

University of California/Stanford University
Government Information Librarians Group
Annual Meeting
Tuesday May 29, 2007

Meeting Participants:

UC Berkeley: Jim Church, Harrison Dekker, Tim Dennis, Susana Hinojosa, , Jesse Silva
UCB IGS Library: Nick Robinson, Frank Lester
UCB Law Library: Alice Youmans
UC Davis: Linda Kennedy, Patsy Inouye, Marsha Meister
UCD Law Library: (did not attend)
UC Irvine: Kay Collins, Yvonne Wilson
UC Los Angeles: Jan Goldsmith, Kris Kasianovitz
UC Riverside: Lynne Reasoner
UC San Diego: Becky Culbertson , Rebecca Hyde, Anneliese Sklar
UC San Francisco: Peggy Tahir
UC Santa Barbara: Sherry DeDecker, Janet Martorana
UC Santa Cruz: Lucia Orlando
Stanford Univ: James Jacobs
Calif Digital Library: Tracy Seneca
Cal State Library: David Cismowski, Janet Coles
CDC Rep: Lucia Snowhill
**CDL Director of
Licensed Content:** Ivy Anderson

9:00 - 9:30 Coffee and morning refreshments

9:30 - 9:45

1. Welcome, introductions and announcements

9:45 - 11:00

2. Reports and Updates

a. Web at Risk project update – (T. Seneca)

- Running usability test today; demo will be tomorrow at WAR meeting
- NDDIIPP funding issue – LC lost funding but later restored and the grant has been officially extended from Jan 2007-09. 935,000 in funding made available to include end-user access tools. NDDIPP II is considered as anything after Dec. 2007
- Exploring continued collaboration possibilities with San Diego Supercomputing Center and Stanford Computer Science Dept.
- 2 pilots of WAS ; each test has had limited capabilities; time crawl; experimenting mainly with workflow and user interface
- 4th pilot release mid-July will focus on collection building; this will be the post-crawl gathering, especially run to grab only individual items vs. whole website
- The remainder of 2007 WAR will focus on infrastructure for realistic production level activity and preservation features. They want to make sure all of the behind the scenes programming is developed, which means we may not see a lot of changes to the interface.
- WAR is responding to concurrent changes in Heritrix crawler (developed by Internet Archive). Heritrix development has been driven by needs of agencies crawling on a much larger scale (i.e. national libraries, and who require pulling in entire web spaces, as opposed to discrete documents

or pages.) A smart crawler project is being developed that will allow the crawler recognize when the page has been changed or remains unchanged. This will increase the amount of information that can be conveyed to curators regarding changes to websites. WAR developers anticipate ability to set up an ongoing crawl of CA government sites and then deposit it into a shared account that individual selectors can dip into for use in their collections.

b. DLC Spring Meeting in Denver - (R. Hyde)

Rebecca distributed a summary of DLC to the group. There was not a lot of new information at this DLC – due to many issues – one being that both Bruce James and Judy Russell retired; here are the highlights:

- GPO unveiled FDLP Desktop – a “mySpace” for FDLP
- Working on developing a better backend to workflow/email processing systems
- Public Access Assessment – will be the new Inspection service
- 5 state conferences – IMLS grant to do online tutorials on finding US government information using Webjunction, freely available software
- Grant to develop training programs for librarians
- IM Reference around the country
- In response to a question, David Cismowski clarified that two California Libraries gave up depository status: Marin & Redwood City

c. CSL/UC Davis joint LSTA grant proposal (L.Kennedy & J.Coles)

- Janet and Linda have submitted a grant request to digitize California Documents - \$73,000 LSTA funds + \$30,000 in-kind contributions from CSL and UC
- Proposed project timeline: July 1, 2007 – June 30, 2008 (If the state budget isn't passed on time, there will be a delay – project will start once the budget has been passed.)
- Proposed titles to be digitized: Governors' Executive Orders, the Department of Fish and Game Inland Fisheries administrative report series; and the debates, proceedings, reports and documents of the 1879 Constitutional Convention.
- This is a good learning process to go through – and the group was encouraged to look toward utilizing LSTA for projects. It is a good way to get materials digitized AND provide public access. They approached the grant writing process as a local history project.
- Materials will be digitized by a vendor; then use Content DM to manage digitized objects. They will be using a Level 2 Content DM license. Metadata will be created in-house and the records will be OpenURL compliant; or they could use bibpurls to do link resolving. Originally, the intent was to upload to Calisphere – but at this time it is not possible due the way OCLC exports objects and metadata (Calisphere can't currently take pdfs). They are looking into other hosting options, perhaps through CALIFA.
- **Action:** Linda and Janet will let us know if/when the proposal is accepted.

