Archiving the Web: What can Institutions learn from National and International Web Archiving Initiatives

Maureen Pennock   Michael Day   Lizzie Richmond

UKOLN   UKOLN   University of Bath
University of Bath   University of Bath

IWMW 2006, University of Bath, 15 June 2006

Today's workshop

- Records Management and the web:
  - Key RM principles
  - Justification for archiving web-based records
  - Breakout 1 - to discuss the types of record found on the web
- An archivist's perspective:
  - Authenticity, accessibility, security, legal compliance
  - Breakout 2 - to discuss drivers and barriers
- An overview of selected national and international web archiving initiatives:
  - Breakout 3 - to develop approaches to preserving web sites
  - Feedback

Web-Based Records

Philosophy

- Archiving web sites & web-based records requires **collaboration from all stakeholders**, including records managers, but also IT managers, web-project managers, webmasters, content editors, content providers, and even senior management, **across the entire life-cycle of the records**

- BUT … there is a difference in approaches between archiving websites and archiving web-based records

What is a record?

- BS ISO 15489 definition: "any information that is created, received and maintained as evidence and information by an organisation or person in pursuance of legal obligations or in the transaction of business"

- Evidence of a transaction

- Anything that:
  - documents a working transaction between two or more parties
  - documents the mission and goals of an organisation
  - was created or received in the course of carrying out the mission and goals of an organisation

Key Records Management issues

- Proper care and management of records throughout their entire life-cycle
- Not all data has to be retained
- Legal information obligations must be met
- Organisational retention schedules - identifies record classes of concern
- Different records and record classes have different retention periods
- Metadata must be stored with records
- Disposal and destruction processes

Leads to archival and long-term storage for some records

Why archive website 'records'

- Records are increasingly posted on the web
- Uniquely available informative records
  - Users may act or take decisions based on this information, with important consequences
- Records of business transactions
- Accountability & transparency
  - To funding bodies
  - To stakeholders
  - For legal reasons
- Historical and culturally valuable

Breakout 1

- Discuss and identify the types of records that can appear on the Web – e.g.:
  - Reports, policy documents etc
  - Information – submission dates, pricing etc

- Discuss and identify the forms can they take – e.g.:
  - Text-based files
  - Web-forms

Feedback I

Archiving the Web

An (inexperienced) archivist's perspective

More definitions…

- ## Records management:
  "…the field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records." (BS ISO 15489 - 2001)

- ## Archives:
  "…documents, irrespective of form, medium or age, intended for long-term preservation because of their continuing value." (BS 5454 - 2000)

What we want from our records and archives …

Authenticity:

- Must be demonstrably reliable as proof
- Creation and capture
- Metadata and context
- Ownership/responsibility
- Version control
- Cataloguing standards

What we want from our records and archives …

Accessibility:

- Must be capable of use over time
- Locate, retrieve and display
- File plans, naming conventions
- Obsolescence
- Migration strategy
- Reduced functionality?
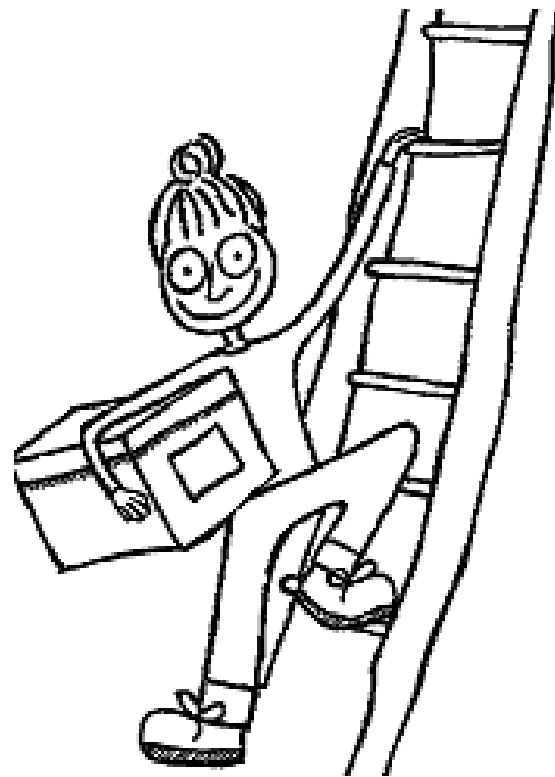
What we want from our records and archives …

Security:

- Must be protected
- Physical damage and unauthorised access
- Robust destruction procedures
- Intellectual control
- Storage environment
- Disaster plan

What we want from our records and archives …

Legal compliance:

- Must not break the law
- Freedom of Information Act 2000
- Data Protection Act 1998
- Copyright issues?
- Defence against litigation
- Legal admissibility

Breakout 2

- What are the main drivers for archiving web-based records?

- Discuss and identify as many challenges or barriers to archiving web-based records as you can:
  - Technical barriers
  - Cultural barriers
  - Socio-economic barriers
  - Organisational barriers

Feedback II

Current Approaches to Archiving the Web

National and International Initiatives

Some basics

- Not all web archives are organised on a records management basis
- Most web archiving initiatives:
  - Emphasise the informational value of the web as a cultural phenomenon or communication medium
  - Highlight the transience of content
  - Focus largely on collecting content, less on providing long-term access (or preservation)
  - Have collection strategies that are based on what can be automatically captured from the client side
  - Have problems with the deep (or hidden) web, i.e. those driven by databases or otherwise interactive … so what about Web 2.0?
  - Tend to ignore differences in type categories or formats
  - Have significant legal problems with providing access

