University of Michigan School of Information
SI 639: Web Archiving
Winter 2019 (1st Session)

Instructor: Lori Donovan
Office Hours: Thursdays, 10:00-11:00am NQ 1278 (starting 1/17) and by appointment.
Email: lorimd@umich.edu
Meeting Time: Thursday, 6-9 p.m. Jan 10, 2019-Feb 21, 2019
Location: NQ 1255
Credits: 1.5
Prerequisite: SI 506 (or taken concurrently) required

OVERVIEW/INTRODUCTION

The Internet is the primary delivery mechanism for a wide variety of digital content. Archivists, librarians, record managers and preservation administrators need to be familiar with tools and appropriate techniques to capture and preserve information delivered through the "surface" web (static web pages, blogs, social media, etc.) as well as the "deep" web (e.g. databases, streaming media, and authenticated resources). The long-term viability and use of such materials requires an understanding of descriptive standards, formats, data structures and appropriate forms of access to a designated community of users in light of intellectual property and copyright considerations.

This 1.5 credit course will expose students to existing and emerging tools for capturing web content with a heavy emphasis on hands-on experience with the current generation of web crawlers. Students will also learn about collection plan development, current preservation strategies and formats, intellectual property issues, and emerging uses of web archives.

Course Objectives

Upon completing this course students will be able to:

1. Understand the challenges associated with the appraisal, acquisition, storage, and provision of access to websites and online content from various platforms.
2. Capture web content using Archive-It and other readily available open source tools.
3. Develop and employ a collection development policy to create a focused archive of web content.

4. Address important legal and intellectual property issues related to web archiving.
5. Speak knowledgeably about standards and best practices for sustainability of archived web content.

GRADING, ASSIGNMENTS, READINGS, and RESOURCES

Grading & Assignments

Component	Due Date	% of final grade
Class attendance and participation	Weekly	20%
Collection Plan Analysis	01/24/19	25%
Web Collection Project		
a. Collecting Goals and Seed Selection	01/31/19	10%
b. Metadata Conventions	02/07/19	10%
c. Post-Crawl Project Update	02/14/19	5%
d. Full Web Collection Report	02/22/19	30%

- Full details of assignments will be provided in class and will be made available in Canvas.
    - Assignments must show an understanding of core course concepts and should integrate points from readings and lectures (with attribution) where appropriate.
    - Written work should display grammar, punctuation, and syntax consistent with a graduate-level professional program.
- Attendance: present in class or excused absence
- Participation: active engagement in discussions, activities, and labs.

Readings

Readings will be available via links to online resources from the syllabus and/or files located in Canvas. Students are expected to have read the materials before class as we will be referring to these in lecture and in the exercises.

Resources

Weekly lecture slides, additional resources for class assignments and weekly discussion topics will be posted on Canvas.

ORIGINAL WORK

Unless otherwise specified in an assignment, all submitted work must be your own, original work. Any excerpts from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on

Academic and Professional Integrity (stated in the Master's and Doctoral Student Handbooks) will result in severe penalties, which might range from failing an assignment, to failing a course, to being expelled from the program, at the discretion of the instructor and the Associate Dean for Academic Affairs.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

The University Faculty Senate (SACUA) in 2006 endorsed the following language for inclusion on course syllabi: **If you think you need an accommodation for a disability, please let me know at your earliest convenience**. Some aspects of this course (including assignments, in-class activities, and teaching strategies) may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://ssd.umich.edu/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information you provide as private and confidential.

STUDENT MENTAL HEALTH AND WELLBEING

The University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (734) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a listing of other mental health resources available on and off campus, visit: http://umich.edu/~mhealth/

CLASSROOM ETIQUETTE

Students are encouraged to bring laptops to class and to use them actively as learning tools. Students should:

- Use laptops for taking notes, conducting research required for activities, and other specific classroom tasks as assigned by the instructor. During class, students should not check email, 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.
- Demonstrate sensitivity to others by refraining from viewing content that might be distracting or offensive to other members of the class.

