University of California Archivists Council

Meeting of June 3, 2008
At UC Berkeley
Earth Sciences and Map Library
50 McCone Hall
9:30-3:30

Agenda

1. Roll Call and Introductions

2. Housekeeping

   - Approval of Minutes of April 12, 2007
   - Approval of Annual Report, 2006-2007

3. Drafting of documents

4. Reports

5. UCAC faculty papers presentation at 2008 SAA

6. Academic Senate - Guidelines for collecting systemwide and divisional records

7. Student Newspaper Project proposal

8. UC Terminology

9. LAUC Retention Schedule

10. Scanning Services

11. Archives Authority

12. Campus eRecords Projects

13. Electronic Theses and Dissertations - Update

14. Archivists' Toolkit

15. Web Archiving Project

16. Plans for next meeting/conference call

Minutes

Attendance: Farrell (B), Neal (B), Skarstad (D), Jackie Dooley (I, recorder), Brown (LA), Jim Dooley (M), Wilson (R), Gartrell (SB), Bunting (SC), Coy (SD, chair), Day (SD Scripps), Mix (SF), Williams (UCOP)

Guests: Chuck Eckman (B), Tracy Seneca (CDL)

1. Introductions

Chuck Eckman, Head of Collections at Berkeley, our liaison to the Joint Steering Committee (collection development), came briefly to meet the group.

2. Housekeeping

Coy will post the April 2007 minutes with minor corrections made at the meeting. No fall meeting was held. The 2006/2007 UCAC annual report was approved.

3. Drafting of documents

Coy will draft the 2007/2008 annual report and send to us for review. Due in September.

Review of existing goals to retain for this year: web archiving service, forwarding article ideas to UC Chronicle, monitor campus ETD programs, development of policy statement on Senate records, monitor HOSC, determine archivists authority to give copyright permission to publish (Brown in the lead on this), monitor campuses' development of institutional repositories.

Only item we're dropping is development of content for UC digital history archive.

4. Reports

A. HOSC conference call on 2 June (Jackie Dooley)

The meeting included a conversation with Rosalie Lack, CDL Head of Digital Special Collections. Rosalie may be setting up a steering committee. UCAC would like to be part of the discussions of digital special collections steering committee and collection development policy for digital special collections. Coy will inform Rosalie that we recommend including a University Archivist on the steering committee. Question: Does CDL accept all file formats for digital objects, or only certain ones? HOSC discussion yesterday was focused on the broad picture, not on what a committee would focus on. We assume policy issues beyond collection development policy.

Next-Gen Melvyl discussion of the fact that notes in archival records are generally not being displayed. Jim Dooley stated that NO note records are being displayed for any records. HOSC will draft a letter to send to ULs and the implementation committee (huzzah to David de Lorenzo at Bancroft for drafting it). Coy will contact HOSC chair Melissa Conway (R) to request that UCAC see the draft and be able to endorse the letter when it goes forward.

Brown asks that HOSC and UCAC minutes be shared with both groups. Very useful for mutual awareness of issues.

B. PAG (Jim Dooley, Mix, Brown)

PAG met with Patricia Cruse re digital preservation and Emily Stambaugh re shared print. PAG is concerned with ensuring that the DPR meet national standards. Steve Abrams has been hired to be in charge of technical issues. DPR is starting to ramp up for more production and development of policy/procedures. Web archiving project content will be bright; i.e., publicly accessible.

PAG is reviewing the disaster plans for both RLFs. Barclay Ogden (B), on behalf of the California Preservation Program, is working with OCLC to develop a workshop on digital preservation. CDL is going to develop a database to track what has been and has not been digitized via the Internet Archive and Google mass digitization programs. Ogden is working on a risk assessment model for collections.

C. UC Advancement conference (Brown)

Brown gave a presentation to a meeting of UC-wide advancement staff on archives perspectives, giving an overview of how we control materials, OAC, finding aids, etc. Stated that development records could be restricted for up to 75 years; we question whether that would hold up if a Public Records Act request were filed. Farrell noted that Berkeley is working to outline legal restrictions, looking at various types of records relevant to the Act; he'll bring to this group when they're done.

D. Next-Gen Melvyl pilot (Brown)

Most of UCLA's records for archival series aren't in OCLC and therefore won't be in the new Melvyl (most aren't in existing Melvyl either). Brown asked to what extent this is an issue at other campuses. Some it is; others it isn't.

