University of California Next Generation Technical Services Initiative
Team 4: 21st Century Emerging Resources
Phase I Report
January 6, 2010

Executive Summary:

- **Expand metadata expertise beyond MARC format.** Many librarians and library staff members have expertise in a wide range of metadata formats already, but there is a strong need for this expertise to be more widely shared. Expertise in non-MARC metadata will be needed for many of this group's recommendations, and will require collaborations between metadata experts, subject experts, and researchers.

- **Collaborative collection development and description of harvested websites.** The Web Archiving Service, developed and hosted at CDL, provides robust tools for harvesting and archiving web content. This service is used across the UC system to collect web content by genre (state documents), events (2007 wildfires in Southern California), and organizations (labor groups). Further collaborative projects should be encouraged.

- **Pilot projects for harvesting blogs and online newspapers.** These materials present special problems for web harvesting tools, such as the Web Archiving Service. We propose two separate pilot projects, one focusing on blogs and the other on online newspapers.

- **Research data curation.** Libraries have not traditionally collected research data. A number of trends point to a need for leadership, and libraries should play a central role on campus in the appraisal, collection, description, access to and long-term preservation of research data. UCSD is implementing DataVerse, and we propose a pilot project to investigate the use of DataVerse across UC.

- **Improve access to, and ensure preservation of, ebooks**. Libraries have a large investment in commercial ebooks. Are there gaps in the description of and access to ebooks? A systematic study is needed to determine if there are any problems. In addition, while many large commercial publishers cooperate with Portico for long-term preservation, not all do. Are there significant ebook publishers not covered by preservation plans, and if so, how should UC libraries respond? We propose to study this issue, leading to recommendations for systematic treatment.

- **Improve access to ETDs.** CDL is working with several campuses on a pilot project to collect, provide access to, and preserve ETDs and any associated auxiliary files. One area that requires attention is how these materials are described. ProQuest provides MARC records, but this can be a very slow process. They provide some rudimentary descriptive metadata in Excel much more quickly. We propose developing tools to convert that information into minimal-level MARC records.

- **Preserve content on CD-ROMs.** This is in many ways not about emerging resources, but disappearing ones. Libraries have received books with accompanying CDs and DVDs for 15 years or longer. Following on pilot projects at UC Berkeley and Indiana University, analyze problem across UC system and develop a system-wide solution.

Recommendations and Analysis:

NGTS-4 is charged with analyzing the selection, acquisition, cataloging and preservation of 21st Century Emerging Resources across UC libraries, looking at ways to improve efficiencies and services in current practice, and determine how changes in policies and technology may affect the future. The working group consists of members from 6 UC campuses and CDL, with wide experience with digital resources. We have met via conference call several times over the past few months, and inventoried digital projects across UC.

In particular, we have focused on several categories of materials, detailed below. We see opportunities for leadership within the UC system in developing and promoting best practices in these areas. Some categories are materials already collected by libraries; other categories include materials traditionally not collected by libraries, but for which there is growing interest and need.

General Recommendation: We feel that there is a strong need to expand the variety of metadata standards used in the UC Libraries, and several of the recommendations below will require familiarity with new metadata formats. We can achieve this by collaborating with communities that are using emerging technologies and to influence the descriptive and other metadata standards that may already be in use. As an organization, we need to become adept at accepting a wide range of formats and transforming them to other formats as needed. As a long-term goal, it will require:

- training of existing staff
- working closely with researchers and specialists in subject domains
- a willingness to rethink position descriptions as vacant positions are filled

In addition, we analyzed these specific formats and publication types:

1) Harvested websites, including scholarly resources

CDL's Web Archiving Service (WAS) provides a rich set of tools for selecting, acquiring and archiving web content. WAS includes features such as the preservation of web content, along with the creation of a persistent URL for any given website captured by WAS. WAS was first developed as a tool for librarians, archivists and curators to collect websites, beginning in October 2008. In July 2009, WAS began to provide access to these websites to the general public, and can now serve as both a preservation archive and access tool for web content.

While these tools are proving useful, their scope extends only to capturing websites in the public webspace. Licensed and otherwise restricted materials are excluded from WAS. Services such as Portico and LOCKSS are devoted to archiving ejournals and are beginning to take on ebooks (see #4 below).

For the purposes of work of this task force, it is important to remember that the problem of archiving web content is shared both inside and outside the UC system, and solutions (or suggested solutions) will come from a wider community among libraries nationally or even internationally.