SCHEDULE

January 10 (Week 1):

- **Basic concepts in Digital Preservation and Web Archiving**
- **The Web Archiving Life Cycle Model**
  - Vision and Objectives
  - Collection Plans
- **Web-Archiving Toolbox:**
  - Native browser functionality
  - Adobe Acrobat

Readings:

Digital Preservation:

- Lavoie, Brian. "The Open Archival Information System (OAIS) Reference Model: Introductory Guide" (2nd Edition) DPC Technology Watch Report 14-02 (October 2014) http://dx.doi.org/10.7207/TWR14-02

Web Archiving:

- Potter, Abbey. "The Why and What of Web Archives." *The Signal: Digital Preservation* (April 29, 2014) http://blogs.loc.gov/digitalpreservation/2014/04/the-why-and-what-of-web-archives/

- LePore, Jill. "The Cobweb: Can the Internet be archived?" *The New Yorker* (Jan. 26, 2015) http://www.newyorker.com/magazine/2015/01/26/cobweb

- National Digital Stewardship Alliance. "Web Archiving in the United Sates: A 2017 Survey" 2017 https://osf.io/ht6ay/ **(Skim)**

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013) http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
  - **Introduction: pp. 1-5**
  - **Vision and Objectives: pp. 5-8**

January 17 (Week 2):

- **Introduction to Archive-It**

- **Appraisal, Selection, and Scoping**
- **Web-Archiving Toolbox:**
    - ArchiveReady

Readings:

- Pennock, Maureen. "Web Archiving" *Digital Preservation Coalition Technology Watch Report 13-01* (March 2013) http://dx.doi.org/10.7207/twr13-01 (Skim)

- Stanford University Library. "Archivability" (2014) http://library.stanford.edu/projects/web-archiving/archivability (includes sections on Access, Capture, and Description)

- Collecting Policies:
    - Library of Congress (rev. Sept 2017) https://www.loc.gov/acq/devpol/webarchive.pdf
    -

- Archive-It "User Guide" https://support.archive-it.org/hc/en-us/categories/201179946-Archive-It-User-Guide

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013) http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
    - **Appraisal and Selection: pp. 22-23**
    - **Scoping: pp. 23-24**

January 24 (Week 3):

- **Acquisition of Content**
- **Group work: seed selection and test crawls**
- **Web-Archiving Toolbox:**
    - HTTrack

Assignment Due: Collection Plan Analysis (individual assignment)

Guest Speaker: Dallas Pillen, Archivist for Metadata and Digital Projects, Bentley Historical Library

Readings:

- Roche, Xavier. "Copying Websites" *Web archiving*. (Masanès, Julien, editor) Berlin: Springer, 2006. **Canvas**

- Jackson, Andy. "The Provenance of Web Archives." *UK Web Archive blog* (November 20, 2015) http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013) http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
  - **pp. 8-12 (Resources and Workflow)**
  - **pp. 25-25 (Data Capture)**

January 31 (Week 4):

- **Post-Crawl Quality Assurance**
- **Description**
- **Web-Archiving Toolbox:**
  - Wget

Assignment Due: Collecting Goals and Seed Selection (group and individual components)

Readings:

- Hockx-Yu, Helen. "How good is good enough? – Quality Assurance of harvested web resources." *UK Web Archive Blog* (October 30, 2012). http://britishlibrary.typepad.co.uk/webarchive/2012/10/how-good-is-good-enough-quality-assurance-of-harvested-web-resources.html

- Reyes Ayala, Brenda. "Current Quality Assurance Practices in Web Archiving" (2013). http://digital.library.unt.edu/ark:/67531/metadc159526/

- Peterson, Christie. "Archival Description for Web Archives." *Chaos -> Order blog* (June 12, 2015) https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/

- Dooley, Jackie, and Kate Bowers. "Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group." (2018) **(Skim)**

https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html

- Examples of Description Guidelines:
  - Library of Congress, "Web Archiving Cataloging" https://www.loc.gov/webarchiving/cataloging.html
  - Bentley Historical Library, "Contextualization of Content" https://sites.google.com/a/umich.edu/bhl-archival-curation/web-archiving/03-contextualization-of-content
  - NYARC, "Metadata Application Profile for Description of Websites with Archived Versions" https://drive.google.com/file/d/1UNY1mvNLxF_X_XtfUpvXIujkn1FQec6Y/view

