Moving Web Archiving into the Classroom at UNC: Experiences & Lessons

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

DigCCurr 2009:

Digital Curation Practice, Promise and Prospects

April 1-3, 2009

Chapel Hill, NC USA

DigCCurr
say : dij-seeker

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

Our Context

- School of Information and Library Science
- Undergraduate, masters & doctoral students
- All students involved in the classes I'll be discussing have been masters students (48 credit hours, which is 2 years, full-time)

Integrating Web Archiving into Existing Classes

- INLS 525 – Electronic Records Management
- INLS 513 – Resource Selection and Evaluation

Electronic Records Management

- Class is about electronic recordkeeping, so emphasis of assignment is on how web archiving fits into a broader recordkeeping regime
- Assignment introduced in Spring 2007 - using both Archive-It and Web Archives Workbench (OCLC)

- Read: NARA Guidance on Managing Web Records (January 2005)

Trying out the software:

- Familiarize themselves with the two tools
- Run test crawls in Archive-It, if they wanted

- Individual paper:
    - Summary of NARA guidance
    - Identify records on a government agency’s web site
    - Indicate what the agency is
    - 3 examples of records on the agency’s site that you think would be important to preserve for several decades (or longer)
    - Attempt to associate the records with records series
    - Analysis and comparison of Archive-It and the WAW

Resource Selection and Evaluation Assignment

- Class is about collection development, so emphasis is on how selection of web content can fit into a broader collecting mission
- Three iterations of the assignment – Fall 2007, Spring 2007, Fall 2008

Round 1 of the Assignment – Spring 2007

- Final product was a paper
- Groups identified 3 different crawls/seeds that they would like to capture on behalf of the institution they were assigned for the major assignment for the class (a collection development policy)
- They set up 3 crawls in both Archive-It and the WAW – then ran the crawls for them
- The paper included:
- Selection and decision making process:
    - Scope
    - Stakeholders
    - Selection criteria
    - Sources of existing information
    - Evaluation
- Use of software to do the crawls:
    - Why they selected these 3 crawls
    - How often they should be crawled and why
    - How and why they might want to narrow, scope or filter the crawl
- Analysis and comparison of Archive-It and the WAW
    - General observations, similarities and differences
    - Considering the process that they described in part 1 (above), how might each of the two tools support that process? What would they expect to be the primary challenges in trying to use each of the two tools to support your decision making and selection process?

Lessons from Spring 2007

- Great value in students seeing how the tasks would actually be done, rather than just talking/reading about them
- Many students were reluctant to jump in and try out the software – partially a result of al working within the same "collection" (didn't want to break things for others)
- Starting the crawls for the groups was probably a mistake

Round 2 of Assignment – Fall 2007

- Just used Archive-It, and provided much more detailed instructions (step-by-step with screenshots)
- Divided the class into six groups - each charged with identifying **four** different crawls/seeds
- There were two groups assigned to a given topic – and both of those groups shared a “collection” within Archive-It
- The two groups on a given topic came together with their own four crawls (4 + 4 = 8 total) and then had to come to agreement on a common set of **five** crawls to actually run through Archive-It
- A designated student initiative the crawl for each topic
- Each of the six smaller groups turned in a paper that addressed:
    - Scope
    - Selection criteria
    - Sources of existing information
    - Evaluation
    - For each of the specific four seeds they identified:
        - Why they were selected
        - How often they should be crawled and why
        - How they might want to narrow, scope or filter the crawl and why

Lessons

- Detailed instructions and using 3 collections was helpful
- Only one student for each topic actually ran the crawl, so most had little incentive to learn how to use the software or understand in much detail how seeds relate to collecting activities
- Topic I had chosen for them (e.g. Beijing Olympics, California Wildfires) were ones already be addressed by other collections in Archive-It, which made the idea of creating this collection less compelling to them

Round 3 of Assignment – Fall 2008

- Substantially similar to Round 2
- Focused topics on North Carolina topics, to make their uniqueness within Archive-It more likely
- Added individual paper component to ensure more engagement of all students in each group:
  - Quantitatively summarize what web resources their group collected, in terms of hosts, URLs, bytes, blocks based on Robots.txt, and file types
  - For each of the group's five seeds, identify and describe:
    - One example of a page that you are glad you collected and explain why it fits your collecting goals
    - One example of a page that you collected, but which you don't believe fits very well into your collecting goals and explain why you believe it doesn't
    - Something about crawl results that surprised them and why
  - How and why they could have changed characteristics of crawls to generate results more appropriate to documenting their topic (in terms of specific parameters that could be set or they would like set within Archive-It)

Thank you to Internet Archive and OCLC!