**University of Michigan School of Information**  
**SI 639: Web Archiving**  
**Winter 2021 (1st 7 week Session)**

**Instructor:** Lori Donovan  
**Meeting Time:** Thursday, 5:30-8:30 p.m. Jan 21, 2021-Mar 4, 2021  
**Location:** [https://umich.zoom.us/j/93805863686](https://umich.zoom.us/j/93805863686) passcode: 936460  
**Office Hours:** Thursdays, 8:30-9:00pm (directly after class) and by appointment.  
**Email:** lorimd@umich.edu  
**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. Web archives are also a source of valuable data for analysis by historians, political scientists, journalists and more, so familiarity with access methodologies for both micro and macro analysis is key.

This 1.5 credit course will expose students to existing and emerging tools for capturing, replaying and analyzing 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**

<table>
<thead>
<tr>
<th>Component</th>
<th>Due Date</th>
<th>% of final grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekly participation</td>
<td>Weekly</td>
<td>15%</td>
</tr>
<tr>
<td>Collection Plan Analysis</td>
<td>02/04/21</td>
<td>20%</td>
</tr>
<tr>
<td>Web Collection Project</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Collecting Goals and Seed Selection</td>
<td>02/11/21</td>
<td>20%</td>
</tr>
<tr>
<td>b. Metadata Conventions</td>
<td>02/18/21</td>
<td>15%</td>
</tr>
<tr>
<td>c. Post-Crawl Project Update</td>
<td>02/25/21</td>
<td>5%</td>
</tr>
<tr>
<td>d. Full Web Collection Report</td>
<td>03/05/21</td>
<td>25%</td>
</tr>
</tbody>
</table>

- Full details of assignments will be 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.
- Participation: active engagement in discussions, activities, and labs. This can include chat or verbal discussion/sharing for those attending synchronously or questions, lab results or other feedback submitted via form, for those accessing course materials asynchronously.

**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, course recordings and 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/](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/](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](https://www.uhs.umich.edu/mentalhealthsvcs), or for alcohol or drug concerns, see [www.uhs.umich.edu/aodresources](http://www.uhs.umich.edu/aodresources).

For a listing of other mental health resources available on and off campus, visit: [https://caps.umich.edu/article/um-mental-health-resources](https://caps.umich.edu/article/um-mental-health-resources)
VIRTUAL CLASSROOM ETIQUETTE

Students are encouraged to attend lecture live when possible, both to enable interaction and discussion on the lecture topics, but also to facilitate breakout activities and group work. 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.
- Enable their cameras during discussions/team work if possible, noting that not everyone has this capability.
- Raise their hands in Zoom during lectures or to chat questions to the whole group or to the instructor privately if they’d rather not ask over audio. During discussion times folks can unmute and discuss in realtime. Breakout groups will be utilized to facilitate group projects and to allow for smaller discussions.
- 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 content.
- Be kind to one another and yourselves!

Optional, but encouraged – sharing your pronouns in your Zoom name (if they’re not there by default you can add them by renaming yourself). This helps us all know how to address one another.
SCHEDULE

January 21 (Week 1):

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

Readings:

**Digital Preservation:**
  [http://dx.doi.org/10.7207/TWR14-02](http://dx.doi.org/10.7207/TWR14-02)

**Web Archiving:**
  [http://www.newyorker.com/magazine/2015/01/26/cobweb](http://www.newyorker.com/magazine/2015/01/26/cobweb)
  - Introduction: pp. 1-5
  - Vision and Objectives: pp. 5-8

January 28 (Week 2):
- Introduction to Archive-It
- Appraisal, Selection, and Scoping
- Web-Archiving Toolbox:
  - ArchiveReady

Readings:


  (includes sections on Access, Capture, and Description)

- Collecting Policies (more in the collection plan analysis assignment):


  - Appraisal and Selection: pp. 22-23
  - Scoping: pp. 23-24

February 4 (Week 3):

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

Assignment Due: Collection Plan Analysis (individual assignment)

Readings:

http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html


- pp. 8-12 (Resources and Workflow)
- pp. 25-25 (Data Capture)

**February 11 (Week 4):**

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

**Assignments Due:** Collecting Goals (group) and Seed Selection (individual)

**Readings:**

http://britishlibrary.typepad.co.uk/webarchive/2012/10/how-good-is-good-enough-quality-assurance-of-harvested-web-resources.html

https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/1145/1004

- 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/

https://elischolar.library.yale.edu/jcas/vol6/iss1/31/
Examples of Description Guidelines:
- 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” (Version 2, Aug 2018)
  https://drive.google.com/file/d/1UNY1mvNLxF_XXtfUpvXIujkn1FQec6Y/view
  - pp. 26-27 (Quality Assurance and Analysis)
  - pp. 20-21 (Metadata and Description)

February 18 (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:
  http://blogs.loc.gov/digitalpreservation/2013/11/anatomy-of-a-web-

  o pp. 17-18 (Preservation)

Intellectual Property Rights:


    o pp. 18-20 (Risk Management)

February 25 (Week 6):
  ● Access and Use
  ● Archiving Deep Web Content and APIs
  ● Web-Archiving Toolbox:
    o youtube-dl
    o twarc
Assignment Due: Post Crawl Project Update (Group Assignment)

Readings:

  - pp. 12-16 (Access/Use/Reuse)

March 4 (Week 7):
- Web Archiving Case Studies
- Collaboration in Web Archiving
- Web Archiving Integrations
- Web Archiving Toolbox
  - Archives Unleashed Cloud

Readings:


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, March 5