July 25, 2011

California Digital Library's Suggestions and Opportunities
SOPAG's Digital Library Services Task Force 2 Report

CDL has reviewed carefully the DLSTF2 report and believes it represents a realistic plan for making progress towards a long term goal. In addition to the CDL staff members who served in various capacities on the Task Force, Executive Director Farley was invited to listen to SOPAG's discussion of the report, and CDL senior staff have discussed the report in detail. Following on the Task Force's approach of building on existing technologies, expertise and infrastructure at UC, the present document outlines what CDL has to offer today in support of the DLSTF2 vision, and suggests several potential paths forward. We are ready to commit resources to implementing new components as well as participate in developing policies and organizational infrastructure.

Complementing the vision outlined by DLSTF2, CDL has been collecting additional use cases in discussions with a variety of campus communities including visual resource curators (as part of the investigation requested by CDC), affiliated libraries like the Transportation Studies library at UCB, and librarians at UC Irvine, UC Santa Cruz, UCLA and others. Through these discussions, we are gaining a better understanding of how the parts for creating a UC Digital Collection could fit together to support various audiences, access permission preferences, workflows, and technical choices.

Where possible, we offer more than one option and expect there may be other ways to approach a task; we welcome further discussion with campuses. In some cases, we believe it is possible to speed up the suggested timeline whereas in other cases, the time and resources may be underestimated.

CDL Services and CAMP Recommendations

The use of the CAMP (Create, Access, Manage, Preserve) framework for evaluating the technical and organizational infrastructure is one that CDL has also used and adapted.

In general, CDL has solutions for Preservation (as noted in the report) provided by Merritt as well as individual microservices, and has already developed many of the guidelines that underpin the Creation phase and could assist with other areas.

Digitization of campus collections:

While the report identifies digitization of campus collections as primarily a local matter, it also refers in several places to the need for coordination of digitization workflows and standards to ensure interoperability across the UC Collection. CDC is also working on a digital collection development strategy as requested by NGTS, which may result in specific recommendations aimed at building the UC digital collection. The CDL has experience in coordinating end-to-end digitization workflows across multiple institutions utilizing third party agents that may be helpful in this area, both for mass digitization and for the Local History Digital Resources Project that we have conducted for many years in collaboration with the State Library and Califa—developing and managing vendor contracts, coordinating and scheduling workflows, developing object and ingest standards and tools needed for aggregation and interoperability, and ingesting content from disparate sources into a common system. CDL would be willing to work with the campuses to better understand what specific services might be helpful in pursuing more aggressive digitization of local collections and whether there are some needs that should be resourced at the systemwide level. We note that one of the recommendations in the CDL Review reflected a desire for additional support from CDL in this area. Additional staffing would most likely be required if CDL were to take on such a role.

Access and Management

CDL has Access solutions that could be readily adapted to support DLSTF2 goals as described below.

It is the Management area (as defined in the report) that is most in need of both technical and organizational infrastructure integration and/or development.

We have provided more detailed responses in the report itself but offer other comments here on the Access and Management functions.

Access

There are multiple ways to accomplish the goals expressed in phase 1 to aggregate existing metadata, both in the short term and later to provide a richer access component. More work on articulating use cases would be beneficial for guiding the best development path, a possible task for POT1's lightning teams. If the Access component envisioned is to provide a more discernable UC collection as well as to ensure a rich environment for viewing and rendering objects, then it is entirely possible to address this goal by developing harvesting or crawling techniques, along with a presentation layer and rendering tools.

1) Melvyl: Many of the goals and recommendations around providing discovery of the vast amount of digitized content across the UC Libraries can be accomplished by ensuring that the metadata resides in Melvyl (based on WorldCat Local) with links back to the digital objects stored locally.

   We already have well established workflows for getting metadata describing digital content into Melvyl. The campus libraries' bibliographic metadata is made discoverable in Melvyl when they catalog in OCLC, or they push their records to OCLC. Digital metadata (either born digital or made digital) that support the OAI-PMH protocol could be harvested by Melvyl via the Digital Gateway for content in OAC/Calisphere and eScholarship, pending further policy and technical assessment. There is a long and established path for bib records to be included in Melvyl, and the Next Generation Melvyl Team has been working with OCLC to improve their OAI harvesting tool. CDL plans to develop guidelines for making metadata available to OCLC's OAI harvesting service, and once the new advisory structure for Melvyl is approved, several policy issues related to this mechanism could be addressed.

   The HathiTrust produces files of metadata that get pushed to OCLC and consumed by OCLC by yet another mechanism. By utilizing all of these methods, many of the goals expressed in Phases 1 and 3 can be achieved.

2) Vertical search crawler: given the potential complexities involved in harvesting and rationalizing metadata (see 3) below), an efficient alternative could be to use a crawler to bring together distributed collections in a common access layer.

   While a crawler may not provide a long-term integration layer, it may provide a low-barrier method for quickly achieving the Phase 1 objective of providing a central point of access to the UC Digital Collection, affording the opportunity to simultaneously start investigating long-term DAMS solutions.