5. Faculty papers presentation at SAA 2008

Both RBMS and SAA rejected Brown's proposal for a meeting session, but the incoming chair of the College and University Archives in SAA has invited us to present a 20-30 minutes session at their section meeting. We would present our guidelines (in contrast with the work promulgated by Yale) and also discuss how Greene/Meissner applies to faculty papers. Neal has prepared a document outlining Berkeley's implementation in terms of collection foci and such. Volunteers to participate: Neal or Farrell, Mix, Gartrell, Brown.

6. Guidelines for collecting Academic Senate records (Farrell)

Draft prepared by Farrell and Brown was distributed. Brown noted that UCLA collected materials from individual campuses until 1992 and offers to send campus-specific records if we would like them. A fair number of campuses are successfully working with their divisional Senates to bring in materials. This document will be very useful in communications with Senate offices.

Question: do we endorse this document? Yes, after receiving comments from the group by the end of June. Intent is to endorse this and post it on the UCAC website.

7. Digitization of student newspapers (Brown)

Brown has proposed digitizing the Daily Bruin in response to a local call for possible candidates for the UC mass digitization project. It was noted that the UC-wide projects aren't yet working with oversized formats. Three campuses have already digitized their newspapers or have this in the pipeline: I, M, SF.

Consider a UC-wide LSTA or NHPRC grant proposal to digitize and make available all UC student newspapers? The argument would be to make available digitally a major resource in the history of California higher education. Brown will pursue this.

8. UC terminology (Brown)

9. LAUC retention schedule (Neal)

Items 8-9 skipped for lack of time.

10. Scanning services in the reading room (Day)

Day allows researchers to bring scanners into the reading room, and she only recently began charging for digital scans made by staff. She asked what others policies and prices are. Only UCLA allows scanners in the reading room; others find it problematic for reasons of security, handling, disruption, etc. Neither Scripps nor UCLA allows scanning by researchers if copyright isn't held by the University; others questioned why that seems necessary. Bunting noted that HOSC plans to compile a matrix of reproduction and permissions services and charges.

11. Archives authority (Farrell)

Farrell compiled a brief documenting listing four UC documents that give campus archives the authority to collect, preserve, and make available the records of the University. These include a letter from Clark Kerr (6/02/64), RMP-1, RMP-2, and RMP-4. This document will be made available on the UCAC website.

QUESTION: My notes don't reflect whether David or somebody else is supposed to further investigate whether all this includes granting permission to publish materials to which the Regents have copyright. Please advise for the final minutes.

12. Campus e-records projects (Round robin)

13. Electronic theses and dissertations (Coy)

14. Archivists Toolkit (Round robin)

Items 12-14 deferred due to lack of time.

15. Web archiving project demonstration (Seneca, Coy, Mix, Williams)

Tracy Seneca is in the lead on this project and gave us an excellent presentation. Our UC Web Archive Project is CDL's first project working with other than U.S. government information, and the first that is an all-campus collaboration. She provided an overview of workflow and an introduction to some of their web capture tools.

The Web-At-Risk project involving the Library of Congress, UC, and several other universities focused on government information. CDL knew they would want to broaden its results for many other types of content. Seneca clarified that the Internet Archive (IA) captures Internet content very broadly but at a shallow level; CDL's intent is the opposite in both respects. CDL wants to be able to explore a variety of file formats beyond HTML. Internet Archive is a force behind many of the web archiving open-access tools that CDL and others use. Pages crawled by IA cannot be navigated in the way that live pages can. Wayback Machine is both the IA crawler and the software that renders web pages for viewing; the latter is a new file format: .warc, which CDL is

using. CDL is tasked with writing IA's tools documentation one aspect of collaboration between the organizations.

The commercial tool ArchiveIt is out there, but UC doesn't want to have our content stored on non-UC servers and so built its own tool. ArchiveIt also carries significant fees.

At present CDL is focused on creating tools that are useable efficiently by UC libraries, doing site-by-site capture of specific and highly curated collections, and enabling cross-campus collaborations that support shared collection development. Each user institution or site has an account (UCAC is an institution; we don't need to go use a general campus account but will each have our own), with various roles defined, including administrators and submitters/curators. Each institution has one terabyte of space and guides its own collection development policy. Coy is our institutional administrator.

Each project determines its own metadata structure; individual projects probably will differ greatly in this regard. Capture is defined by entering seed URLs (home page for a particular domain), frequency and other settings, and site-level metadata. Useful for curators to have extensive knowledge about the sites they're capturing in order to set the most useful parameters. Can set crawls to capture only certain file types. The crawl will capture all pages within that domain that have been linked upward to the home page, up to 23 hops from the original link. (The IA crawler goes up to about seven hops; specialized tools such as CDL's tend to focus on providing depth.) Will capture any embedded files, such as style sheets. Can set to capture external linked pages, but no hops beyond the linked URL. Reports don't yet state how much space the native captured files encompass. Files are compressed after being brought in and automatically are added to the DPR.

