

ISO Draft Technical Report Web Archive Quality issues and statistics



Recap of main findings and recommended Quality indicators

Gildas Illien, BnF, convenor of TC46/SC8/WG9
IIPC General Assembly 2011 – The Hague

Overview

- 1- Baseline, scope and motivations
- 2- General structure of the Technical Report
- 3- Core statistics and quality indicators

1- Baseline, scope & motivations 1/3

Goals:

- Obtain international recognition of Web archives as Heritage and Research Library material
- Clarify definitions and propose statistics & indicators to facilitate :
 - better understanding & advocacy of web archiving initiatives in a wider environment, either national, regional or international;
 - international measures and comparisons;
 - best evaluation practices within institutions;

1- Baseline, scope & motivations 2/3

- **This is not a standard, but a technical report**
 - However, key statistics for web archiving will be added to Library general statistics ISO standards
- **Scope : Heritage & Research institutions**
 - e.g institutions, foundations, contractors...
- **Targeted audience:**
 - Library practitioners;
 - Decision-makers and other stakeholders;
 - Terms : the effort is more to translate computing terms from IT to librarians & managers than the other way around.

1- Baseline, scope & motivations 3/3

- Defining « documents »
- Finding common statistics for bulk and selective harvesting
- Considering activities such as Cataloguing and Permission management
- Finding appropriate statistics for in-house and online usage
- Accessing value, risk and costs along with organisation issues
- Selecting quality indicators

2- Structure of the Technical Report, links to IIPC current work



- General
- Terms & Definitions
- Criteria for statistics & quality measures
- Statistics
 - Collection assessment & development
 - Collection description and usage
 - Collection preservation
 - Web archiving management
- Quality issues & indicators
- Usage and benefits
- Bibliography
- Advocacy & Outreach TF
- Harvesting WG; curators' forum
- Access WG
- Preservation WG
- QA issues
- Advocacy & Outreach TF

3- Quality indicators : Key questions to evaluate a web archiving program

- Is there a clear definition of the intended coverage or target of the archive: do we know what to collect?
- Consistency and quality of collecting: do we collect what we want?
- Effective harvesting procedures: do we make best use of our resources?
- Searching possibilities and usability of the archive (as to topics or collections): how accessible and searchable/organized is the archive?
- Long-term preservation procedures: can we guarantee the investment will be saved over time?

Quality indicators 1/5 : collection assessment and development

- Existence or not of a collection policy.
- Percentage of increase of certain file formats meaningful or critical to a given mission or target in the collection over a given time period.
- Percentage of a collection that has disappeared from the live web during a given period (by means of large scale gap analysis or sampling).
- Temporal coverage of the collection (based on date of oldest archive).
- Size & growth of the collection:
 - by number of URLs.
 - by number of terabytes .



Quality indicators 2/5 : collection assessment and development

- **Average harvest frequency** for domain crawls / for selective crawls over a year period.
- When not addressed by Legal Deposit laws, annual growth of the **number of agreements or permissions** achieved with rightholders.
- **Average percentage of missing files** per archived URL (e.g., a .jpg file in a web page), by sampling.
- **Labor cost per Terabyte** - can show various things depending on context : either that selective harvesting making things more expensive or that the institution who invested in QA and selection is likely to hold a collection of higher quality.
- Difference between the original number of URLs to crawl (e.g., a list provided by a certain national domain name registry) and the number of URLs successfully harvested.

Core statistics 3/5 : collection description and usage

- Percentage of collection in white or grey Archive.
- Percentage of collection that has been indexed (whatever the chosen means of indexing or description: e.g automated or catalogued).
- In case of on-site access to the collection, number of terminals available for researchers.
- Growth of unique users for a given period.
- Growth of total number or average of search requests for a given period.
- Growth of the total number of viewed pages for a given period.
- Growth of average time spent per session for a given period.



Core statistics 4/5 : collection preservation

- Percentage of collection that has at least one replication.
- Percentage of collection with preservation metadata assigned.
- Number of “lost” files from the collection after a given period of time.



Core statistics 5/5 : costs & management

- Labour cost per collected terabyte - should include:
- Working time spent in tasks related to the processes of web archiving, considering the salaries of the people involved.
- Number of people assigned to the project full-time.
- Number of people assigned to the project part-time.
- Consumable material spent in different tasks.
- Expenses in software developments.
- Expenses in hardware (servers, CPUs for users access...)

