A Survey of Web Archiving - A Recordkeeping Perspective

Paul Wu Horng-Jyh
Nanyang Technological University

ICA Kuala Lumpur
22 July 2008

Structuration Theory and Recordkeeping

Every ***social activities*** have ***intended functions*** as well as ***unintended consequences***

Thus

- ***Current*** recordkeeping is to preserve the *structure* of the documents recording the activities
- ***Regulatory*** recordkeeping is to preserve the *function* recorded in the documents of the activities
- ***Historical*** recordkeeping is to preserve the *total* (intended as well as non-intended) consequences

Web Archiving

To preserve Web documents for

Current, ***Regulatory***, and ***Historical (CRH)*** recordkeeping goals

First, **records** are:

"... the documents that you *make*, *receive* and *use* in your *activities*, and that you keep because you may need them later or because you want to have *reliable evidence* of what you have done"

- *Creator Guidelines* published in IP 1 & 2

Current web archiving

- A current web archiving perspective involves looking at web archiving processes from the viewpoint of what needs to be done to ***capture a record*** and ***fix it in its context of creation*** so that it can be ***recalled***, ***re-presented*** and ***distributed*** for as long as it is of continuing value

Regulatory web archiving

- A regulatory web archiving perspective involves looking at web archiving processes from the viewpoint of how they can be ***standardised***, ***controlled*** and ***monitored***

Historical web archiving

- A historical web archiving perspective focuses attention on what has to be done to ***maintain*** this record and ***manage its meaning*** over time, whether that be for a nanosecond or a millenium.

Degrees of Recordkeeping and Web Archiving Processes

- **Degree 1**: Fix Web contents as static Web pages & as viewed by the creators/users
- **Degree 2**: Create metadata for the Web pages and/or websites
- **Degree 3**: Integrate Web publishing with organizational recordkeeping workflow with accompanying metadata
- **Degree 4**: Organize and describe the relationship among a collection of websites as fonds or interlinking series

	Degree 1	Degree 2	Degree 3	Degree 4
Current Recordkeeping	X	X	-	-
Regulatory Recordkeeping	X	X	X	-
Historical Recordkeeping	X	X	X	X

Parties and Recordkeeping Goals

	Creator	Preserver	Researcher	IT Specialist
Current Web Archiving	1	2	3	1
Regulatory Web Archiving	2	1	2	2
Historical Web Archiving	3	2	1	3

Legend: 1, most; 3, least, appropriate

Pragmatic Considerations

- As much as possible, a record should be kept with all three recordkeeping perspectives/goals
- Each goal requires cooperation of multiple parties
- With limited resources, each party may need to cover for other parties and take up necessary skills
- Depending on available resources, priority task may be set differently for each party

Action Research is used to study the non-optimal cases how each party may collaborate pragmatically

Website Records in Singapore Legislation Activities

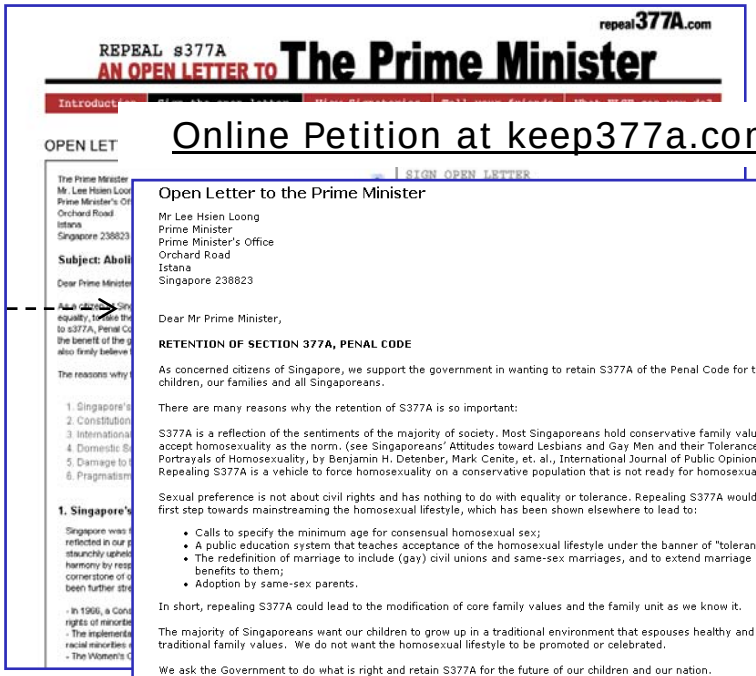
- **Online parliament papers on penal code debate record online petitions at keep377a.com & repeal377a.com**