At present we have just one site list, which will quickly become unmanageable. We can instead set up multiple projects, each of which could have its own metadata structure. CDL will be providing better tools for managing site lists, which will lessen the need for separate projects. They don't envision building tools to enable moving content around between institutions or projects. This is likely to prove problematic for our content unless we build a highly expansive structure at the outset.

Copyright law section 108-study group looked at web archiving exceptions for libraries, but no results yet. Original grant for this service required inclusion of rights information; not clear yet what will be appropriate. Regardless, copyright-restricted pages can be crawled unless the host has a robots.txt file that prohibits capture.

UCAC has so far used 8% of our one terabyte. We noted that one terabyte won't last long once we get rolling. Space will become an issue. Space isn't expensive per se, but maintenance of saved content by the service center does have real costs. Audio and video files are particularly space hogs. Cost model for maintaining a project thats shared across campuses, such as ours, hasn't yet been determined. Tricia Cruse working on this with delivery date of November 2008. Must be determined before the grant that currently funds WAS expires a year from now. Charging will be at the institution level, not the project level.

After crawling, can view reports indicating sites captured, display content, navigate within a site, search full text of a captured site. Can limit a search by file type, such as PDF or audio. Once you build a collection, can search all the sites simultaneously, as well as setting various parameters and deciding what content within the collection to make available. Can select sites for a collection item-by-item or en masse.

Working now on duplicate reduction to eliminate files such as PDFs and multimedia that were already captured but would not have changed; software uses checksums to detect whether any

changes have been made. HTML files take up little space so aren't such a problem to recapture even if they haven't changed. Reports indicate new, changed, etc. files so that curator can manage parameters of the crawl. Looking at a report, there are levels: collection, site, crawls. We currently have only one project within our institution, so as curators we will all be able to e.g. set up collections.

Soft release planned existing projects this summer. Won't start any new accounts until a couple months of pilot experience. Will work with CDC to identify tier 1 and tier 2 projects, identify what comprises a meaningful collection plan, rank new projects, etc. As an existing project, we don't need approval from CDC at this point to add more content.

No access capability yet other than viewing contents from within project participants environment.

Tracy encourages us to continue working at we currently are rather than making decisions about how to split content into projects or other structural issues. Their work with CDC will identify the types of policy decisions that will be necessary to define a collection development plan for a project, such as metadata structure, what comprises a collection, and much else.

They're about to migrate all the data from test servers to production servers. We can't see the existing content until that happens the soft release due in August. Steve can then issue logins and passwords. [NOTE: Message went out from CDL this week that the release has been delayed until fall.]

Our UCAC project group developed a collection development plan that remains in draft; Steve will send it to us. CDL has a template for use in drafting a plan for our project. We can see all existing plans (Google web at risk Wiki). Yvonne Wilsons (UCI) and Kris Kasianovitz (UCLA) and Tamiment (NYU) are very good.

This project will be a very big deal for UCAC. How much do we need to restrict ourselves so that we don't spend our lone terabyte on the content that isn't the most important? Focus on core records, such as general catalogs and student newspapers?

Our project group will work some more on the draft collection development plan, then well have a conference call in about a month. Tracy will send us an inventory of the content that has already been crawled; Steve has notes on this also.

We're in the Wild West stage of web archiving. This will be interesting!

16. Next meeting

Jim Dooley noted that CDC encourages all its subgroups to have a maximum of one in-person meeting per year. We'll meet by conference call in October. Go to quarterly conference calls (and one annual in-person meeting), so next one in February. Spring 2009 at Merced?

[17.] UCOP reorganization and downsizing (Williams)

In the context of UCOP streamlining and downsizing, ca. 400 of ca. 1600-1900 UCOP employees have applied for voluntary severance and retirement. Connie retires as of June 27th. Many people have received layoff notices, including half of her staff. She is discouraged and uncertain about the future of UCOP records management. Does higher management understand the importance of recordkeeping? The importance of university records in the context of the variety

of legal issues that have involved evidence in University records apparently is not recognized. Her unit will report, at least temporarily, to the CIO's assistant. Not clear how well deal with our communication with UC records managers (who haven't met in quite some time). We'll need to figure that out.

Submitted by J. Dooley, recorder
6.20.2008