School of Information and Library Science
University of North Carolina, Chapel Hill
INLS 690-208: Web Archiving
[Last Updated: 2013-08]

Meeting Time: Tuesday 2:00-4:45, August 20-Octobor 8
Location: Manning 304
Credits: 1.5
Perquisite: None
Instructor: Ayoung Yoon
E-Mail: ayyoon [at] email [dot] unc [dot] edu
Office Hours: Tuesday 4:45 to 5:30, or by appointment
Course Web Site: https://sakai.unc.edu/

Course Description

This course aims to provide knowledge of the role and potential of the Web as a source for archival collection development. As we live in the Internet world, archivists need to be familiar with the tools and appropriate techniques for preservation of information delivered through the "surface" Web (static Web pages, blogs, etc.) and information that is part of the "deep" Web (e.g., databases, authenticated resources, etc).

Through lectures, presentations on specific topics, analysis of Web archives, and hands-on work, students will gain insight on relevant issues arising from the nature and characteristics of the Web and how to make correct and fruitful use of the Web in their archival work. This course also discusses a variety of topics, issues, and challenges around Web archiving and preservation. About half of this course will expose students to existing and emerging tools for capturing Web content, with an emphasis on laboratory practice using Web crawlers. Students will also learn about current preservation formats of Web-based content.

Course Objectives

Upon completing this course, students will be able to:

1. Discuss the role and the potential of the Web as information, and characteristics of the Web for archiving and preservation.
2. Recognize the challenges of acquiring, downloading, storing, and providing access to Web-based content.
3. Become proficient at setting up a Web crawl using Archive-It.
4. Create a Web archive.
5. Increase their awareness of legal and policy constraints on Web archiving.
6. Become familiar with standards and best practices for sustainability of archived Web content.

Classroom & Laptop Etiquette

Due to the nature of this class—that is, lab sessions—use of personal or lab computers is necessary. Students are encouraged to bring laptop computers to class for lab sessions. However, students should not use the computers for purposes outside the classroom. Students can use them actively as learning tools over the course. Students should:

- Use computers for taking notes, conducting research required for activities, and other classroom-specific tasks as assigned by the instructor. During class, students should not check e-mail, chat, IM, play games, or perform other off-task activities.
- Engage in class activity as actively as they would in any other class. The computer should not become a barrier to one-on-one interaction but instead should help facilitate the exchange of ideas and engagement in classroom contact.

Readings and Sources

Readings are on Sakai or links on the World Wide Web. It is expected that students will have read the materials before class, as we will be referring to them in lectures and in the exercises.

This class has one required Textbook. This textbook is available electronically via Sakai.

- Julian Masanés, ed. *Web Archiving,* Springer, 2006.

Other useful sources for the class:

- PoWR, the Preservation of Web Resources Handbook. The JISC-PoWR Team. (2008).
- WARC implementation guidelines v.1. Clément Oury, Bibliothèque nationale de France (National Library of France). (2009).
- Web Archiving Resources. Harvard University Library. (2009). http://hul.harvard.edu/ois/systems/wax/resources.html
- Top 10 Tips For Preserving Web Sites. Cultural Heritage briefing document no. 32, UKOLN. (2008).
- Web Archiving. Alex Bal. UKOLN, University of Bath.
- International Internet Preservation Consortium (IIPC). http://netpreserve.org/about/index.php
- Preserving Access to Digital Information (PADI): Web Archiving, The National Library of Australia. http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html
- Web Archiving Bibliography, Austrian On-Line Archive. http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html
- International Web Archiving Workshop (IWAW). http://bibnum.bnf.fr/ecdl/
- Web Archives Cooperative: Making Web Archives Useful Today. http://infolab.stanford.edu/wac/

Assignments and Grading

Grades will be based on class attendance and participation and a series of assignments.

Assignment	Due Date	Percent
Class attendance and Participation	Ongoing	15%
Review of Web archives	Week 2. (Aug 27) Before class	30%
Final Web Crawl Project		55%
a. Selection / Scope update: from the lab session	Week 3. (Sep 3) After class	5%
b. Metadata update: from the lab the lab session	Week 4. (Sep 10) After class	5%
c. Preservation plan: from the lab the lab session	Week 6. (Sep 24) After class	5%
d. Full Web Crawl Report	Week 8. (Oct 8) Before class	35%
e. Presentation	Week 8. (Oct 8) During class	5%

Class Attendance and Participation (15% of total grade)

Class attendance is mandatory. Students are also expected to complete all required readings. Students should be prepared to enter into class discussions and raise questions reflecting their reading and interests. This is important since a portion of the lab sessions (or exercises) will be based on readings and lectures. We will also have a series of lab sessions to work on a group project, discuss issues on Web archiving and preservation, and hear about each other's group projects. Learning from each other's experiences is important as well as learning by yourself. Please share your thoughts and experience during the lab discussion as well as listen to others.