d. LOCKSS-DOCS - (E. Cowell via Email)

The Round Robin report from Elizabeth Cowell, Stanford University indicated LOCKSS has a group of partners working on a pilot project for government information, and they are interested in working with the UC campuses as well. This email generated much discussion, and some questions:

- Are there any other campuses with an interest in using LOCKSS? UCB has signed on; UCSC interested; UCD had interest a few years ago; CSU Fresno (Carol Doyle) would also like to use it.
- UTLANIC running a web crawling site for Latin America
- A lot of discussion was generated about the service and the costs and workload associated with running a LOCKSS box:

- The group agreed it would be useful to gather more information and keep this topic on our agenda. **ACTION:** UCB will report regularly on their progress.
- **Action:** Lucia Orlando will ask Elizabeth to give an update on LOCKSS and the project alliance partnership so we can get all of the details about becoming a LOCKSS partner.
- How does this support the idea of open access to this material especially if you need to be part of the partnership to get access to all materials – not just those that you add?
- From a CDC point of view – how would this increase collaboration among UC campuses? What does this service do – or not do – in comparison to other systems (especially free systems)?
- Is this a way to capture CA documents?

e. Shared Cataloging Project - (B. Culbertson)

- Statistics for this year: over 1000 serials; 3250 monos for CA Docs cataloged
- Staffing: 1/2 time cataloger doing this work; Donal Sullivan hired; he is also a programmer
- SCP has been distributing records with digital archive link since last June
- As soon as materials with online access show up in CSP, SCP is cataloging them.
- Ivy noted that CDL is very interested in the potential of this project for CA docs. CDL also interested in how WAR might take care of problems associated with Content DM.
- SCP has identified principles for persistence at this point and wants to make decisions that are consistent with CDL's progress;
- Project: UCD working on old SCP records with bib purls (Global) or pids (used locally). SCP will change the URL to a DAL behind the bibpurl, so records won't have to be redistributed.
- BIG Thanks to Yvonne Wilson – for all of her work on identifying agencies and submitting them to Becky. All of us are encouraged to email Becky with suggestions. Please be sure to check list to make sure that the agency isn't already on the list. If you find specific publications, also email Becky.
 - The “Priority List of California Agencies” cataloged by SCP is available on the UC GILs website: <http://www.library.ucsb.edu/hosted/gils/projects.html> **Action:** Becky will send updated list to be added to the UCGILs website.
 - Janet gave Becky CSL's agency priority list. Becky will share her list with Janet.
 - Workflow and this list – when doing collection development, we need to know what agencies will be covered and on which timeline materials are being cataloged.
- Question: Is there a prioritization process for SCP projects?
Answer: Depends on the project. If we want to suggest a new agency for SCP to monitor and catalog, just email Becky. For new projects, e.g. adding records for all SourceOECD material, then a more formal submission goes through Ivy's group and JSC. Ivy's group will be doing more analysis of priorities and general analysis of these kinds of projects and associated workload. .
- Email Becky to suggest or recommend a new agency.
- **Action:** there was general consensus that people want to review list and adjust priorities.

f. GILS CDC Report/ JSC nominations- (L. Snowhill and I. Anderson)

- JSC/CDL working on their workplan. It was noted that most requests for most resources have been filled.
- An email has been distributed outlining what has been by CDL licensed thus far. Lexis Nexis Academic is on the list and is an administrative change only.
- SourceOECD – look for email from Lucia Snowhill; 8 campuses subscribe to online; 4 subscribe to print too; IEA Statistics as well. Wendy and JSC are trying to determine what kind of joint subscription they should license. The group suggested that IEA Statistics be an opt-in for a full electronic subscription for everyone, and 1 shared print copy.