   CDL developed a prototype vertical search crawler service for the UC Portal pilot project sponsored by UCOP in 2009 (see http://libraries.universityofcalifornia.edu/planning/slasiac/052109/IPBS_description_SLASIAC_15May09.pdf; project no longer supported) and for a demonstration of how to bring together distributed collections for the Water Resources Center Archive. Although neither of these prototypes is still running, CDL will be applying the same technology as a prototype for the Digital Public Library of America (DPLA) beta sprint (see http://blogs.law.harvard.edu/dpla/) and would be happy to explore its applicability to DLSTF2. Such a service could enable discovery across a targeted group of content hosted in a variety of digital library platforms (for example, websites using Omeka, CONTENTdm, or locally developed content management systems or websites). We are also investigating the possibility of using this method to crawl or otherwise include publicly-accessible resources contained in the HathiTrust repository.

   A crawler system like this could be built entirely with existing open-source applications (as we will be demonstrating for the DPLA beta sprint), including Apache Nutch, Apache Solr, Lucene, and related technologies. Unlike the high level of coordination with metadata providers or search targets required for metadata harvesting, this approach may provide an effective access layer for relevant content with a low barrier to entry and rapid deployment. By demonstrating the type of content available, it could provide a quick win that would inform decisions about whether and how to aggregate the UC Digital Collection in a more fully featured manner. It is also worth noting that in response to the Digital Public Library of America call for proposals, Geneva Henry of the Center for Digital Scholarship at Rice University, has been contracted by DLF to research different methods of aggregating digital content which could inform UC's choices.

3) Harvesting: If the Melvyl option and/or the crawling option are not sufficient for phase 1, then harvesting and displaying metadata (as suggested in the report) is possible.

   CDL has the capacity to harvest metadata and convert to formats such as Dublin Core or MODS as recommended by DLSTF2. We have gained experience with the OAI-PMH protocol and "best practices," in particular, through research and development projects such as the Hewlett Foundation-funded American West initiative (http://www.cdlib.org/services/dsc/projects/amwest.html), and we also have extensive experience ingesting and rationalizing metadata from disparate sources in OAC and Calisphere.

Depending on the quality of metadata constituting the UC Digital Collection, we believe it could take considerably longer than the six months estimated by DLSTF2 to implement a harvesting framework. Based on our experience with ingesting heterogeneous metadata from multiple sources, we anticipate significant work may be required to map and normalize metadata so that records are suitable for indexing and display. Metadata harvesting also presupposes that each UC campus has systems in place to expose metadata in a harvestable format (OAI-PMH or otherwise)— this may require additional implementation time in some cases.

For an access layer for the harvested metadata (recommended in Appendix A, Phase 1), CDL could offer XTF for indexing and to provide a unified search tool. See http://xtf.cdlib.org/xtf/ for examples of how XTF is being used in CDL services as well as by a number of other institutions. The capacity of XTF could be extended to provide richer access to different content types. XTF is already optimized for search engines, and CDL is beginning a project to research additional SEO techniques.

SOPAG expressed interest in subsetting the collection by discipline or type for export or combination with other content as another reason for creating a UC collection. It would be possible to support this function with XTF as well. As stated above, the real challenge is in creating metadata to serve the purpose intended. By aggregating and exposing into other environments, we will learn a great deal about metadata gaps and successes.

It is also possible for XTF to index and display content and metadata stored in Merritt, UC's systemwide repository service. For library collections not yet in Merritt, some form of deposit into Merritt would need to be undertaken.

Although not specifically addressed in DLSTF2, CDL assumes that the UC Digital Collection would overlap and exist separately from existing XTF-based services such as OAC, which includes content from other non-UC institutions and is heavily customized to support finding aids, and eScholarship, which is optimized as a publishing platform. Calisphere, which is perhaps the closest analog to the UC Digital Collection as envisioned by DLSTF2, could be scoped, re-oriented and/or re-branded to provide a UC Collection view. CDL is currently working to accommodate additional types of digital objects (e.g. audio and video materials) into all of these existing services.

CDC has commissioned a study of campus needs for visual resources in preparation for evaluating UC participation in ARTstor's Shared Shelf service. It is possible that the findings from this study (to be completed soon) may indicate the need for an access layer optimized for this type of material.

Management

The Management recommendations focus on two main components, a DAMS and a rights management framework.

Digital Asset Management System (DAMS)

CDL would be interested in assisting with the assessment of the DAMS identified by DLSTF2, in addition to other potential alternative solutions, as determined. CDL can offer flexible support with the development of a basic DAMS in early phases, integration with existing systems used by campuses (such as WebGenDB at UCB and UCSD's DAMS), and/or hosting a DAMS such as Islandora or a collection management tool such as CollectionSpace if further investigation favors one or more of these as a systemwide solution. CDL has already begun preliminary discussions with UCB (which is developing CollectionSpace—see http://www.collectionspace.org/) and is investigating Islandora in partnership with UCLA. UCSC and CDL are also discussing how to integrate Merritt and Omeka. It may be desirable for CDL to install some of these or other candidates as an aid to further evaluation.

