**Changing world of scholarly communication: new roles and challenges for libraries**

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

Michael Neubert
Digital Projects Specialist
mneu@loc.gov

1



- What
- Why
- How
- More

2



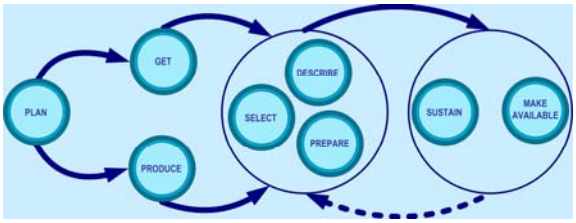
What

- Scholarly communications
- Web archiving (capture, harvesting)
  - More archival
  - Covers entire *Digital Life Cycle*

3



Digital Life Cycle Stages




Dimensions: Work activities, policies & best practices, and enabling technologizies

4



Why

- Collection building / selection plan
- Cavaet

5



How

- Software
  - Capture software
  - Playback
  - Curator tool
  - Discovery system

6

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

How

- Contractors
  - Archive-It (Archive-it.org)
  - CDL Web Archiving Service (was.cdlib.org)
  - Commercial companies

7

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

How

- Make no assumptions
- Permissions
- Technical limitations

8

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

More

9

**Changing world of scholarly communication: new roles and challenges for libraries**

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

Michael Neubert
Digital Projects Specialist
mneu@loc.gov

10

*Critical questions to answer for archiving scholarly blogs, online forums, and other digital communications*

SLIDE 1: We are going to look at web archiving as applied to scholarly blogs, online forums, and other communications. If you have any further questions later, feel free to contact me.

SLIDE 2: I would like to provide you some answers to the "what, why, how" questions that you may have about this topic along with providing some pointers to further information.

SLIDE 3: What is web archiving, and what kind of scholarly communications are possible to archive using this tool?

Scholarly communications is often understood broadly to include any exchange of information among scholars themselves and between scholars and the audience for their research, up to and including formal publication. For web archiving the focus is on openly accessible website-based communications or publications such as blogs, online discussion forums, or social media.

Web archiving – the terminology used varies. My colleagues and I try to use the phrase "web archiving" but one may also hear "web capture" or "web harvesting" – these are all the same thing.

For librarians, the phrase "web archiving" reminds us that many aspects of this activity are more familiar to our archivist colleagues than to librarians who focus on formally published materials such as books and serials. The ephemeral nature of much of what is contained in websites and the frequent lack of clarity about our right to copy and reuse legally what is on the Internet – do we need permission? – may be more familiar to archivists.

Web Archiving is a library activity or function with its own complete Digital Life Cycle.

SLIDE 4: A Digital Life Cycle Model such as this is one used at the Library of Congress – there are others – is helpful as a framework for thinking through how to organize a web archiving project. The life cycle stages run from Plan, Produce, and Get through Sustain and Make Available. Each stage has three dimensions – work activities (or work flow),

policies and best practices, and enabling technologies. You can perform a simple gap analysis with a table with the stages in columns and the dimensions in rows. For example, is there a policy for sustaining or preserving the archived web sites?

SLIDE 5: It is best to have a clear statement describing why you are doing a particular web archiving project that is consistent with the institutional goals for collection development. There should be a written statement explaining the selection parameters for any web archiving collection that anyone can read and easily understand. Web archiving can be both about the content as information and also about capturing for future users how the particular information was presented on websites. So for a blog it can be both the blog entries by the blog author or authors and also how a multi-part discussion takes place in the "comments" section of the blog.

SLIDE 6: Web archiving involves as many as four different software activities, although only two are essential. Much of the software used by those doing this activity is open source programs, but there are private companies using proprietary approaches as well.

Capture software – first you need a program that will "capture" (or harvest) the content from websites. The most widely used program is called Heritrix. If we give the program instructions to crawl all of XYZ.org, it starts at the home page and pulls in all the files required to display XYZ.org and then follows links it sees within XYZ.org to do the same with all pages of that site. We might give it instructions to also capture pages "one hop," or one click, out from XYZ.org – in this regard the capture software does exactly what it is told. It cannot capture underlying software that requires user interaction, so things like the search function for a site are not captured. It saves all these bits and pieces of the site in a format for web archiving known as WARC.

Playback software – in order to view the WARC files, you need separate playback software. The most commonly used is called "the Wayback Machine." The software development efforts for playback and capture are separate tracks so sometimes new releases of the Wayback Machine will be able to display content harvested by the capture software not previously displayable in the archived version.