- CDC is reviewing negotiation and value pricing of materials. Weighing the impacts of: UC silos, digitization, shared print, and born digital initiatives. They have formed a number of small working groups to address specific areas and are trying to negotiate which directions to go. A value-based pricing model group is working with Ivy. There is also an eBooks group looking at eBook subscriptions. What do we want with resources like PLOS; RLF duplication and criterion issues - more targeted issues than just persistence? A newspapers group (Kay Collins is chairing this). A group is addressing how to manage shared print. Will shared print be more prospective? It will involve managing all aspects of it which means working with campus groups to do retrospective and prospective. Final determination depends on the direction of the shared print program.
- The Persistence policy for RLF's has been developed, now they are working on criteria for trusted copies. The question about allowed duplication between RLFs was raised. These are not dark archives – most of the group favors having duplication among the two storage facilities – since the materials circulate, can be come damaged or lost. Also, this provides better access to campuses in the North and South (don't have to wait for a copy from the North if a Southern campus request it). CDC is looking at when it is appropriate to intentionally duplicate –and when should there really only be 1 copy

11:00 - 12:30

3. Digital Collection Projects & Collaborations

- a. Web at Risk project (T. Seneca) [replaced in agenda with morning report]
- b. OCLC Digital Archive (N. Robinson)
 - OCLC Digital Archive (OCLC DA) A report from the field – how UCB IGS Library has been using the OCLC Digital Archive (See PowerPoint)
 - Mission and collection policy (historic collection of local budgets, planning documents, etc.) informed IGS' decision to use OCLC DA
 - Born Digital objects are as ephemeral as the print items which the IGS library has had a strong collection. IGS has a commitment to provide access through catalog records for their materials; they also catalog in OCLC; and have a strong practice of creating analytics for records. They also want to capture discreet documents.
 - 4 UCB affiliate libraries are subscribing to OCLC DA: Transportation, Water Resources, IGS, and Law. They split the annual fees and storage fees 4 ways; all but Law are OCLC cataloging units; they earned cataloging credits and have been able to help pay for the subscription through these credits.
 - In addition to utilizing the OCLC DA, IGS was able to test and participate in two other OCLC projects: Conent Cooperative Project and the Workbench Beta Field test (please contact Nick if you have specific questions about either of these).
 - See slide on workflow: Create dig archive record > Analysis of objects in order to set up spider to harvest objects > Harvest > Ingest. You could switch this workflow though, this is how IGS has chosen to do it. Once captured and cataloged, the url in bib record points to OCLC DA metadata and this points to the actual object in the Data Archive – but the click through's are seamless to the user. You can capture discreet objects, entire website, or even select specific links on a page to capture, while others are not selected for capture.
 - Access to digital object via Worldcat record – shows up in Melvyl approx 10 days; also in Open Worldcat; potentially in Google, Yahoo, etc. The url can be copied into webpages; database etc. so access to archived copy is very portable.
 - The 4 affiliates are working on collection planning together and have a monthly users group meeting to share collection plans and experiences. Rights Management issues also arise so they

formed a DRM task force to work on obtaining materials from non-governmental organizations with permission.

- Current materials they are collecting: County Grand Jury, Budget, Financial Reports, California Planning documents. They hope to collect publications from the following agencies/organizations: CRB, LAO, Little Hoover, Senate Office of Research, Public Policy Institute of California (PPIC), CA Policy Inbox – blog that tracks publications from an extensive list of organizations. This list will help inform priorities for collecting and getting permissions for capture
- Cataloging issues – single record vs. multiple records. Ownership of base record creates need to create a new record sometimes.
- Web Archives Workbench: 4 tools for discovery, properties, analysis, harvest. This will be released as open source in July 2007. See: webarchives.oclc.org/WAW
- Changes to OCLC Dig Archive – next implementation will be a preservation only; access copies will via CONTENTDM
- Is the workload sustainable: it is significant; this solution won't be scalable for all CA State and Local docs – but has met the needs for core collection needs (local depository)
- Storage of digital objects – can be at OCLC or other repository;
- WAW harvester works better than the OCLC Digital Archive e – Janet found out recently
- Changes to staff? same level of staffing as in the print world. PT cataloger; 2 lib assts 3 & 4 who work about 90% of time; students to do discovery.
- UCB Doe/Moffat to collaborate with IGS as they embark on the LOCKSS project.
- How is Nick using WAR vs. OCLC – whole websites; political blogs; vs. OCLC cataloging at the item level.

c. SCP/CSL/OCLC Archive (P. Inouye)

