School of Information and Library Science
University of North Carolina, Chapel Hill
INLS 690-208W: Web Archiving
[Last Updated: 2014-08]

Meeting Time: August 25- October 13, 2014
Location: Web
Credits: 1.5
Perquisite: None
Instructor: Ayoung Yoon
E-Mail: ayyoon [at] email [dot] unc [dot] edu
Course Web Site: https://sakai.unc.edu/

Course Description and Objectives

This course aims to provide knowledge of the role and potential of the Web as a source for archival collection development, as it has already been part of archival collections for many government, educational, and cultural institutions. Archivists need not only to react to the needs of these many institutions but also to be proactive to these changes and their needs for collection development.

Through reading materials on a variety of topics, issues, and challenges surrounding Web archiving and preservation, analyzing Web archives, and performing hands-on work using a Web archiving tool, students will gain insight on relevant issues arising from the nature and characteristics of the Web and how to make correct use of the Web in their archival work. This course will expose students to existing and emerging tools for capturing Web content, with an emphasis on laboratory practice using Web crawlers (Archive-it). Students will also learn about current preservation formats of Web-based content.

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. 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.).
3. Recognize the challenges of Web archiving.
4. Become proficient at setting up a Web crawl using Archive-It.
5. Create a Web archive.
6. Increase their awareness of legal and policy constraints on Web archiving.

7. Be familiar with standards and best practices for sustainability of archived Web content.

Email

Please check the course listserv on a regular basis if not daily. This will be our primary means of communication.

The address is: inls690_208W14@sakai.unc.edu

If you have a question of general interest to the class such as "What do you mean by X in assignment Y?" please send this to the class list and I will answer it so that everyone can benefit.

How we will conduct "Class"

While online courses afford great convenience, they also demand extra effort from both instructors and students. Because there is no face time for lectures, discussions, group work, or other activities, all this must take place through the course site on Sakai. This involves extensive writing and creation of slides, videos, and other media we will use to communicate our ideas and questions. The syllabus, assignments, and many of the readings (unless available online) will be posted on Sakai. Each week I will provide slides or notes on important points, but much of the work of the course will take place in the forums via discussion of the readings, videos, and slides, and your own work (small assignments).

Online Etiquette (Netiquette) Guideline

- http://www.indiana.edu/~icy/netiquette.html
- http://www.brighthub.com/education/online-learning/articles/26946.aspx

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 Participation and Work Log (Sakai Forum)	Ongoing	15%
Review of Web archives	Week 1. (Aug 28)	30%
Web Crawl Project		55%
a. Archive-It training	Week 2.	5%
b. Project description and selection / scope update	Week 3. (Sep 11)	10%
c. Metadata update	Week 4. (Sep 18)	10%
d. Quality control	Week 5 (Sep 25)	5%
e. Preservation plan	Week 6. (Oct 2)	10%
f. Project and tool evaluation	Week 8. (Oct 15)	10%
g. Issue Tracking (Sakai Forum)	Ongoing	5%

Class Participation and Work Log (15% of total grade)

Students are expected to complete all required readings. This is important, as a portion of the assignments (or exercises) will be based on readings.

In addition, it is important to work on your small assignments each week in order to complete your term project (Web crawl project) on time, and not to do all the work at the last minute.

> ***Work log***: a *Work Log* board in Forums (Sakai). You should describe what you have done for the week and how much time you put into that work. If you run into any issue, please fill the issue tracking form as well. (Also read others' issue tracking in case your colleagues have the same issue and have resolved it.)
>
> ***Questions*** and ***General Discussion*** boards in Forums (Sakai): Students are encouraged to use these spaces for sharing their thoughts on certain topics, providing reflections on any of readings, and raising any questions about anything relevant to the course or the tool. I'll also raise some questions for you to think and respond in ***General Discussion*** board (may not every week).

Assignments

* All assignments must be turned in through the Sakai class website (**except for your work log and issue tracking**). 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 (Aug 28)

- A brief paper (2-3 pages, font-size 12, double-spaced)
- Instructions will be distributed a week before the deadline.

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

Students will work on term projects throughout the semester.

- **Archive-It Training** (Watch the video in week 2)

The project is to develop a Web archive using Archive-It. All students are required to watch the Archive-It training video run by an Archive-It specialist. This is important, as you need to learn about the tool in order to develop your own Web archive throughout the semester.

- **Small assignments** (weekly exercises)
    - Project description and selection / scope update: Due Week 3 (Sep 11)
    - Metadata update: Due Week 4 (Sep 18)

- Quality control: Due Week 5 (Sep 25)
- Preservation plan: Due week 6 (Oct 2)
- Project and tool evaluation: Due Week 8 (Oct 15)

Students will complete small tasks each week based on topics of the week for their Web archives. The class assignments will lead students to follow a step-by-step process to develop a Web archive, and thus it is important to be on track with the small assignments and not get delayed.

