Next-Generation Technical Services (NGTS) POT1 LT3C: Evaluate Crawler and Harvesting Solutions

Final Report

December 12, 2012

POT 1 LT 3C Group

Garey Mills, UC Berkeley Lisa Schiff, California Digital Library (CHAIR) Brian Tingle, California Digital Library

Table of Contents

Executive Summary
Discussion of Charge
Task 1: Scope a new application based on harvesting and crawling4
Defining Harvesting and Crawling4
Refinement of Requirements5
"Phase 1" Prioritized Requirements for a Harvester/Crawler Based Solution
Scoping Experiment
Harvesting7
Crawling10
Harvesting Plus Crawling13
Effort Estimate16
Results17
Task 2: Develop an appropriate metadata schema that will accommodate discovery and display needsand SEO goals21
Task 3: Determine and test for feasibility with collection owners/managers a proposed workflow to
notify UCLDC about services/content for inclusion in UCLDC
UCLDC Registry Concept Survey
Responses to Questions
Summary and Findings24
Specific Findings25
Recommendation25
Appendix A: Prioritized Discovery and Display Requirements
Appendix B: Registry Concept Survey

Executive Summary

The Power of Three 1 (POT1) Lightning Team 3C (LT3C) was presented with the following charge:

"Based on the discovery and display requirements from Lightning Team #3.A, Lightning Team #3.C will develop a scope of work estimate for building a vertical crawler and access system, similar to the <u>Digital Public Library of America Beta Sprint</u>. This solution would serve as the initial UCLDC discovery and display interface (possibly in addition to other discovery solutions, e.g. WorldCat Local) and could potentially morph into the access interface scoped out by Lightning Team #1.A."

To fulfill our charge, LT3C developed an experimental system, combining harvesting and crawling strategies to build a very basic discovery and display system and developed a prototype of a Registry concept designed to both alert the UC Library Digital Collection (UCLDC) of new collections and to provide a mechanism for capturing collection level metadata. The experimental system was evaluated against a refined list of discovery and display requirements and assessed in terms of the relative worth of such an effort and the Registry concept was vetted through a survey of collection owners.

The resulting analysis led to the conclusion that it would be viable to develop an initial, basic access interface for the UCLDC with a combination of harvesting, crawling and a Registry. Such a system would require at least nine person months to develop.

As a final caveat, while the system described above would be a productive initial step towards the development of the complete UCLDC, it would not in any way fulfill the need for a systemwide DAMS designed to support robust stewardship of UC digital materials, which is one of the fundamental goals of the UCLDC.

Discussion of Charge

The Power of Three (POT1) Lightning Team 3C (LT3C) was charged with exploring the amount of effort required to implement a crawling and/or harvesting based access system that could provide a jumping off point for the UC Digital Library Collections access system. Our formal charge read as follows:

"Based on the discovery and display requirements from Lightning Team #3.A, Lightning Team #3.C will develop a scope of work estimate for building a vertical crawler and access system, similar to the <u>Digital Public Library of America Beta Sprint</u>. This solution would serve as the initial UCLDC discovery and display interface (possibly in addition to other discovery solutions, e.g. WorldCat Local) and could potentially morph into the access interface scoped out by Lightning Team #1.A."

Three discrete tasks were associated with the above charge:

1.3.10. Develop a scope of work for building a new application (the UCLDC) that meets the requirements established by POT LT 1.B and 3.A.

1.3.11. Develop an appropriate metadata schema that will accommodate discovery and display needs and SEO goals.

1.3.12. Determine and test for feasibility with collection owners/managers a proposed workflow to notify UCLDC about services/content for inclusion in UCLDC.

The LT3C team interpreted the charge and task list to mean that we should scope out the extent and feasibility of harvesting and/or crawling solutions for achieving the discovery and display requirements identified through the work of previous Lightning Teams and that we should also attempt to find reasonable mechanisms for getting new collections included in the UCLDC at that service's earliest incarnation, even if that first incarnation just attended to discovery and display services.

Task 1: Scope a new application based on harvesting and crawling

Defining Harvesting and Crawling

As a first step, LT3C members spent some time discussing the relative merits of harvesting and crawling. We defined harvesting to mean capturing metadata about an item in a collection via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface. Because the OAI-PMH standard is widely used within the library community, it was likely that at least some, if not many, collections intended to be included in the UCLDC would make information about their materials available in this way. Other harvesting mechanisms certainly exist, but they would be specific to each given collection/application, which would make it impossible to build a service around them. However, despite the high level of adoption of the OAI-PMH standard, there is no universal way of alerting the world to the availability of an OAI-PMH service for a given collection. Additionally, the manner in which metadata is exposed in OAI-PMH varies from service to service. Thus while promising to a certain extent, harvesting comes with serious limitations.