- This ties into what UCD is doing with the old SCP records and adding to OCLC Digital Archive.
- They have lists of titles in SCP up to a certain month – around 2007 – a “capture file” and are working from the file.
- The project impact will go beyond UC to include public libraries and CSU's
- Using WayBack machine to pull in content if it is no longer available from old SCP records.
- Suggested that we could also ingest WayBack machine links into the archive.
- As of July the process is as follows: download to desktop; catalog and pull in the object – won't have to run the harvester; these harvester tools will disappear as of July; and will use the ContentDM
- Can't access html content via ContentDM
- 3 Levels of use:
 - Level 1 \$9800 can archive up to 10k but metadata is also an object; compound objects also counts as objects
 - Level 2 \$20000 up to 50,000 objects
 - Unlimited license \$50,000 one time fee; maintenance fee \$9800/year (other fees for dark archiving??)

12:30pm - 1:30pm LUNCH on our own

d. Coordination and collaboration on digital projects (all)

Should we purchase the Readex Serial Bibliographic Records? (M. Meister)

- Marcia passed out a handout which outlined the cost and questions which we have concerning the purchase. The cost would be a one-time purchase of \$67,760, if we purchase by 6/30/07. We would start with 150,000 records with a final goal of 354,000 records. We discussed the fact that this price

seems very high compared to the price we paid for CIS records. We thought if the records were purchased just for MELVYL, or as if we were one campus, the price might be lower; but most vendors will always consider us 10 campuses, or will price us by the number of possible users.

- Becky Culbertson said there is a possibility we could create our own records at a much lower cost, but we would need a “Harvester,” which only UCLA owns. Becky thought if UCSD had the Harvester they could make the records, but a Harvester costs \$25,000, and requires special IT staff. Ivy said we need to formalize a recommendation so it can go to JSC, which might refer it to HOTS. The survey request from JSC will be coming out in 4-6 weeks.
- **Action:** Kris to find out more about the UCLA Harvester.

Clarification about the “Harvester” – provided by Stephen Davison, Head of UCLA Library Digital Projects via email, June 5, 2007:

Yes, we have a metadata harvester, built specifically for the Sheet Music project. It doesn't harvest automatically, and it is configured to harvest sheet music metadata from specific sites. If we were to reconfigure it for another purpose such as yours it would require programming resources. If you are interested we could look into it. The question would be whether to build upon what we have or to purchase from a vendor I guess. There are pros and cons to each. S short presentation about it:

<http://ismir2003.ismir.net/papers/Davison.PDF> . Also <http://digidev.library.ucla.edu/sheetmusic/OAIProject.html>

Further Clarification from Becky, provided via email, June 5, 2007

I also found out that the cost of the III software was \$20,000 not \$25,000.
(116) XML Harvester

XML harvester provides an automated cataloging tool which can create MARC records from XML metadata stored on remote servers. These MARC records are loaded into the Millennium database with URL's that offer links to the digital objects stored on the external server where appropriate. The XML Harvester is OAI-PMH compliant.

Includes Profiling Services necessary to load MARC records created by the harvesting process.

- Other questions: Why weren't the price of records negotiated at the time of the Readex Serial set purchase? And if our new catalog is through OCLC, could it then take Dublin Core Records? Sherry suggested that Wendy Parfrey tries to renegotiate the price. Alternatively, we can try renegotiations with Readex, but once you do that there is the expectation that if you get a good price, you will buy the product, so we need to make sure we want it. Becky brought up the MARC records for the Goldsmith-Kress collection, which were priced at \$50,000 for 61,000 records; Becky can get them from OCLC, but not at the title level.
- **Action:** Lucia Orlando asked Becky to look into putting forth a proposal with all the various options and we can discuss it further on email.

Digitization Projects (Ivy Anderson)

- **Google Project** – mass digitization - no selection – all RLF based – 325,000 sent to Google of which 210,000 are scanned of these, 15,000vol/week are being processed; less than 2 week turnaround. Commitment is to digitize 2.5 million volumes. Those materials under copyright only have “snippets,” while those not under copyright are complete. Probably a lot are duplicates of what's being done on other campuses, but they are not de-duping. Their mission is aimed at discovery not getting the best digitally rendered copy. Still more learning on both sides
- **Microsoft project – MS/OCA** – UC signed a separate license with Microsoft – done by Internet Archive under the auspices of OCA. There are more restrictions on the content: not allowing