Assignments

* All assignments must be turned in through the Sakai class website. Late submissions will not be accepted unless students have consulted with the instructor prior to the late submission.

1. Review of Web archives (30%): Due Week 2(Aug 27), before the class

- A brief paper (2-3 pages, font-size 12, time new roman, double-spaced)
- Instructions will be distributed in Week 1.

2. Web Crawl Project (55% of total grade)

Students will work on term projects throughout the semester.

The project is to develop a Web archive using Archive-It.

2-1) 3 small assignments (5% each)

- Selection / Scope update: Due Week 3 (Sep 3), after the class
  - A brief report (1-2 pages, font-size 12, time new roman, double-spaced)
  - Instructions will be distributed during the lab session in week 3.
- Metadata update: Due Week 4: Due Week 5 (Sep 10), after the class
  - A brief report (1-2 pages, font-size 12, time new roman, double-spaced)
  - Instructions will be distributed during the lab session in week 4.
- Preservation plan: Due Week: Due week 6 (Sep 24), after the class
  - A brief report (1-2 pages, font-size 12, time new roman, double-spaced)
  - Instructions will be distributed during the lab session in week 5.

2-2) Final project report (35%): Due Week 8 (Oct 8), **before** the class

Write a final report of your collection with analysis of tools and your experiences. Paper should be 10-15 pages (double-spaced, 12-point Times New Roman font), and you can attach an Appendix, if needed.

Your report should include, but not limited:

- Indicate the completion data of the crawl or crawls that you are analyzing as your best shot at a collection.
- Project description
  - Describe the scope and content of your collection (topic, scope, intended audience, extent, exclusions).
  - Discuss any content that you would like to have included but could not harvest with the web crawling tool (e.g. hidden web content). Briefly describe a plan for how you might acquire this content. Would you continuously build the collection or create one or more snapshots? When and how often would you add content?
  - Based on what you have learned in the course, how would you revise your collection scope?
- Quality control, organization, metadata
  - How would you address issues that you found in post-crawl analysis?
  - Explain how you would organize, catalog, or add metadata to the collection to improve its usability and appearance.
  - Are there specific URLs or web pages that you would add item-level metadata to?
- Provide a preservation plan for the collection
  - What preservation level do you assign to the most common file types in your collection?
  - What is your overall preservation strategy (convert to standard formats, migration, maintenance plan, etc.)?
- Your impression, evaluation on the tool
  - How much do you satisfied with the tool? Does it do what you expect to do? Any difficulties or problems using the tool? How can it be improved?

Citations should conform to a standard style manual or commonly accepted disciplinary format such as the Chicago Rules of Style or the Publication Manual of the American Psychological Association (APA). A bibliography or works consulted list is required.

Papers will be evaluated on the following criteria:

- Appropriateness of your decisions on your collection
- Integration of recourse to support your decisions (not limited to the class readings)
- Clarity of writing
- Presentation (citation, proofreading, bibliography, etc).

2-3) Presentation (5%): Week 8 (Oct 8).

Each person has 20 minutes to present a summary and highlights from the projects.

Presentation should include:

- Brief project description (with screen shot(s) of your collection)
- Procedures of crawls
- Your decisions on collection scope, quality control, organization, metadata, and preservation plan
- Lesson learned & evaluation on the tool

Presentation will be evaluated on:

- Clarity of content (both slides and oral presentation)
- Clarity of presentation (no filler words, extended pauses, etc.)
- Speaker's poise – no fidgeting, making eye contact with audience, etc.

*** Special Needs:** If you need an accommodation for a disability or have any other special need, please make an appointment to discuss this with me. I will be most able to address special circumstances if I know about them early in the semester. My office hours and contact information are listed at the beginning of this syllabus.