Crawling was defined to mean the use of robust third-party services that are available to capture websites and extract, or scrape, information from them. The best known example (and the one that we ultimately experimented with) is <u>Nutch</u>. Unlike harvesting, crawling does not require knowing more than a site's basic URL (or a preferred, honed set of URLs) in order to capture information. However, crawling, in contrast to harvesting, can only retrieve the basic level of metadata typically expressed on web pages, essentially titles and keywords, and even those will have varying levels of quality, consistency and availability.

Because of the above differences between harvesting and crawling, the LT3C team decided that these two strategies were actually complementary and that instead of choosing between them for our scoping experiments, we should attempt to combine them in order to capitalize on the strengths of each. Before designing and engaging in such an exercise though, we felt it was necessary to review the requirements from LT1A and LT3A in order to understand the standard we were ultimately going to be measuring harvesting/crawling against. We were particularly concerned with identifying essential requirements as well as pulling out features that would be essentially impossible to meet with harvesting/crawling. We felt strongly that it was important to calibrate expectations of these strategies with what they could reasonably accomplish.

Refinement of Requirements

As alluded to above, LT3C members reviewed the requirements from Lightning Teams 1A and 3A in order to identify those that could be reasonably met by harvesting and crawling solutions. This culled or prioritized list served two purposes: 1) to provide a standard against which to assess the ultimate value of the effort identified in our scope of work estimate (i.e., for a likely outcome, is the expected amount of invested effort worth the anticipated results) and 2) to provide a target against which we could estimate the scope of work.

We identified those elements that could be addressed (though perhaps to a limited degree) as "Phase 1" requirements and have listed them below. The fully annotated set of requirements can be found in Appendix A.

"Phase 1" Prioritized Requirements for a Harvester/Crawler Based Solution

Search

- **Basic search**: Every page should include a single text box for simple keyword searches that may include single or multiple search terms. When a keyword search is submitted, the following fields will be searched: title, subject, description, contributor, date, format, rights.
- **Scope**: By default, searches should be conducted across all collections with the option of limiting to a specific collection.
- Multilingual search: Search should accommodate multiple languages. Unicode support.

Search Results

• Item level information: Each item in a result set should be accompanied by the following primary metadata: title, subject, description, contributor, date, format, rights.

- **Facets**: Facets should serve to refine or expand search results and should be made available for the following primary metadata: title, subject, description, contributor, date, format, rights. Sorting: Default sorting of search results should be by relevance; users should have option to sort by additional sorting criteria: collection, author, title, date.
- **Pagination**: Result sets should be paginated with users able to navigate back / forth through pages of results.

Object View

(NOTE: it is probable that a crawler/harvester solution will always direct users to the original home/host site in order to interact deeply with a specific piece of content).

- **Context**: Objects should be displayed in a view that provides UC Libraries Digital Collection, UC campus, and potentially collection-branding.
- **Thumbnails**: Images should be represented by thumbnails that when clicked open to a full view of the image within an image viewer.
- **Object level citation**: All objects should have an object-level citation. A "Citation" link or icon should be available that when clicked will display citation information.
- **Social media**: A link or icon should be available that when clicked will allow the user to send objects to social media targets (e.g., Facebook, Delicious, Pinterest).

Attribution

- UC Libraries: The UC Libraries attribution/brand should always be present; all pages should have a branding area at the top that will include at minimum the UC Libraries brand.
- UC campus: UC campus attribution/branding should be present on all pages associated with that campus.
- **Contributing institution**: Objects contributed by or associated with a given entity will be identified on the object level page in the area containing associated primary metadata.

Feedback / Communication / Inquiries

(NOTE--a significant amount of support infrastructure is implied here, especially since many of these requests are likely to be for the content owners, not for the UCLDC system itself. For 3C purposes, an email address may be sufficient, as a placeholder until this larger structure is established.)

• Help / feedback: A link or icon should be available from all pages that when clicked provides a feedback form for submitting comments and questions to the UC Libraries Digital Collection staff.

Contributor / Collection Information

• **Contributing institutions (needs to be driven by a registry)**: A full alphabetical list of contributing institutions should be made available, with each entry linked to a customized landing page including full contact information. The right to perform administrative

activities relative to the landing page (e.g., institution contact information) should be granted to the contributing institution.

