

Archiving the Web: the mass preservation challenge

Catherine Lupovici

Chargée de Mission auprès du Directeur des Services et des Réseaux
Bibliothèque nationale de France



Tools and Trends: International Conference on Digital Preservation,
1-2 November 2007, Koninklijke Bibliotheek, Den Haag

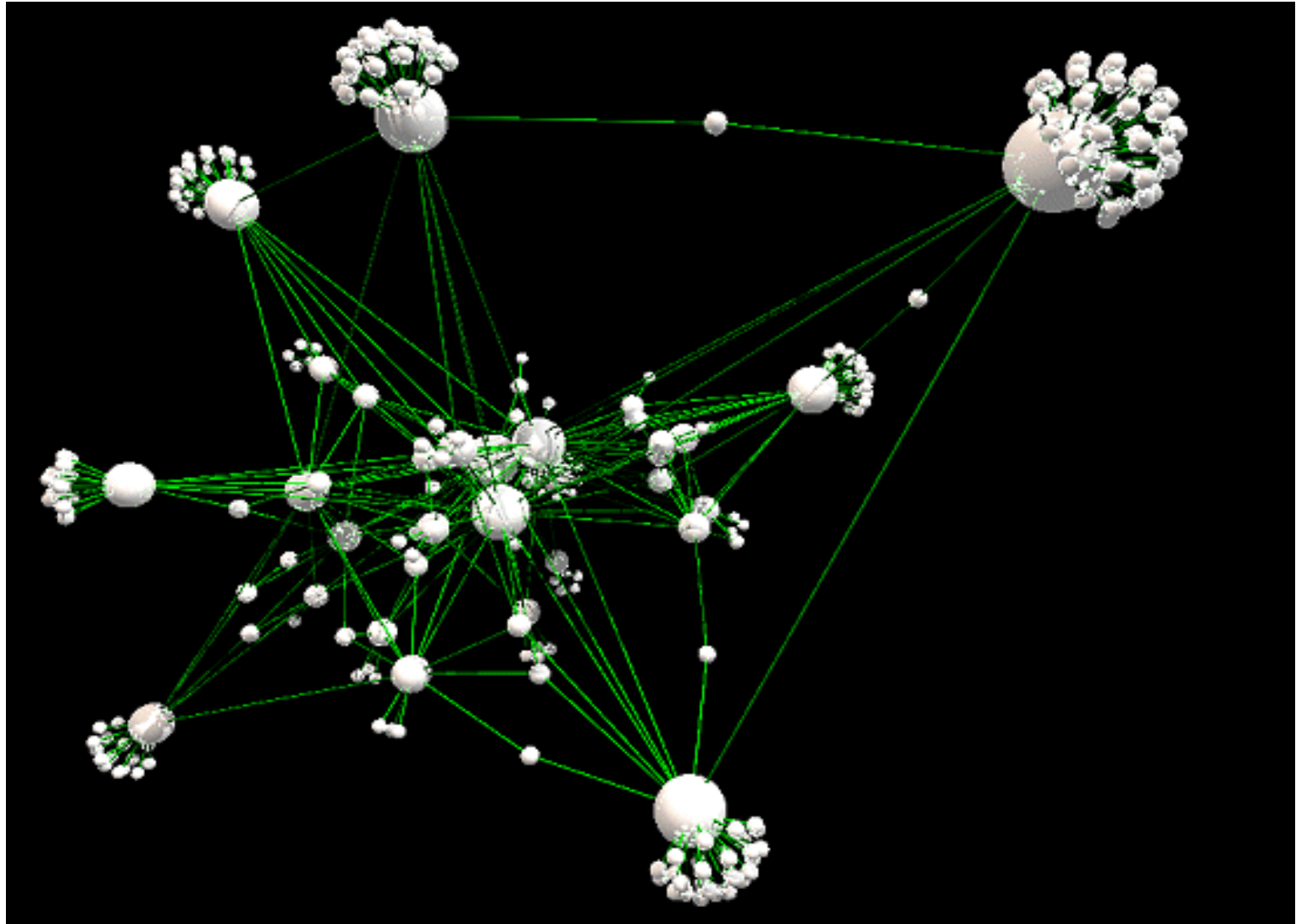
Preservation of the Web

- Initial concern has been to capture the Web to be able to save it. Emphasis put on the crawlers capabilities
 - NEDLIB harvester
 - Heritrix developed by International Internet Preservation consortium (IIPC) members
- Second concern has been to provide access to users: search and navigate through time
 - Rendering problems identified
- Long term preservation is now coming on the front, with the web archives to be ingested in the Digital preservation repositories
 - Denmark started to work on its trusted Digital repository with the Web archiving

The Web content specificity

- Web content
 - Classical types from different previous domains (Publishing, Institutional records, Audiovisual ...)
 - Continuous emerging types: blogs
- Web content technical characteristics & challenges
 - Mass, continuing resources, dynamic content and streaming
 - Surface web or visible web (open web content) and deep web (restricted access for the robots or non harvestable contents for instance accessible via data bases)
 - Interlinking

Links are part of the web content



Web size estimates

- OCLC Web Characterization study:
<http://wcp.oclc.org>

	1998	2000	2002
Unique sites	2 636 000	7 128 000	8 712 000
Public sites	1 457 000	2 942 000	3 080 000

- Indexable Web
 - 200 million pages in 1997 (K.Bharat and A.Broder)
 - 11,5 billion pages end January 2005 (A.Gulli and A.Signorini)
- A snapshot of .fr domain November 2006
 - 271.7 million files, 2 928 000 hosts, 7.23 Tb

Web archiving models

- Whole domain approach through broad crawl
 - Internet Archive, National Library Sweden
- Selection: collection of specified site snapshots
 - National Library of Australia, UKWAC
- Specific selection through thematic approach
 - Presidential elections at LoC
- Deposit (legal or voluntary deposit codes)
 - Netherlands e-deposit
- Combine approaches
 - Denmark, France

Web preservation characteristics

- Scale. Web archiving produces massive and technically heterogeneous digital collections even with a selective policy
- Packaging of data
 - ARC files format that is used to store web crawls as sequences of content blocks harvested. Each block has a header with some metadata (mime file type)
 - WARC which is an extension of ARC. Currently at the ISO Committee Draft status

WARC format

- Ability to store arbitrary metadata linked to other stored data (e.g., subject classifier, discovered language, encoding)
- Support for data compression and maintenance of data record integrity
- Ability to store all control information from the harvesting protocol (e.g., request headers), not just response information
- Ability to store the results of data migrations linked to other stored data
- Ability to store a duplicate detection event
- Ability to store globally unique record identifiers
- Support for deterministic handling of long records (e.g., truncation, segmentation).

International Internet Preservation Consortium

- IIPC launched in July 2003 <http://netpreserve.org>
 - Technical standardisation at the domain scale: functional architecture and standard APIs, archival format, metadata, permanent identification
 - Strong basic tools for all the chain from acquisition to access processes
 - Open source, free licence tools
 - Provide a forum for sharing knowledge about Internet content archiving both within the Consortium and beyond
- Working group on Preservation created in January 2007

IIPC preservation WG program of work

- Identification of applicable standards and tools not necessarily developed solely for web archives
- Test the use of existing format tools on ARC and WARC files
- Liaison with other related efforts
 - Planets (EU funded)
 - Web at Risk, California Digital Library (NDIIPP)
- Report by end of December 2007

Conclusion

- The work achieved so far leads to post process the collections to ingest them in the trusted repositories (a large scale migration test)
- The main question for the future is how to do it in the workflow of collecting, pre-ingest and ingest at the Web scale.