Important note on plagiarism

Unless otherwise specified in an assignment, all submitted work must be your own, original work. Any experts from the work of others must be clearly identified as a quotation, and a proper citation provided. Be aware of the University of North Carolina policy on plagiarism. All cases of plagiarism (unattributed quotation or paraphrasing) of anyone else's work, (e.g. from published materials) will be officially reported and dealt with according to UNC policies (Instrument of Student Judicial Governance, Section II.B.1. and III.D.2, http://instrument.unc.edu).

Evaluation

Based on UNC Registrar Policy for graduate-level courses (http://registrar.unc.edu/AcademicServices/Grades/ExplanationofGradingSystem/index.ht

m#grad), both assignment and semester grades will be H, P, L or F. Few students will obtain an "H," which signifies an exceptionally high level of performance (higher than an "A" in an A-F systems). The following is a more detailed breakdown used for class assignments:

H Superior work: complete command of subject, unusual depth, great creativity or originality

P+ Above average performance: solid work somewhat beyond what was required and good command of the material

P Satisfactory performance that meets course requirements (expected to be the median grade of all students in the course)

P- Acceptable work in need of improvement L Unacceptable graduate performance: substandard in significant ways F Performance that is seriously deficient and unworthy of graduate credit

F An unacceptable performance. The F grade indicates that the student's performance in the required exercises has revealed almost no understanding of the course content.

Course Schedules

Week 1. (Aug 20): Introduction & Basic concepts in Web Archiving; Surface and Deep Web

- Introduction to the class
  - Course logistics
  - Introduction of the students and instructor
  - Review of syllabus and assignments
- Basic concepts in Web Archiving
  - What is web archiving?
  - What do we need to know about Web for archiving it?
  - Surface Web and deep Web

Required readings

- Masanès, Chapters 1 and 9
- Lyman, Peter. Archiving the World Wide Web. In CLIR (Ed.), *Building a national strategy for preservation: issues in digital media archiving.* (pp. 38-51) Council on Library and Information Resources and Preservation Program, the Library of Congress. (2002).

Additional readings

- Smith, Elizabeth H. Lost in Cyberspace: Have Archives a Future? Paper delivered at the *Australian Society of Archivists Conference*, Melbourne, 19 August 2000. (not in UNC library)
- Rosenzweig, Roy. Wizards, Bureaucrats, Warriors & Hackers: Writing the History of the Internet. *American Historical Review 103*(5) (December 1998): 1530-52. Available at www.pne.people.si.umich.edu/PDF/ahrcwreview.pdf
- Gillies, James & Cailliau, R. *How the Web was born: The story of the World Wide Web*. Oxford: Oxford University Press. (2000).
- Castells, Manuel. *The Internet Galaxy. Reflections on the Internet, Business, and Society*. Oxford: Oxford University Press. (2001): 1-63, 247-82.
- O'Neill, E. T. Trends in the Evolution of the Public Web. *D-Lib Magazine*. (2003). Available at http://www.dlib.org/dlib/april03/lavoie/04lavoie.html
- Piche, Jean-Stephen. Doing What's Possible With What We've Got: Using the World Wide Web to Integrate Archival Functions. *The American Archivist 61* (Spring 1998): 106-22.
- Cho. J and Garcia-Molina, H. The evolution of the web and implications for an Incremental Crawler. Paper presented at the *Proceedings of the 26th International Conference on Very Large Data Bases*. (2000).
- Masanès, J. Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine 8*(12). Available at http://www.dlib.org/dlib/december02/masanes/12masanes.html

Week 2. (Aug 27): Archiving different types of Web materials

*** Review of Web archives assignment DUE before class**

- Types of materials on the Web
- Who does archives Web? Why they do? What are the needs?
- Web and organizational memory
- Archiving personal web & social media

Lab session (3:30)

- Archive-IT: **Guest speaker** from Internet Archives

Required readings

- Frank, P. *How Federal Agencies can Effectively Manage Records Created Using New Social Media Tools*. IBM center for the business of government. 2010: 15-34. http://observgo.uquebec.ca/observgo/fichiers/57504_GRI%201.pdf
- Lee, Christopher A. Collecting the externalized me: Appraisal of materials in the social web. In *I, Digital: Personal Collections in the Digital Era*, edited by Christopher A. Lee, 1-26. Chicago, IL: Society of American Archivists, 2011.
- Nathan, L. P., & Shaffer, E. Preserving Social Media: Opening a Muti-Disciplinary Dialogue. (n.d.).