- Collection description (needs to be driven by a registry): A document describing the collections included in the UCL Digital Collection should be available on the site. User guides (needs to be human generated): A document describing how to use the features the UCL Digital Collection should be available on the site.
- **Contributor guide (needs to be human generated)**: A document providing guidance for how to contribute to the UCL Digital Collection should be available on the site. **Technical documentation (needs to be human generated)**: A high level description of the components driving the UCL Digital Collection should be available on the site.

General

• **Identifiers** Each object should have a unique, permanent identifier. **Search engines**: Content should be optimized for and discoverable via search engines.

Scoping Experiment

All harvesting, crawling and indexing experimentation was conducted on a machine set up in the Amazon Web Services (AWS) cloud. This resource allowed us to quickly have a low-cost, shared space available to all team members, across institutions. It also provided the only feasible platform from which to launch and conduct the large-scale processing jobs required for web crawling.

Harvesting

LT3C identified an easy to use harvesting tool, JOAI and then, using the Appendix B, Collections Ready for Surfacing created by POT1 LT3A, identified collections that could be readily harvested through an OAI-PMH interface. This was more challenging than expected as few sites have a documented OAI-PMH interface. Without directly contacting collections owners, we were able to identify 13 collections, with at least one for each campus, and attempted to harvest those. Some errors were encountered, from problems in collections to bugs in the harvesting tool. Ultimately, 12 collections were harvested resulting in 23,065 records. Gathering information from collection owners about availability of protocols such as OAI-PMH initially would make the harvesting process much more efficient.

An additional challenge to the initial use of JOAI was the discovery of a bug that generated invalid harvested records for some sources. After contacting the JOAI staff, LT3C team member Garey Mills was able to develop a patch that not only fixed the bug for the Lightning Team's purposes, but that ultimately got included in a new distribution (version 1.1.3)of JOAI.

Two of the four collections whose OAI interface is hosted by the Internet Archive had problems that we could not resolve. The contact information given by the interface (via the OAI Identify command) named the Internet Archive as the technical contact, and email sent to them was never answered or, apparently, acted upon. We would suggest that the lack of response was due to the fact that the errors were in the formatting of the information served, not in the OAI repository software, and that the Internet Archive was not able to fix the data, being at one remove from the

problem. There should either be a data contact as well as a technical contact, or institutions should not outsource this service.

The metadata harvested is in Dublin Core format, called in OAI parlance 'oai_dc'. OAI-PHM is flexible enough so that it can generate other metadata formats with a minimum amount of configuration, which means that it could conceivably serve as part of a solution that specified a different, perhaps more complete, metadata format. That would require that the participating institutions make that new metadata format available, which would be difficult, but OAI could harvest 'oai_dc' from some collections and the new, more complete, metadata format from others. In this way that OAI could provide part of the graduated solution that we are proposing: crawling some collections, collecting 'oai_dc' from others, and exhaustive metadata from those collections capable of providing it. However, a serious caveat is required here: while this approach might reduce the amount of effort for collection owners, the ingest process side of the equation would become increasingly more complicated as the system would have to accommodate multiple metadata formats.

After content was harvested, it was indexed in Solr. Figure 1 is a screenshot of the main JOAI page for managing the harvesting and Figure 2 is a screenshot of the resulting index. Basic metadata such as title, creator and URL displayed easily using the default Solr schema. The absence of other basic fields, such as date and keywords indicates the variability of harvested content, even in the presence of a standard such as OAI-PMH.

Figure 1: JOAI Harvesting Administration Page

🥹 jOAT Ha	rvester Setup and Sta	tus - Mozilla Firefox				- B 🛛
<u>File E</u> dit	<u>View History Bookma</u>	ks <u>T</u> ools <u>H</u> elp				
€ ⇒	🛞 ec2-50-18-28-83.us	west-1.compute. amazon a	aws.com:8080/oai/admin/	harvester.do	🏠 🔻 😋 🔀 🕆 calistoga spa inn	🔎 🎓 🤻 🔻
🔎 Most Visi	ed 🔞! Yahoo! Mail: The be	st 🚹 Welcome to Fa	cebook 🗍 eschol-dev	.escholarsh 🗍 SUBI: Available b	branc 🦲 Dashboard - Pivotal Tr 🍥 eScholarship: Universi 📈 voo2do - Your Dashbo.	🗌 My Journals 🗌 Log In 💦 »
⊘ Disable+	👗 Cookies+ 🗡 CSS+ 🚺	Forms+ 🕎 Images+	〕 Information+ 📃 Misc	ellaneous+ 🥖 Outline+ 🥢 Resize	er 💥 Tools- 🔳 View Source- 🔟 Options-	🖸 🖸 🖌
🞑 (187 unr	aad) - sflschiff - Yahoo! Mail	× 🔀 Circles - G	oogle+	× 🛃 Facebook	× 👩 /urls/ - public-digital-collection - CDL's DP × 🛃 jOAI Harve	ster Setup and Status × +
jOA	HARVE	STER	Õ	6		<u>^</u>
THE JAVA-	BASED OPEN ARCHIVE	INITIATIVE DATA P	ROVIDER & HARVEST	ER		
Overview	Search	Explore Da	ata Provider 🔻 Harve	ster 🔹 🔻 Documentation 🔻		

