesting than for bulk harvesting. Though web archiving is a genuine task of national libraries (or national archives), other institutions are also engaged in collecting web resources, mainly for special collections. Most of the measures that have been defined in the Technical report will also be relevant for smaller archives, collected by selective harvesting.

The data will in any case be useful for internal management, for allocation of staff and other resources, for storage and preservation planning. Statistics of usage and cost data will be especially interesting for the library's authorities. Measures demonstrating the size and growth of the archive, the coverage of the legal mandate and the salvage of disappeared websites can interest the public, the media, politicians, and all institutions responsible for cultural heritage.

References

BERMES, E. and ILLIEN, G.« Metrics and strategies for Web heritage management and preservation », in Proceedings of 75th IFLA General Conference and Assembly, Milano (Italy), 23-27 August 2009, online: http://conference.ifla.org/past/ifla75/92-bermes-en.pdf.

IIPC Access Working Group: Use cases for Access to Internet Archives,2006, online:http://www.netpreserve.org/sites/default/files/resources/UseCases.pdf

MASANÈS, J. (ed.),*Web Archiving*, Springer, Berlin, 2006.

OURY C. « New collections, new measures: metrics and quality indicators for web archives », in: *Proceedings of the International Internet Preservation Consortium (IIPC) General Assembly*, Washington (USA), May 2012, online: http://netpreserve.org/sites/default/files/resources/IIPCGA_ISO_Workshop.pdf