Parliament speech by PM of Singapore at parliament.gov.sg

" … There was a Petition to remove section 377A. It accumulated a couple of thousand signatures which were presented to this House. Therefore, there was a counter-petition to retain it, which collected 15,000 signatures - at least, according to the newspapers. I have not counted the signatures - 16,000. …"

- By *Prime Minister Lee Hsien Loong*

Online Petition at repeal377a.com



Online Petition at keep377a.com

Open Letter to the Prime Minister

Mr Lee Hsien Loong
Prime Minister
Prime Minister's Office
Orchard Road
Istana
Singapore 238823

Dear Mr Prime Minister,

RETENTION OF SECTION 377A, PENAL CODE

As concerned citizens of Singapore, we support the government in wanting to retain S377A of the Penal Code for the good of our children, our families and all Singaporeans.

There are many reasons why the retention of S377A is so important:

S377A is a reflection of the sentiments of the majority of society. Most Singaporeans hold conservative family values and do not accept homosexuality as the norm. (see Singaporeans' Attitudes toward Lesbians and Gay Men and their Tolerance of Media Portrayals of Homosexuality, by Benjamin H. Detenber, Mark Cenite, et. al., International Journal of Public Opinion Research) Repealing S377A is a vehicle to force homosexuality on a conservative population that is not ready for homosexuality.

Sexual preference is not about civil rights and has nothing to do with equality or tolerance. Repealing S377A would in fact be the first step towards mainstreaming the homosexual lifestyle, which has been shown elsewhere to lead to:

- Calls to specify the minimum age for consensual homosexual sex;
- A public education system that teaches acceptance of the homosexual lifestyle under the banner of "tolerance";
- The redefinition of marriage to include (gay) civil unions and same-sex marriages, and to extend marriage and parenthood benefits to them;
- Adoption by same-sex parents.

In short, repealing S377A could lead to the modification of core family values and the family unit as we know it.

The majority of Singaporeans want our children to grow up in a traditional environment that espouses healthy and wholesome traditional family values. We do not want the homosexual lifestyle to be promoted or celebrated.

We ask the Government to do what is right and retain S377A for the future of our children and our nation.

Yours faithfully,

15,559 Signatories

Current, Regulatory, and Historical Recordkeeping in S377a

- Social political actors in a historical act:
  - Singapore Government (SG)
  - Repeal 377a petitioners (R377a)
  - Keep 377a petitioners (K377a)
  - Social, political scientists and historians (SPSH)
- SG needs to perform ***current*** recordkeeping on keep377a.com and repeal377a.com websites
- SG needs to perform ***regulatory*** recordkeeping on Singapore parliament websites
- SPSH need to perform ***historical*** recordkeeping on all relevant websites

Representative Web Archiving Programs, Projects and Organizations

- International Internet Preservation Consortium (IIPC)
- Internet Archive (IA)
- PANDORA Project and National Library of Australia (NLA)
- ECHO Project and National Digital Information Infrastructure & Preservation Program (NDIIPP)

Mostly, focusing on Selection and Collection of Web materials, few on Description and Arrangement

International Internet Preservation Consortium (IIPC)

- In July 2003 the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA) acknowledged the importance of international collaboration for preserving Internet content for future generations. This group of 12 institutions chartered the IIPC to fund and participate in projects and working groups to accomplish the Consortium's goals. The initial agreement was in effect for three years, during which time the membership was limited to the charter institutions. Since then, membership has expanded to include additional libraries, archives, museums and cultural heritage institutions involved in Web archiving. For a complete list visit our members page.

  The goals of the consortium are: To enable the collection, preservation and long-term access of a rich body of Internet content from around the world.

- To foster the development and use of common tools, techniques and standards for the creation of international archives.
- To be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content.
- To encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation

URL: http://www.netpreserve.org/

Recordkeeping in IIPC

- Degree 1 Activities:
  - Web Harvesting Technology: *Heritrix*
  - Web Archives Accessing Technology: *Wera & OpenWayback*
  - Web Curation Systems: *NetarchiveSuite*, *PANDAS*, *Web Curator Tool*, *Web Archiving Service Archive-It*

- Degree 2 Activities:
  - WARC (ISO 28500) format: The WARC (Web ARChivewarc" and have the MIME type application/warc. The WARC file format is a revision and generalization of the ARC format used by the Internet Archive to store information blocks harvested by web crawlers.