There are many definitions and variations of DAMS, and systems may be optimized for a type of content or purpose. Regardless of definitions, some possible configurations include the following:

1) CDL creates a basic tool for assigning metadata, rights, and identifiers via EZID, in bulk or item by item; supports any file format; allows editing and deleting of metadata (and associated objects). EZID already supports assigning an identifier, assigning and maintaining associated metadata, and maintaining target URLs, and Merritt supports multiple versions of a file. Support for various authority files could be added in later phases after requirements are fleshed out.

2) CDL integrates existing tools such as CollectionSpace (optimized for managing print and digital collections) or Islandora (based on Drupal, typically used for web content management) with Merritt. Campuses and/or CDL could host these tools.

Rights Management

CDL is eager to contribute to future investigation of policy and operations issues. Many of the same issues overlap with eScholarship and UC Shared Images so there may be synergies in supporting both needs.

Currently, Merritt supports the assignment of rights to individuals and institutions at a collections level and soon at an object level, and XTF can support display decisions at the object level although a modest amount of work is needed to define permissions at the level of the user in addition to the IP level. A consistent, UC-wide practice for recording rights metadata will provide the means to express the full range of rights assertions that can be attached to content. Then we can further develop the technical means to enforce any rights restrictions in order to support more granular access control to material.

Organizational Recommendations

CDL is willing to participate actively in the exploration of these and other options and expects that SOPAG will determine the appropriate models. We have comments on a few areas where we can offer particular expertise.

- Project management for each phase:
  - CDL is participating in the project management structure already put in place by SOPAG for NGTS.
  - We suggest there is also a need for technical and service management for any services provided on a systemwide basis, whether hosted by CDL or campuses.
- Outreach and training for all campuses regarding new processes and procedures
- Development of Centers of Expertise.:
  - By virtue of Stephen Abrams' work on JHOVE2, we have expertise on technical information for most object formats. It may be possible for us to host a "digital forensics machine" to assist with managing old formats (like floppy disks).
  - We also have a staff member, Jason Colman, who has extensive expertise in digital video materials (including videography, editing, compression, and web delivery).
  - We suggest that a column on metadata expertise be added to the inventory since that will be a critical skill to identify.
- Development of new cost sharing models for UC-wide digital projects and infrastructure support:
  - Ivy Anderson's Collections team has expertise in modeling different coinvestment scenarios and CDL has begun to explore models for supporting services.
  - UC3 is exploring a new cost model that will allow Merritt users to pay once and store forever.
- Development of a marketing plan to help drive users to the collections.
  - CDL could help in developing this plan.
- Development of a method for collecting feedback that would inform future improvements to the UC Digital Collection.
  - CDL has experience in collecting feedback for services and could consult on various methods.

Phases and Timing

While the phases proposed seem reasonable, CDL has a few observations and alternatives to consider.

The proposal for Phase 1 is to harvest metadata from existing digitized collections, in particular from campuses that already have DAMS. The benefits to those campuses include: 1) aggregation with other UC collections; 2) improved search engine optimization.

However, it may be desirable to put more emphasis in this phase on providing at least a basic DAMS for metadata and objects for those campuses that do not yet have that capability. Depending on the technique selected for aggregation and indexing/presentation (see earlier discussion), it may be possible for CDL to work on both of these requirements in phase 1.

Another consideration is that campuses can store objects in Merritt in phase 1, for those that have existing digital content but no DAMS solution yet. There is no technical reason to delay putting objects in Merritt. Once they are stored there, it is possible to normalize, index and display them with XTF or other tools.

In terms of born digital objects, again there is no reason for campuses to delay putting these into Merritt if they have existing workflows and management systems. The Web Archiving Service is an example where there is a creation tool, an access service and a connection to Merritt for web published digital content. Work is under way to create MARC records for web-harvested content, at least at the collection level, which could feed into Melvyl for discovery. Other types of born digital content have more complex issues and definitely need attention later in this process.

Alternative phases might be:

Plan A
Phase 1:
a) DAMS: CDL hosts one or more existing DAMS products for evaluation and integration with Merritt for campuses without this component or those ready to migrate. If these prove inadequate, CDL could build a basic DAMS.

b) Access: Two possibilities exist for an access layer and either or both could be pursued. We assume that the Melvyl option will happen in any case, although it may take some time to fully rationalize the OAI harvesting service.

1) Aggregate metadata for all campuses through Melvyl (HathiTrust already included; could use OCLC's OAI harvesting service for Calisphere/OAC, eScholarship). Incorporate into the Digital Object Guidelines information for working with OCLC's harvesting service.

2) Aggregate metadata from all campuses for indexing and display using a vertical search crawler; could include OAC/Calisphere, eScholarship. This would provide a discrete view of the UC digital collection, apart from Melvyl and would be a less resource intensive option than additional OAI harvesting. It could also include access for all campuses, regardless of the status of their DAMS solution.

Phase 2:
a) DAMS: Extend DAMS component for features not available in phase 1(e.g., authorities, crowd sourcing, etc).

b) Access: Develop richer presentation/rendering environment for digital objects deposited into Merritt or available from the DAMS component using XTF; expand SEO, branding options.