Harvester Setup and Status

Setup

Add a harvest to get metadata XML files from OAI data providers. If only providing metadata files, harvester set up is not necessary.

Add new harvest Create a harvest to get files and to specify when and where the harvest is performed.

Status

View harvest history and progress View a listing of all past harvests performed, current harvests in progress and their details.

Harvest Repository	Metadata Format 😯	SetSpec 🛛	Harvest Interval	Manually Harvest 🕑	Harvest Settings ?
California Agricultural Experiment Station Publications Base URL: http://archive.org/services/oai2.php Harvested to:/usr/share/tomcat7/webapps /oai/WEB-INF/harvested_records Last harvest: 1824 files, Sun Oct 21 01:30:00 UTC 2012 View harvest history and progress	oai_dc	collection:californiaagriculturalexperimentstationpublications	Automatic (Every 2 days at 1:30 AM UTC)	New All	Edit Delete
UCD Bulletin of Calif. Dept. of Water Resources Base URL: http://archive.org/services/oai2.php Harvested to: /usr/share/tomcat7/webapps /oai/WEB-INF/harvested_records View harvest history and progress	oai_dc	collection:cawaterres	Automatic (Every 2 days at 1:20 AM UTC)	New All	Edit Delete
UCD Bulletin of Calif. division of Mines and Geology Base URL: http://archive.org/services/oai2.php Harvested to: /usr/share/torncat7/webapps /oai/WEB-INF/harvested_records Last hopsect 557 kites_200_oct 31.011/00.01.010	oai_dc	collection:caminesgeo	Automatic (Every 2 days at 1:10 AM UTC)	New All	Edit Delete
ıd: Fre 🦊 Next 👚 Previous 🖌 Highli	ight <u>a</u> ll 📃 Mat <u>c</u> h case	() Phrase not found			

POT1 LT3C Final Report 9



Figure 2: Solr index of Harvested Content

Crawling

The crawling experiment was designed to collect metadata for collections not accessible through an OAI-PMH interface. The first step was to develop a seed list of URLs. To create this, we extracted all of the links from the LT 3B Surfacing list, eliminating those that did not lead to clear collection pages (e.g. <u>http://archive.org</u>) as links of this type require more manual analysis to create specific crawling rules. From approximately 112 collections, we were able to identify 79 seed URLs, some of which pointed to entry pages for multiple collections.

Crawling large numbers of websites involves requesting a tremendous number of pages, requiring in turn a high level of available resources if the crawl is to take place in a reasonable amount of time. LT3C used the widely used <u>Nutch</u> crawler from Apache, which recommends running the crawler on <u>Hadoop</u>, a distributed platform. Setting up the Hadoop cluster and Nutch

involves climbing a significant learning curve. Fortunately, Brian Tingle, one of the LT3C members, had previous experience with deploying these tools, which we were able to leverage. The crawl itself took approximately four or five hours and resulted in just over 35,000 web pages ready for indexing.

Once the crawl was completed, the content was indexed, again in Solr. The screenshots below show the very baseline results. Note that these results do not include any work to try to improve the records as they are indexed in Solr (for instance, ensure that the values in <title/> elements get indexed and displayed as titles). That type of improvement would be an obvious next step. We decided not to do this, because our ultimate goal was to combine both the harvested and crawled content, so it made sense to work on display enhancement at that next stage.