The capture and playback software, generally operated by contractors, are the essential ones.

A more robust web archiving program will want to have a separate curator tool – software to manage archiving a larger number of URLs over time. Web sites can be harvested just once but more typically are harvested at some frequency, from daily to weekly or monthly or just once a year. A curator program is a dedicated tool to keep track of these things and other instructions for harvesting each site that exports the instructions for the capture software and can interact with the playback software.

In addition there can be a separate program for supporting more advanced discovery of archived web sites - these are mostly in use by large web archiving programs, which isn't typical at present for U.S. universities – likely this will change.

SLIDE 7: In the U.S., universities and other organizations are typically using contractors to perform the capture and playback for them. A university typically will receive copies of the WARC files for preservation purposes and later it would be able to display them separately from contractor-supplied playback.

I am not here to endorse any particular contractor organization. I will mention the two non-profit organizations that I am know work with the university web archiving "market" at present. The Archive-It service, which is part of the Internet Archive – the contact is Lori@archive.org (Lori Donovan). And there is the California Digital Library Web Archiving Service which does provide services outside of the UC system – the contact is Rosalie.Lack@ucop.edu. There are now private companies providing web archiving, simply Google "web archiving service" and you will find companies such as PageFreezer and Hanzo Archives and others. Private companies doing web archiving typically archive sites belonging to the customer and not those of other organizations which affects their technical approach.

SLIDE 8: A few more comments.

"Make no assumptions" – an example: since people have heard that the Library of Congress is receiving and archiving "all the Tweets from Twitter" they assume there is no need to harvest Twitter pages as part of the social media presence of an organization. This is an incorrect assumption since LC is acquiring the Tweets as "big data" and this data does not include the ability to reconstruct the original appearance and linked-to

content of the web pages in Twitter. If you want the Tweets as they appeared as web pages, you should harvest the relevant Twitter pages.

"Permissions" – Both of the two main contractor organizations say that making copies of publicly available websites for research is OK under fair use if they don't make copies of the portions of the site that are closed to search engines through instructions embedded in the site's homepage. You can get more details when you consult with these organizations, but if you are going to do web archiving and consider the resulting archived content an acquisition for your collections you would presumably want to talk to your institution's lawyers. Here it may be relevant that in area studies you would be planning to archive sites from other countries and you would want to do some analysis of the relevant laws in those countries also.

"Technical limitations" – Even people immersed in this activity are not always able to correctly determine in advance if a site is "archivable" without doing a test crawl, and for many projects that is the only way to determine if a particular site is not possible to archive successfully with today's technology. For example, sometimes discussion forums with long threaded discussions of different topics will be archived acceptably and others won't – the way to know is to test and assess the results against the collecting goals.

SLIDE 9: On the last slide I have a Bit.ly link that will take you to a copy of these slides and the text of my remarks as well as some additional links for reading about web archiving generally.

As a final comment – web archiving is a refreshingly forward looking activity. By comparison books and serials were published in the past once we receive them, while developing a web archiving effort is looking forward in time to acquiring materials that are mostly not yet produced or available that you anticipate appearing and having value for researchers. It is an interesting alternative to the usual.

SLIDE 10: Thank you.

Resources for further reading

http://libguides.northwestern.edu/content.php?pid=262419 Scholarly Communication in more detail with a discussion of self-archiving, not treated in my presentation.

http://dlib.org/dlib/march12/niu/03niu1.html *An Overview of Web Archiving* by Jinfang Niu; D-Lib Magazine, March/April 2012. *Well organized introduction with useful links.*

http://www.infotoday.com/cilmag/dec11/Grotke.shtml *Web Archiving at the Library of Congress* by Abbie Grotke, Information Today, December 2011. *If one is interested in the Library of Congress efforts.*

http://www.netpreserve.org/ International Internet Preservation Consortium. *An international membership organization that supports web archiving – more university libraries are joining, although initially membership was mostly national libraries.*

https://webarch.cul.columbia.edu/ Web Archiving Policies and Practices in the U.S. A legacy website for a conference organized at Columbia University in 2012 that sought to broaden exchange of information and best practices about web archiving, particularly among university libraries.

Non-profit organizations working with universities to perform web archiving on a fee-for-service basis

https://archive-it.org/ - Archive-It, part of the Internet Archive. Contact lori@archive.org

http://was.cdlib.org/ - Web Archiving Service of the California Digital Library. Contact Rosalie.Lack@ucop.edu