Additional readings

- Suderman, Jim. Committing the Web to Memory: Transmitting Web-based Records over Time. Paper presented at IV Coloquio del Papiro a la Biblioteca Virtual, 21-25 March 2005, Havana, Cuba.
- Smithsonian Institution Archives. To Preserve or Not to Preserve: Social Media. http://siarchives.si.edu/blog/preserve-or-not-preserve-social-media.

Week 3. (Sep 3): Exploration of existing web archives / Selection & Scoping

- Review of existing web archives
  - What are the different approaches current web archives take?
- Selection policy and criteria
- Different selection approaches: domain, topic or event, media type and genre based

Lab session (3:30)

- Working on the initial selection policy for the projects.
- First lab assignment distributed.
- **First lab assignment DUE after class**

Required readings

- Please review **at least 3** from the following examples before coming to the class
  - The Internet Archive (IA): http://www.archive.org/
  - MINERVA, Library of Congress: http://www.loc.gov/minerva/
  - PANDORA (Preserving and Accessing Networked Documentary Resources of Australia), National Library of Australia with nine other Australian libraries and cultural collecting organizations: http://pandora.nla.gov.au/index.html
  - UK Government Web Archive: http://www.nationalarchives.gov.uk/webarchive/
  - UK Web Archive: http://www.webarchive.org.uk/ukwa/
  - Topical archives developed using two different approached
    - The September 11 Digital Archive: http://911digitalarchive.org/
    - September 11 Archive: http://september11.archive.org/
  - Media, form, and genre based
    - U.S. Fish and Wildelife services, Digital Media Archives: http://images.fws.gov/
    - NCSA Digital Video Archive: http://archive.ncsa.uiuc.edu/MEDIA/vidlib/
- Masanès, Chapter 3
- Library of Congress Collection Policy Statements Supplementary Guideline. (2008). http://www.loc.gov/acq/devpol/webarchive.pdf.

Additional readings

- The Web-at-Risk: A Distributed Approach to Preserving our Nation's Political Cultural Heritage Content Identification, Selection, and Acquisition Path, Collection Plans, http://web3.unt.edu/webatrisk/cpg.php
- Brown, A. Chapter 3. Selection. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006)
- Guidelines for a Collection Development Policy using the Conspectus Model. International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development. (2001).
- Lyle, J. (2004, September). Sampling the Umich.edu domain. Paper presented at the *4th International Web Archiving Workshop (IWAW04)*, Bath, UK. URL: http://iwaw.europarchive.org/04/Lyle.pdf
- Qin, J. et al. Building domain-specific web collections for scientific digital libraries: A meta-search enhanced focused crawling method. *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. (2004).
- Schneider, S.M. et al. Building thematic Web collections: Challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. Paper presented at the *3rd Workshop on Web Archives*. (2003).

Week 4. (Sep 10): Acquisition and Collection Methods / Metadata and Description

- Dynamics of websites and different technology
- What information should be provided? (Metadata!)
- How are users of web archives' needs different from users of active websites?

Lab session (3:30)

- Working on the metadata for the projects.
- Second lab assignment distributed.
- **Second lab assignment DUE after class**

Required readings

- Masanès, Chapter 4
- Bragg, Molly and Lori Donovan. "Archiving Social Networking Sites w/ Archive-It." Available at https://webarchive.jira.com/wiki/pages/viewpage.action?pageId=3113092
- Beryl A. Howell, (2006) "How to Use the Internet Archive," Journal of Internet Law (February): 3-9

Additional readings

- Brown, A. Chapter 4. Collection methods. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006)
- Brown, A. Chapter 7. Delivery to users. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006)
- Fitch, K. Web site archiving: An approach to recording every materially different response produced by a website. Paper presented at t*he AusWeb 2003: The Ninth*

Australian World Wide Web Conference, Sanctuary Cove, Australia. (2003). Available at http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/1258/1543

- Election 2002 Web Archive Cataloging and Description. MINERVA. Library of Congress. (2004). Available at http://lcweb4.loc.gov/elect2002/catalog/2860.html
- PANDORA cataloging manual. Available at http://pandora.nla.gov.au/manual/cattoc.html

Week 5. (Sep 17): Quality Control and Post Capture Processing

- Why is post-collection processing necessary?
- Different methods of quality control (types of tests)
- Issues and challenges