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013) http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
  - **pp. 26-27 (Quality Assurance and Analysis)**
  - **pp. 20-21 (Metadata and Description)**

February 7 (Week 5):

- **Intellectual Property Rights/Web Archiving Ethics**
- **Digital Preservation Issues**
- **Web-Archiving Toolbox:**
  - WARCreate
  - webrecorder.io

Assignment Due: Metadata Conventions (Group Assignment)

Readings:

Preservation:

- Taylor, Nicholas. "Anatomy of a Web Archive." *The Signal: Digital Preservation Blog* (5 November 2013) http://blogs.loc.gov/digitalpreservation/2013/11/anatomy-of-a-web-archive/
- Wright, Richard and Miller, Ant. "The Significance of Storage in the "Cost of Risk" of Digital Preservation." *The International Journal of*

Digital Curation (4:3) 2009.
http://www.ijdc.net/index.php/ijdc/article/view/138

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013)
  http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
  - **pp. 17-18 (Preservation)**

Intellectual Property Rights:

- Association of Research Libraries. "Section Eight: Collecting Material Posted on the World Wide Web and Making It Available," *Code of Best Practices in Fair Use for Academic and Research Libraries*. (Jan 2012)
  http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf

- "The Section 108 Study Group Report" (March 2008) **pp. 80-87**
  http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

- Jules, Burgis, Summers, Ed, Mitchell, Jr., Dr. Vernon. "Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations." (April 2018)
  https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013)
  http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
  - **pp. 18-20 (Risk Management)**

February 14 (Week 6):

- **Access and Use**
- **Archiving Deep Web Content and APIs**
- **Web-Archiving Toolbox:**
  - youtube-dl
  - twarc

Assignment Due: Post Crawl Project Update (Group Assignment)

Readings:

- Milligan, Ian. "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives." *International Journal of Humanities and Arts Computing*, Volume 10 Issue 1 (March 2016) https://www.euppublishing.com/doi/abs/10.3366/ijhac.2016.0161

- Reynolds, Emily. "Web Archiving Use Cases." (2013) http://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf

- Neubert, Michael. "Users, Use Cases and Adapting Web Archiving to Achieve Better Results." *The Signal Digital Preservation Blog* (May 8, 2015). http://blogs.loc.gov/digitalpreservation/2015/05/users-use-cases-and-adapting-web-archiving-to-achieve-better-results/

- Bragg, Molly, Hanna, Kristine, et al. "The Web Archiving Life Cycle Model." (March 2013) http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf
    - **pp. 12-16 (Access/Use/Reuse)**

February 21 (Week 7):

- **Web Archiving Case Studies**
- **Collaboration in Web Archiving**

Guest Speaker: Sumitra Duncan, Head, Web Archiving Program, The Frick Art Reference Library

Readings:

- Davis, Corey. "Archiving the Web: A Case Study from the University of Victoria." Code4Lib Issue 26, 2014-10-21 (2014). http://journal.code4lib.org/articles/10015
- Colburn, Alston. "Case Study: Washington and Lee's First Year Using Archive-It." Journal of Western Archives Volume 8 Issue (2017) https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1077&context=westernarchives
- Hight, Cliff, et. al. "Collaboration Made It Happen! The Kansas Archive-It Consortium". Journal of Western Archives Volume 8 Issue (2017) https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1078&context=westernarchives
- Gafney, Dylan. "Make It Weird": Building a Collaborative Public Library Web Archive in an Arts and Counterculture Community. (October 2018). https://archive-it.org/blog/post/make-it-weird/

- Library of Congress Collaborations: https://www.loc.gov/programs/web-archiving/about-this-program/web-archiving-collaborations/

Final Project due via Canvas by 5:00 p.m. on Friday, Feb. 22