Using California state government documents as an example, a pilot study was conducted by CAMCIG (Cataloging and Metadata Common Interest Group), developing workflows to identify and describe them. In addition, following recommendations from GILS (Government Information Librarians), they used WAS to harvest and archive these documents, providing permanent access and persistent URLs. This model, although at an early stage, has been successful in at least increasing the amount of cataloging generated.

http://libraries.universityofcalifornia.edu/hots/camcig/CalDocsFinalReport-2A.pdf

Besides government documents, librarians from multiple campuses have formed other WAS collections, some based on events, such as the H1N1 flu outbreak or California wildfires, and some based on multi-campus research units such as the California Institutes of Regenerative Medicine.

Recommendation: Propose additional cooperative projects, including, for example, documents from municipal, county and regional government agencies, or GIS data. Develop a collection development policy for collaborative projects that provide systematic coverage of areas such as local governments.

2) Blogs, online newspapers, and other dynamic web content

This category is a subset of web content discussed in the previous point, but it presents special complexities. Weblogs are an established means of scholarly communication, but most have no regular publication schedule. Readers' comments create an opportunity for debate and clarification that add to the value of the original post. Online newspapers have their own complexities. Some online newspapers present an ever-changing face with no static "final edition," while others do have a more routine publication schedule with stable editions and finalized content. Newspapers also include opportunities for readers' comments. Many publishers of web content do not have a preservation or archiving strategy. They may welcome the opportunity to work collaboratively to develop a strategy and a workflow to achieve the preservation of their content.

Unfortunately, these aspects of blogs, newspapers and other dynamic websites provide difficulties for current tools used in web harvesting. Ideally, blogs should be harvested only when new content is added, rather than following a regular schedule for crawling. In addition to the blog posts themselves, the comments can appear a considerably long time after the original post, and should be captured as they are added. Web harvesting tools can capture blog posts and comments, but lack the functionality for the kind of conditional capture that would match the characteristics of the medium. NARA does not provide much guidance in this area. While recommending that blogs from federal agencies be captured and archived, NARA does not provide guidance on how best to do it. There are research projects such as ArchivePress (working with WordPress blogging software) that may provide tools in the future.

Recommendations: For blogs, we propose a pilot project with WordPress in the UC system. For newspapers, we propose a pilot with a web publisher to collaborate on work flow as one way of capturing their content, while also investigating the use of WAS as a tool for capturing online newspapers. The Center for Bibliographical Studies and Research at UCR has, as part of the California Digital Newspaper Collection <http://cdnc.ucr.edu/newsucr> and California Newspaper Project <http://cnp.ucr.edu/>, investigated collecting born-digital newspapers, and we should collaborate with them in this investigation.

3) Research data

Faculty members use and generate data in the course of their research. There is a growing list of requirements by granting agencies such as NIH and NSF for researchers to preserve the data generated during the course of funded research projects, yet there are few systematic efforts to do so. The DataVerse Network Project, under development at Harvard, allows users to search across or browse through the virtual data archives that are hosted at Harvard or at an individual institution. In its research, NGTS-4 noted at least 18 dataverses (as the virtual data archives are known) from a variety of UC campuses in the Harvard-hosted archives. The UCSD Libraries plan to host their own DataVerse Network, to be made public in the near future.

Two examples show the immediacy of the issue:

- The Neuroscience Information Framework, an NIH-funded grant project centered at UCSD, is, as explained on their website:

  > a dynamic inventory of Web-based neuroscience resources: data, materials, and tools accessible via any computer connected to the Internet. An initiative of the NIH Blueprint for Neuroscience Research, NIF advances neuroscience research by enabling discovery and access to public research data and tools worldwide through an open source, networked environment.

  There are no librarians listed among the project staff.

- Another NIH-funded project, the eagle-i project, plans to

  > build a national research resource discovery network that will allow biomedical scientists to quickly find previously invisible but potentially valuable research resources (e.g., technologies, animal models, equipment, cell and tissue banks, training opportunities).

  One of the partners, the Oregon Health and Science University Library "will lead the Data Curation Team to build the ontologies and vocabularies used to describe research resources and make them easier to find."

  > "This award will fund eight new researchers working in Oregon on the eagle-i project," said Chris Shaffer, University Librarian. [...] "Libraries have a long history of organizing research information and publications.

> Through this project, we are extending library expertise into the research enterprise in new and exciting ways."

