National Library of Medicine

NDSR Project Developing a Thematic Web Archive Collection

Goal Summary

The Resident will create a collection of Web content on a specific theme or topic of interest to the Resident and relevant to NLM collecting strategies, such as medicine and art or the e-patient movement. The project will involve developing a new Web collection from start to finish, and will include testing and refining anticipated workflows and timeframes for the development of a new Web collection. Documentation by the Resident of the processes of creating a new thematic collection will aid in the development of effective workflow and quality control review procedures to optimize the quality of Web collecting at the National Library of Medicine. The work of this project may serve as a case study for the development of a new Web collection and will likely be of great interest to other institutions considering Web collecting initiatives within their own organizations.

Specific Goals / Objectives

For this project, the Resident will:

- Identify a topic/theme for Web collecting within the scope of NLM Collection Development strategies (http://www.nlm.nih.gov/tsd/acquisitions/cdm/).
- Develop a collecting proposal, including research and identification of 30-50 seeds to crawl, with detailed recommendations for the crawling frequency and duration necessary to fully capture the desired resources.
- Add seeds to NLM's Archive-It collection (http://www.archive-it.org/organizations/350) and initiate a test Web crawl to identify and address Web capture problems in advance of an actual crawl.
- Conduct a Web Crawl and review captured content, analyzing and documenting results to determine whether desired content was adequately captured.
- Make adjustments to crawling instructions to improve future capture of content; monitor/review scheduled crawls.
- Review and make recommendations for the enhancement of preliminary Web collecting workflows as described in the April 2012 report of the NLM Web Collecting and Archiving Working Group.

Timeframe & Deliverables

Months 1-4

- Meet with members of the NLM Collecting and Archiving Working Group for an in-depth introduction to the activities of the working group and NLM Web collecting efforts so far;
- Gain a practical understanding of using Archive-It through training from NLM staff with Archive-It experience, Archive-It's online training modules, and an in depth review of Archive-It's detailed documentation;
- Collaborate with NLM staff in the History of Medicine Division and Technical Services Division to identify a topic/theme for Web collecting within the scope of NLM Collection Development strategies; and
- Develop and present to the Web Collecting and Archiving Working Group a Web collecting proposal, including research and identification of 30-50 seeds to crawl. The proposal will be reviewed by the working group and other relevant NLM staff for content and for technical and data budget considerations. Once approved, and if needed, the Resident will seek copyright permissions with guidance from NLM.

Months 5-6

- Add seeds to NLM's Archive-It collection and initiate a test Web crawl to identify and address Web capture problems in advance of an actual crawl;
- Conduct a Web Crawl and review captured content, analyzing and documenting results

based on a review of Archive-It Crawl reports and a manual review of content to determine whether desired content was adequately captured; and
- Make further adjustments to crawling instructions, if needed, to improve future capture of content; monitor/review scheduled crawls as needed according to collecting proposal.

Months 7-8
- Prepare a detailed report of the selection strategy, the procedures followed, specific challenges encountered, and adjustments made throughout the collection development process;
- Make recommendations to the Web Collecting and Archiving Working Group for further development of preliminary workflows and procedures for Web collecting; and
- With NLM mentor, develop a strategy and preparations for communicating results/findings of the NDSR project to NLM leadership and staff, as well as to the wider digital preservation community through press releases, blog postings or other publications, and presentations at meetings or conferences.

Project Deliverables
- Web collecting proposal, including research and identification of 30-50 seeds to crawl
- A report documenting the collection development process and detailed analysis of the results
- Recommendations for further development of preliminary workflows and procedures for Web collecting at NLM
- A new NLM Web collection of 30-50 Web resources

Resources Required

1 Mentor (Moffatt), 1 Resident

Access to select staff within the National Library of Medicine, including the leadership of NLM's History of Medicine Division and Technical Services Division, and the Web Collecting and Archiving Working Group.

As needed, contacts with other related organizations who have demonstrated interest and expertise in web collecting and archiving.

Context

In 2011-2012 the NLM Web Collecting and Archiving Working Group engaged in a pilot project to better understand the processes and challenges of collecting born-digital Web content to expand the Library's collecting strategy for digital formats. An NLM news announcement about this initiative is available at http://www.nlm.nih.gov/news/nlm_web_content_collection.html. One of the recommendations of the Working Group was that NLM should develop curated collections around a particular theme or topic that could be planned well in advance and added to over time.

Required Knowledge and Skills for Residents

Graduate degree in Library and Information Science, or equivalent.
Additionally, the successful candidate will have the following:
- General knowledge of Web archiving and collecting practices, digital preservation principles
- Experience using Microsoft Windows computers and office productivity software such as Microsoft Office

Preferred Knowledge or Experience

Familiarity with Web archiving programs/services, such as Archive-It