Internet Archive (IA)

The **Internet Archive** (**IA**) is a nonprofit organization dedicated to maintaining an on-line library and archive of Web and multimedia resources. Located at the Presidio in San Francisco, California, this archive includes "snapshots of the World Wide Web" (archived copies of pages, taken at various points in time), software, movies, books, and audio recordings. To ensure the stability and endurance of the archive, IA is mirrored at the Bibliotheca Alexandrina in Egypt, the only library in the world with a mirror. The IA makes the collections available at no cost to researchers, historians, and scholars. It is a member of the American Library Association and is officially recognized by the State of California as a library.

URL: www.archive.org

Recordkeeping in IA

- Degree 1 Activities:
  - Web Harvesting Technology: *Heritrix*
  - Web Archives Accessing Technology: *OpenWayback*
  - Web Curation System: *Web Archiving Service Archive-It*
  - Web Collection: wayback machine at www.archive.org

- Degree 2 Activities:
  - WARC (ISO 28500) format

PANDORA Project at NLA

- PANDORA derives from its goal of "Preserving and Accessing Networked DOcumentary Resources of Australia".
- PANDORA is a highly selective archive: started June 2001, it contained just 19,257 web sites. Nevertheless, it already constitutes a strongly representative sample of Australian web publishing by
  - academic,
  - government,
  - commercial and
  - community organizations.
- Several of the web sites captured in the archive - including the official web site for the Sydney Olympic Games - have already disappeared from the live Internet.

  URL: pandora.nla.gov.au

Similar efforts:

Minerva Project of Library of Congress URL: http://www.loc.gov/minerva/

National Archives and Records Administration URL: http://www.webharvest.gov/collections/peth04/

Snapshots of PANDORA (I)

Snapshots of PANDORA (II)

Recordkeeping in PANDORA

- Degree 1 Activities:
  - Web Curation Systems: *PANDAS*
  - Web Collection: pandora.nla.gov.sg
- Degree 2 Activities:
  - Integration with National Bibliographic Database (Kinetica)
  - URL: http://www.nla.gov.au/librariesaustralia/

ECHO Project at NDIIPP

- Articulating a rationale and methodology for selecting digital materials for preservation, whether Web-accessible or not, as aggregates, rather than at the item level, based on archival principles, and using provenance, functional analysis, and context analysis to facilitate meta-tagging for retrieval.
- This methodology differs from approaches utilized to date. Manual, item-level selection fails because information professionals cannot keep up with the enormous number of resources on the Web. A fully automated approach to capture all the Web results in substantive materials being buried under a mountain of ephemeral, redundant, or irrelevant information.
- Instead, this selection methodology is based on an archival approach to the Web. In this approach materials are managed as they are in paper-based archives: as a hierarchy of aggregates rather than as individual items. This approach reduces to a more practical size the sheer volume problem of preserving Web materials, while maintaining a scalable degree of human involvement.

  URL: http://www.ndiipp.uiuc.edu/

Similar efforts:

Research Project on "Guidelines for Electronic Records Management on State and Federal Agency Websites" by Charles R. McClure and J. Timothy Sprehe, 1998

Recordkeeping in ECHO

- Degree 3 Activities:
  - Archival Description Framework: The Arizona Model

    - Materials are managed as a hierarchy of aggregates. In general, archivists do not manage collections at the item level unless the individual items are of great importance.
    - Respect for provenance requires that documents from one source are not mixed with documents from another source.
    - Respect for original order requires that documents be kept in the order that the creator used to manage the materials.
    - Respect for provenance and original order ensures that documents remain in context, and that the context can yield a richer understanding of the individual documents.

    Reference: Richard Pearce-Moses and Joanne Kaczmarek. (2005) “An Arizona Model for Preservation and Access of Web Documents” *DttP: Documents to the People* 33:1 p. 17-24.

Concluding Remarks

- State-of-the-Arts Web Archiving focus on Degree 1 and 2 Recordkeeping activities – selection and collection of Web materials
- Few Degree 3 Recordkeeping activities in Web Archives: Arizona Model in ECHO that considered provenance and functions when selecting and describing Web sites
- No known Degree 4 Recordkeeping activities
- Singapore Team intends to case-study eLearning space that require all degrees of recordkeeping activities:
  - IT Specialist (and Preserver): Center of Education Development (CED)
  - Creators (and Preserver): Division of Information Studies: instructors, students, and speakers
  - Researchers (and Preserver): Singapore Internet Research Center at School of Communication and Information.