More details on the eagle-I project can be found here:
<http://www.ohsu.edu/xd/education/library/about/whats-new/ohsu-library-awarded-funds.cfm>

Recommendation: This area of research data is one where NGTS-4 feels that the UC libraries should provide leadership within UC, providing a service to acquire, describe and preserve datasets. Working with the UC VP for Research and Research Officers and Principal Investigators on each campus, librarians could identify important research datasets for acquisition, description and preservation. UCI has used the NIH mandate as an opening for forging a close relationship with the Office of Research, and is included as grant-funded projects are proposed.

In particular, we propose investigating the use of the UCSD DataVerse Network instance as a pilot for the UC system (both access and preservation). This pilot should address questions about how well the system would scale up to serve as a UC-wide resource, and the role of CDL in long-term preservation of the data.

4) EBooks

EBooks have the most systematic treatment of the e-resources under analysis by NGTS-4, but there remain questions about access and preservation. UC and CDL have several models for acquiring ebooks. The main stream comes from commercial publishers such as Springer, Safari, and others. CDL licenses or purchases individual titles or entire categories of the publishers' lists, and as such is included in the charge for NGTS-1. The Shared Cataloging Program catalogs them and makes the records available to individual campuses, and the records are then loaded into Melvyl. CDL maintains OpenURL links to the items through UC e-links (SFX).

In addition, UC is a partner in both the Google Books Project and the Open Content Alliance, which are creating millions of digital versions of books from its collections. UC and CDL are members of the HathiTrust, founded at the University of Michigan, a preservation archive for mass digitization projects such as Google and the Open Content Alliance.

While Portico and LOCKSS include ebooks in their collections, the coverage is not systematic. We need to determine where gaps exist in long-term preservation of these materials, and develop a strategy to provide preservation services.

Recommendation: In collaboration with NGTS-1, study the entire range of ebooks within the UC libraries to determine whether description, access and preservation are adequate.

5) Electronic Theses and Dissertations (ETD)

These materials are of interest to all campuses because they are manifestations of scholarly communications unique to UC, and represent the culmination of graduate education on the campuses. Each campus sets requirements for formatting and media on its own, and their guidelines vary widely across UC. Some campuses do not accept electronic submission of theses and dissertations; some campuses accept either print or electronic submission; at least 3 campuses (UCD, UCI and UCSF) require theses and dissertations to be submitted in a single PDF file. Dissertations are submitted to ProQuest, which digitizes the print copy or accepts a PDF version. As of now, ProQuest does not accept any supporting or auxiliary materials, such as research data.

Within CDL, both the Digital Preservation Program and eScholarship have been working to understand the issues surrounding ETDs, and determine what is needed to begin to include ETDs within both the eScholarship publication platform and the Digital Preservation Repository. The interactions between graduate schools and libraries, and agreements with ProQuest, complicate the situation. CDL is currently working with UCSF to begin to ingest ETDs into the DPR.

Recommendation: In collaboration with NGTS-3, document the work flow of theses and dissertations on each campus. CAMCIG has documented the cataloging practices on each campus:
http://libraries.universityofcalifornia.edu/hots/camcig/CampusPracticesForThesesAndDissertations.pdf

According to a study at UCI, ProQuest provides rudimentary descriptive information and links to ETDs up to 6 months before MARC records are available. We recommend investigating whether this information could be used to create minimal-level records, and whether this process could be automated.

6) CD content

Libraries have yet to address in any systematic way long term access to and preservation of CDs. UC libraries hold a significant number of titles on CD even though CD has a limited shelf life. For several years, the Library Data Lab at UC Berkeley, has been copying the content of CDs onto spinning disk using the ISO file format. The advantages of this format are as follows:

- facilitates data management by bundling the entire contents of a CD into a single file,
- allows an accurate reproduction of the original disk to be created with standard CD ripping software
- facilitates network access to both the data and software contained on the CD.

A project to address long term preservation of the data has recently begun. A related project, the SUDOC Virtualization Project, is ongoing at the Indiana University.

Recommendation: Using Berkeley and Indiana's ongoing projects as possible models, develop a set of recommended procedures, software, and format standards to promote preservation of and access to data that currently reside on CD. Investigate technological approaches to identifying duplicate holdings across collections. Identify potential legacy file format and software dependency issues that may require special attention to assure long term access. Determine whether there are significant digital collections on other obsolete media require attention.

Members of NGTS-4

- Perry Willett (CDL), Chair
- Carol Hughes (UCI), NGTS Steering Committee liaison
- Harrison Dekker (UCB)
- Brad Eden (UCSB)
- Ken Furuta (UCR)
- Ardys Kozbial (UCSD)
- Sue Perry (UCSC)
- Lisa Sibert (UCI)
- John Ridener (UCB)
- Roger Smith (UCSD)