commercial downloading, data-mining or harvesting, etc. by commercial entities. They are very supportive of non-commercial reuse, that is, making content available to other non-commercial entities. Volume is much lower and more selective of content, for example selective material in specific subject areas is and out of copyright materials only; at roughly 100,000 and 200,000 volumes for this year. They are digitizing both SRLF and NRLF materials. Some materials are rejected, if they are too brittle to scan, or if they have foldouts. They are developing ways to capture materials that contain un-scan able information, e.g. foldout pages, poor condition, maps, etc. Selection is based on broad subject areas that MS is interested in and English language only, but they might be open to CJK soon. Selection is very labor intensive in this project. CDC has put into place a Mass Digitization Advisory Committee, chaired by Robin Chandler, to identify and prioritize collections to be digitized; reviewing which campuses and collections on the campuses will be added to the mass digitization of these materials.

- Discussion and Issues:

1. Creating access to records? Should they be put into Melvyn and then flow back to the campuses? Should they be put into WorldCat? Some analysis has been done by Karen Coyle – to determine workflows – potentially not putting records into Melvyl – but do a pilot at a single campus.
2. Will we be able to use a single record approach if we go through OCLC? Not sure. Have to find out from OCLC.
3. Will UC campuses have access to copyrighted materials? Michigan is still waiting to find out what the outcome of the lawsuits will be, therefore CDL not presently developing any access servers at UC. CDL is still working out the operational issues; not yet addressing the access issues.
4. Are hearings copyrighted and accessible? Google is not making them available, but the U. of Michigan is (since hearings may have some copyrighted materials within them). A lawsuit has already been settled on this issue, ruling that it was “fair use” to publish a hearing even if it contained copyrighted material.
5. Are we duplicating what other universities are already digitizing? There may be good reason to duplicate digitization, especially, if we cannot get access to materials via Google; it depends on what is being made accessible.

4. Public Services and Training (all)

Projects and opportunities for collaboration:

- a. Revised GILS webpage – see <http://www.library.ucsb.edu/hosted/gils/index.html>

Action: Everyone take a look at the website and send comments. Please pay special attention to the directory information. Send updated contact information for your campus to Sherry DeDecker.

- b. Websites - we tabled this for discussion via email

- c. Chat Reference (J. Silva)

- See round robin report and Report on their Chat Reference Services:
<http://sunsite3.berkeley.edu/wikis/govinfo/index.php?n=Main.ChatRefReportF2006>
- Chat Reference use has been going up exponentially. Government Information librarians have been encouraging librarians within the library to utilize it for GI questions. They are hoping to work collaboratively with IGS and other units that have a lot of GI questions.
- Their open chat hours depend on staffing; they'd like to do more hours, but don't have the staff. Currently they are doing Tuesday-Thursday, 1-5. It was suggested that all the UC's collaborate on this; they would also like to expand on their own campus, and involve more librarians. People

want night hours, but someone would have to stay late or answer from home. If we collaborated, each campus could take a night.

- They are utilizing Meebo for their chatbox because you can chat with all major IM applications.
- They still maintain email reference.

d. Blogs and Wikis (H. Deckker and T. Dennis)

- Tim “skinned” a test wiki of the UC GILs page, using pm wiki
- The group all agreed that using a wiki for many of our pages, especially the directory pages, is an excellent idea. Is it feasible to move the GILS pages to a wiki? UCB expressed willingness to maintain it.
- New [DataServices website](#) managed by library wiki software -- pmwiki
- Government and statistical CD-ROM re-digitization project. Project includes copying CDs onto library server, providing access in the DataLab, and, depending on copyright, remotely.
- Working with Indiana Project <http://cgi.cs.indiana.edu/~geobrown/svp> They will be adding the interfaces to run the CD's; will be able to keep track what it takes to run each CD. There may be a call for unique CD titles to add content. Unmediated use and install that creates a virtual desktop – where the user interacts with the data on a virtual desktop Using Vmware – install once on a library server and then when someone accesses the material, it opens up in a virtual desktop.

5. 2:45-3:00: Wrap-up and review of action items

Lucia Orlando announced we need two new people for the Steering committee, one Member-at Large, and one from the Northern campuses.

Action: Lucia will send a call for nominations via email.

Minutes submitted by Kris Kasianovitz, June 5, 2007

Morning minutes: Kris Kasianovitz; Afternoon minutes: Jan Goldsmith