Lab session (3:30)

- Looking at Archive-It report
  - Blocked sites
  - URLs that are out of scope
  - Sites that you expected but were missed
  - Problematic file types

Required readings

- Brown, A. Chapter 5. Quality Assurance and Cataloging. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006).
- International Internet Preservation Consortium. Sketching and Checking Quality for Web Archives: A Frist Stage Report from BnF. (2006):13-34. bibnum.bnf.fr/conservation/bnf-qualityforwebarchives-feb06.pdf

Additional readings

- Martin, C., Lasfargues. F., and Medjkoune, L. What if Web Archiving were as reliable as pushing a simple button? IMF, France. Available at http://www.museumsandtheweb.com/mw2011/papers/what_if_web_archiving_were_as_reliable_as_push
- Willer, M., Buzina, T., Holub, K., Zajec, J., Milinovic, M., and Topolščak, N. Selective Archiving of Web Resources: A Study of Processing Costs. *Program: Electronic Library and Information Systems* 42(4). (2008): 341-364.

Week 6. (Sep 24): Preserving Web Sites

- Challenges of preserving websites (technical, financial, and organizational)
- Strategies (passive vs. active; emulation vs. migration)
- Significant properties of web

Lab session (3:30)

- Review of preservation statement
- Third lab assignment distributed
- **Third lab assignment DUE after class**

Required readings

- Masanès, Chapter 8
- PoWR, the Preservation of Web Resources Handbook. The JISC-PoWR Team. (2008).
- International Internet Preservation Consortium. Long-Term Preservation of Web Archives - Experimenting with Emulation and Migration Methodologies: IIPC Project to Evaluate Emulation and Migration as Long-Term Preservation Solutions for Web Archives. http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf.

Additional readings

- Hockx-Yu, H. and Knight, G. What to Preserve?: Significant Properties of Digital Objects. *International Journal of Digital Curation, 3*(1). (2008).
- Digital Preservation Testbed White Paper: Emulation: context and current status. (2003).
- Chapter 4. Assessing Risk- Factors to Consider. In Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government. National Archives of Australia. (2001): 20-22.
- Kenney, A.R. et al. Preservation Risk Management for Web Resources. *D-Lib Magazine*, 8(1). (2002). Available at http://www.dlib.org/dlib/january02/kenney/01kenney.html

Week 7. (Oct 1): Access and Use / Legal and Ethical Issues

- Web archives access tools
- Who uses web archives? How are web archives used?
- Intellectual properties / Privacy
- Content reliability

Lab session (3:30)

- Working on web crawling and the project

Required readings

- Masanès, Chapter 2 and 6
- Charlesworth, A. Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia. JISC, The Wellcome Trust, University of Bristol. (2003).

- A. Website: Five ways to stay out of trouble. Copyright & Fair Use, Stanford University Libraries. Available at http://fairuse.stanford.edu/Copyright_and_Fair_Use_Overview/chapter6/6-a.html
- IIPC. (n.d.) Lega Issues. Available at http://netpreserve.org/web-archiving/legal-issues / (Engle, E. (May 30, 2012). Legal Issues in Web Archiving. Available at http://blogs.loc.gov/digitalpreservation/2012/05/legal-issues-in-web-archiving/)

Additional readings

- Glanville, L. Web Archiving: Ethical and Legal Issues Affecting Programmes in Australia and the Netherlands. *Australian Library Journal* 59(3). (2010): 128-134.
- Copyright on the Internet. Thomnas G. Field Jr. Franklin Pierce Law Center. (2002). Available at http://law.unh.edu/thomasfield/ipbasics/copyright-on-the-internet.php
- Kavcic-Colic, A. Archiving the Web - some legal aspects. 68th IFLA Council and General Conference, Glasglow. (2002).
- Copyright Statement for the September 11 Web Archive, Library of Congress. Available at http://lcweb2.loc.gov/diglib/lcwa/html/sept11/sept11-overview.html#copyright
- Preserving Access to Digital Information (PADI): Intellectual property rights management, The National Library of Australia. Available at http://www.nla.gov.au/padi/topics/28.html
- U.S. Copyright Office (December 2004) "Copyright Registration for Online Works", Circular 66. Available at http://www.copyright.gov/

Week 8. (Oct 8): Presentations of Crawls / Class Evaluation