Instructions will be distributed a week before the deadline of each assignment. You should write a one-page report (font size 12) each week about what you have worked on using Archive-It, following the instructions for each week.

- **Issue Tracking**: On going

In addition to the weekly report, students should document issues that they experience during the Web archiving development. It is likely that your colleagues will also have the same issues that you experience, so put your issue log on the Sakai forum, named "issue tracking" so that everyone in the class can see what issues others have and help (by replying to each other's posts) to resolve some of the issues. An issue log form is available on the Sakai forum (see "sample issue tracking"). If you run into any issue, check others' issue tracking. Your issue is not likely something that only you are encountering. Collaborative efforts are always welcome to resolve the issue.

*** 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 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.htm#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.

According to UNC Registrar Policy, undergraduate grades are based on the following definitions:

A	Mastery of course content at the highest level of attainment that can reasonably be expected of students at a given stage of development. The A grade states clearly that the students have shown such outstanding promise in the aspect of the discipline under study that he/she may be strongly encouraged to continue.
B	Strong performance demonstrating a high level of attainment for a student at a given stage of development. The B grade states that the student has shown solid promise in the aspect of the discipline under study.
C	A totally acceptable performance demonstrating an adequate level of attainment for a student at a given stage of development. The C grade states that, while not yet showing unusual promise, the student may continue to study in the discipline with reasonable hope of intellectual development.
D	A marginal performance in the required exercises demonstrating a minimal passing level of attainment. A student has given no evidence of prospective growth in the discipline; an accumulation of D grades should be taken to mean that the student would be well advised not to continue in the academic field.
F	For whatever reason, 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. A grade of F should warrant an advisor's questioning whether the

	student may suitably register for further study in the discipline before remedial work is undertaken.
AB	Absent from final examination, but could have passed if exam taken. This is a temporary grade that converts to an F* after the last day of class for the next regular semester unless the student makes up the exam.
FA	Failed and absent from exam. The FA grade is given when the undergraduate student did not attend the exam, and could not pass the course regardless of performance on the exam. This would be appropriate for a student that never attended the course or has excessive absences in the course, as well as missing the exam.
IN	Work incomplete. This is a temporary grade that converts to F* at the end of eight weeks into the next semester unless the student makes up the incomplete work.
W	Withdrew passing. Entered when a student drops after the six-week drop period.

Course Schedules

Week 1. (Week of Aug 25): Introduction & Basic concepts in Web Archiving; Surface and Deep Web

Review of Web archives assignment DUE on Aug 28

- 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. (Week of Sep 1): Archiving different types of Web materials

Watch video "Archive-It training"

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

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. (Week of Sep 8): Exploration of existing web archives / Selection & Scoping

Project description and selection / scope update Due on Sep 11

- 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

Required readings

- Please review **at least 3** from the following examples and think about their different approaches
  - 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. (Week of Sep 15): Acquisition and Collection Methods / Metadata and Description

Metadata update Due on Sep 18

- 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?

Required readings

- Brown, A. Chapter 4. Collection methods. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006)
- Niu, J. (March/April, 2012). An Overview of Web Archiving. D-Lib Magazine, 18(3/4). Available at http://dlib.org/dlib/march12/niu/03niu1.html (Read the section on *Acquisition*, and *Description and metadata*)
- Bragg, Molly and Lori Donovan. "Archiving Social Networking Sites w/ Archive-It." Available at https://webarchive.jira.com/wiki/pages/viewpage.action?pageId=3113092
- Election 2002 Web Archive Cataloging and Description. MINERVA. Library of Congress. (2004). Available at http://lcweb4.loc.gov/elect2002/catalog/2860.html (Review their use of metadata)

Additional readings

- Masanès, Chapter 4
- Romaniuk, L. (Winter, 2014). Metadata for a Web Archive: PREMIS and XMP as Tools for the Task. *Library Philosophy and Practice, 1098*. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2755&context=libphilprac

- 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
- PANDORA cataloging manual. Available at http://pandora.nla.gov.au/manual/cattoc.html
- Beryl A. Howell, (2006) "How to Use the Internet Archive," Journal of Internet Law (February): 3-9

Week 5. (Week of Sep 22): Quality Control and Post Capture Processing

Quality Control Due on Sep 25

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

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. (Week of Sep 29): Preserving Web Sites

Preservation plan due on Oct 2

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

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.
- Hockx-Yu, H. and Knight, G. What to Preserve?: Significant Properties of Digital Objects. *International Journal of Digital Curation, 3*(1). (2008).

Additional readings

- 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. (Week of Oct 6): 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

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

- Brown, A. Chapter 7. Delivery to users. In *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publishing. (2006)
- 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. (Week of Oct 13): Evaluation of the project

Evaluation of the project and tool Due on 15