July 2010 Partners Meeting Breakout Session

Web Archiving Tools & Access, session #3
NDIIPP Annual Meeting
Wednesday, July 21, 2010
2:45pm – 4:00pm

Presenters: Michael Nelson, Old Dominion University
Kate Odell, Internet Archive
Jennifer Ricker, State Library of North Carolina
Mike Smorul, University of Maryland

Attendees: 33

Memento: Time Travel for the Web
Michael Nelson, Old Dominion University

- Memento: refers to an archived version of a resource
- TimeGate: refers to a transparently negotiable resource that supports the datetime dimension.
- TimeBundle: refers to a resource via which an overview of all Mementos for an original resource URI-R is available.
- There are two components to the Memento Solution:
    - Component 1: Navigation towards an archived resource via its original resource, by leveraging content negotiation. DONE
    - Component 2: A discovery API for archives that allows requesting a list of all archived versions it holds for a resource with a given URI. DONE
- Goal: Memento wants to make navigating the Web's Past Easy

Web Archiving and Access
Mike Smorul, University of Maryland

- Being able to navigate archives that are usable and flexible is key; to do so, we need new search paradigms
- WebArc manager
    - We have no idea what we actually have in our several terabyte clouds—what can we do?
    - goals
        - develop a tool to help manage webarc collections
        - show statistics of a series of crawls
        - open API to easily query collection
- WarcManager (server)
    - REST-based access to index
    - Index of DAT/ARC entries
    - URL Searching, ARC browsing,
- Javascript Client

- Simple Web-Accessible Preservation (SWAP)
    - Web-accessible distributed storage
    - ARC page retrieval
    - 1Gbps, 2200requests/s
- storage design
    - fairly simple, allows us to get away with not having a central database to log all the files
    - use 320redirects to find the correct server that's holding information
- time machine for the web
    - fast parallel indexer to handle large scale crawled web contents, coupled with a new compression scheme.
    - fast search of contents based on unstructured queries involving temporal specifications.
    - presentation of pertinent summary information in ranked order according to the temporal context.
- Additional information:
    - http://adapt.umiacs.umd.edu [papers, results, etc.]
    - e-mail: msmorul@umiacs.umd.edu
- questions at session:
    - Q: What hardware are you using? A: SWAP-house system helps us get around having a simple file server/central hard-drive
    - Q: How do you download information? A: The individual file or the entire ARCfile. Cannot currently create an original ARCfile [looking to enable that during next revision; waiting on case studies from the LOC]. Right now, it still only permits browsing.

A Flickr of Hope: Harvesting and Archiving Social Networking Sites
Jennifer Ricker, State Library of North Carolina

- three statutes identify DCR as authority of public records
- Obama in DC; Perdue in NC
    - interest in e-governance, Web 2.0, & social networking sites (SNS) skyrockets.
    - outreach opportunity; communication forum
- documenting social change, social reaction to important events, and a cultural phenomena
- social networking sites as public records?
    - view websites and the various documents posted on them (.doc, .pdf, etc.) as publications.
    - social networking sites are being used to conduct state business (marketing, information push, etc.) and, therefore, meet the statutory definition of public record.
    - decided not to take the "wait and see" approach
- State Library staff willingly offers itself out as beta testers of just about anything we can. It enables us to stay up-to-date on new tools and have our voice heard at the design stages.

Kate Odell, Internet Archive

- Internet Archive
  - digital library
  - mission statement: Universal access to human knowledge
- tools behind Archive-It
  - Heritrix
  - Wayback Machine
  - NutchWax
- website kept simple and low-barrier, to ensure maximum access to the wealth archived therein
- social media
  - currently, 20 partners are archiving from these sites
  - they are a moving target: the ballgame changes every few months, and new challenges to archiving these pages, intact, arise
  - great screenshot examples of web2.0 archived pages
- challenges discussed: content behind log-ins, not "archive-friendly" sites, sites change technical structure and policy often, scoping rules must be created [and that's complex!]
- overall approaches: trial and error, quality reviews, collaboration, document document document
- large challenge ahead: many formats for archiving have to be negotiated
- internet universe to be archived: when we think of the scale of our work, it is daunting
- www.archive-it.org http://www.facebook.com/ArchiveIt
- Kate Odell
  Partner Specialist, Internet Archive
  kate@archive.org

Action Items

Smorul: Cannot currently create an original ARCfile. Looking to enable that during next revision [waiting on case studies from the LOC, which Gina Jones is involved with]. Right now, the application only permits browsing.

Ricker: Interested in harvesting public records from more social networking sites, such as vimeo, scribd, facebook, etc. [do not harvest any social media site that doesn't host original content, or that which is not appropriate for business usage], and overcoming challenges related to harvesting those they already focus on [including keeping record of who all has social networking accounts and where].

Odell: Collaboration is needed, across all institutions/agencies/organizations, because no one ins/age/org can do this alone; we need to produce and record best practices and lessons learned to share with other partners.