Approaches to collection

- Broadly four main collecting approaches (not mutually exclusive):
  - Domain capture (harvesting)
    - Using specialised crawler programs to collect sites within national (or other) identifiable domains
    - Often based on the 'national' web domain
    - Can usually only deal with the surface web
  - Selective capture (harvesting)
    - Capturing selected web sites on a given frequency
    - Can usually only deal with the surface web
  - Selective capture (conversion or re-engineering)
    - Typically requires access at the server-side
    - Can deal with the deep web
  - Deposit by website owner

Two main models

- Harvesting model
  - Used by national and research libraries, university special collections (e.g., DACHS) and the Internet Archive
- Records management model
  - Addresses the issues raised earlier in this session
  - May be more appropriate for specific institutional records …

Some Examples …

Internet Archive

- Non-profit organisation, based in US
- Wants to offer permanent access to digital online materials of all types
- Founded in 1996, has been collecting since then … much content donated by Alexa Internet
- Collects sites by crawling and harvesting web sites
- Sites can 'opt out' by way of robots.txt file on the web server
- Most content is freely available to the public, e.g. through the Wayback Machine
- Interface issues: only the URL indicates that the page is archived
- Website: http://www.archive.org/

National Library of Australia

- The PANDORA Archive
  - Builds on existing NLA collection policies
  - Provides long-term access to selected online publications and websites
  - Permission is sought from site owners in advance
- PANDAS (v3) –PANDORA Digital Archiving System
  - Open Source Software used for managing the process of gathering, archiving and publishing website resources
  - Offers end-to-end archiving workflow
  - Supports modularity: currently mostly used with HTTrack, but other harvester programs can be plugged-in
  - Assigns persistent identifiers and metadata to each item when registered
- Website: http://pandora.nla.gov.au/

UK WAC

UK WEB ARCHIVING
CONSORTIUM
www.webarchive.org.uk

- UK Web Archiving Consortium (6 members)
  - British Library, National Library of Scotland, National Library of Wales, The National Archives, Wellcome Library, JISC
- Collects Web content selectively
  - Uses modified PANDAS collection/harvesting software developed by the National Library of Australia
  - Underlying harvesting program is currently HTTrack
  - Permission is sought from site owners in advance
  - The collections are publicly accessible
  - Persistent Identifier URLs
  - Central repository of metadata
  - Single partner assumes responsibility for each site
- Website: http://www.webarchive.org.uk/

Nordic Web Archive

< NWA >

- A collaboration between the Nordic national libraries (Denmark, Finland, Iceland, Norway, Sweden)
- Considerable expertise available:
  - For example, the Swedish Royal Library pioneered the national domain capture approach
- Main focus on developing access tools
  - NWA Toolset (open source)
  - Work now taken forward as part of the WERA viewer application developed as part of the International Internet Preservation Consortium
- Website: http://nwa.nb.no/

IIPC (1)

netpreserve.org
international internet preservation consortium

- International Internet Preservation Consortium
  - Builds co-operation between the Internet Archive and national and research libraries
  - Co-ordinated by the Bibliothèque nationale de France
  - The British Library is the only current UK member, other national library partners include the Library of Congress, the Library and Archives Canada and the national libraries of Australia, Denmark, Finland, Iceland, Italy, Norway and Sweden
  - Reflects those with current experience of Web archiving
  - Both working-groups and tool development
  - Phase II will enable new partners to join the consortium

- Website: http://netpreserve.org/

IIPC (2)

netpreserve.org
international internet preservation consortium

- Phase I - developing the IIPC toolkit
  - Standards and tools for supporting:
    - **Acquisition** - archival quality crawler (Heritrix); portable database extraction and migration tool for database-driven deep web sites (DeepARC)
    - **Managing collections** - analytical and prioritization tools for automatically focusing harvesting; curation tools to provide a non-technical interface for selecting, monitoring and verifying archived web sites
    - **Collection storage and maintenance** - tools for manipulating formats; a standardised storage format (WARC), standards for metadata
    - **Access and finding aids** - browse interfaces (WERA) and search facilities (NutchWAX)

The National Archives (UK)

- Managing web resources (December 2001)
- ERM toolkit for government agencies
- Practical steps for active records management and sustainability
  - Useful identification of web-based records
  - Scenarios
  - How websites differ from other records
  - Management control mechanisms
  - Model action plan
  - Sustainability
- Website: http://www.nationalarchives.gov.uk/

National Archives of Australia

- A Policy for keeping records of web-based activity (January 2001)
  - Provides clear directions to Commonwealth agencies to implement mechanisms for creating, managing and retaining web-based records of value
- Guidelines (March 2001)
  - Challenges and responsibilities
  - Types of web-based resources
  - Fundamentals of good record-keeping
  - Assessing risk – factors to consider
  - Strategic & technical options
  - Storage & preservation - issues & strategies
  - Determining the best option

Australian Government

National Archives of Australia

Managing web-based records

- Fundamentals:
  - Information Audit and Risk Assessment
  - A systematic approach
    - Develop policy
    - Formulate plan for capture, maintenance, and preservation
    - Implement appropriate website maintenance procedures
  - Assign and document responsibilities
  - Identify records
  - Determine retention requirements
  - Capture records into recordkeeping system
  - Add metadata
  - Transfer content and metadata into archive as appropriate

* Based on NAA Guidelines for Archiving Web Resources

Breakout 3

- Scenarios for each group
  - Read brief
  - Identify main actions for each stage of life-cycle that play a role in archiving web-based resources
  - Identify aspects of a successful long-term preservation strategy
  - What aspects of a harvesting model could be of use? How? Why?
  - What other technical development is needed?

Feedback III

Your approach?

Go forth and archive!

Maureen Pennock
M.Pennock@ukoln.ac.uk

Michael Day
M.Day@ukoln.ac.uk

Lizzie Richmond
L.Richmond@bath.ac.uk