🕑 Solritas - Mozilla Firefox					
Ele Edit View Higtory Bookmarks Tools Help					
🗧 🔶 🕐 ec2-50-18-28-83.us-west-1.compute.amazonaws.com:8081/solr/browse			☆ マ C Google	<u> </u>	₩ •
💭 Solritas	+				
Apache Solr Examples: Simple <u>Spatial Group F</u>	<u>By</u>			<u>Solr</u>	Admin 🌥
F	ind:		Submit Query Reset		
	□B	post by Price			≣
Query Facets		35033 results found in 2 ms Page 1 of 35	04		
Range Facets	[Historical Overview] More Like This More Info: http://animation.library.ucla.edu/historicalEssay.html				
Pivot Facets	ORE. #dil	âcir(h)			
Clusters Run Solr with java -Dsolr.clustering.enabled=true -jar start.jar to see results	Image: The Project More Like This More Info: http://animation.library.ucla.edu/AboutOverview.html URL: \$url.get(0) s				
	This This More Info URL: \$url	net Archive: Digital Library of Free Books, Movies, Music & Wayback Machine] More Like http://archive.org/ get(0)			
	More Info URL: \$url	net Archive Frequently Asked Questions) More Like This http://archive.org/about/faqs.php get(D)			
× Find: entity	🕨 <u>N</u> ext 🔒 Pre	vious 🖌 Highlight all 🔲 Match case 🛛 🎲 Reached end of page, continued from top			
х					3

Figure 3: Solr Index of Crawled Content

🥹 Solritas - Mozilla Firefox					
<u>File Edit Vi</u> ew Hi <u>s</u> tory <u>B</u> ookmarl	s <u>T</u> ools <u>H</u> elp				
♦ ♦ @ ec2-50-18-28-83.us-	est-1.compute. amazonaws.com :8081/solr/collection1/browse?q=gold+rush	🏫 ▼ 😋 🚼 - Google	🔎 🎓 🧚 🔻		
Solritas	+				
Apache Solr Examples: Simple <u>Spatial Group I</u>	Υ		<u>Solr Admin</u> ▲		
F	ind: gold rush	Submit Query Reset			
	□Boost by Price				
Query Facets	215 results found in 1	18 ms Page 1 of 22			
Range Facets	Days of Cal Blue & Gold, 1895] More Like This More Info: http://bancroft.berkeley.edu/CalHistory/yearbook.1895.html				
Pivot Facets	OKL. Juniger(0)				
Clusters	[H. Works Gold Rush letter, 1855.] More Like This				
Run Solr with java -Dsolr.clustering.enabled=true -jar start.jar to see results	More Info: http://www.oac.cdlib.org/search?style=oac4;titlesAZ=h;idT=001299816 URL: \$url.get(0)				
	C Room Six: Gold Rush Women] More Like This More Info: http://bancroft.berkeley.edu/Exhibits/Goldrush/room_06.html URL: \$url.get(0)				
	[Room Six: Gold Rush Women] More Like This More Info: http://bancroft.berkeley.edu/Exhibits/Goldrush/roomD6_penknife_lg.html URL: \$url.get(0)				
× Find: entity ×	• Next 👚 Previous 🖉 Highlight all 🗌 Match case 🔅 Reached end of page, continued from top		⊻ ⊛		

Figure 4: Crawl Only Search Results



Figure 5: Example Object from Search Result

Harvesting Plus Crawling

Harvest and crawling represent different, yet potentially complementary, approaches to pulling in metadata records and digital object references from disparate sources. Harvesting promises a richer set of metadata, while crawling supports a wider scope of collected materials. The primary challenges to combining these strategies are the differences in the structure of the records and the variability of the robustness of the records. As described previously, harvested content will include at least a majority of the <u>Dublin Core</u> metadata fields (e.g. Title, Creator, Date, Subject, Description, etc.), while crawled content will contain variable levels of title and keyword information.

After gathering metadata records via both strategies, a Solr indexing schema was modified to allow for the creation of a single index of both records from both sources, resulting in a system with 57,663 records available for searching and browsing. Figure 6 below shows the first page of that combined system. Note that the facets are those available through the harvested metadata only, and do not include records gathered from the crawl.

🥙 Solritas - Mozilla Firefox			- 2 2	K
Eile Edit View Higtory Bookmark	s <u>T</u> ools <u>H</u> e	eb		
♦ ♦ @ ec2-50-18-28-83.us-w	/est-1.compute	e. amazonaws.com:8081/solr/collection2/browse?8q=	🔎 🏦 🤻	٣
[]] Solritas	+			
Apache Solr Evamples: Simple Spatial Group E	10		<u>Solr Admin</u> ≜	
	4			
F	ind:	Submit Query Reset		
		Boost by Price		
Field Facets		57663 results found in 10 ms Page 1 of 5767		
creater co	[Hist	orical Overview] More Like This		
Cancian, Frank (3083)	More Info	o: http://animation.library.ucla.edu/historicalEssay.html		
DeCou, Branson (American,				
(2471)				
<u>Scheffler, Harold W., 1932-,</u> <u>Photographer</u> (1370)	Abo [Abo	ut the Project) More Like This		
Rorty, Richard (975)	More Info	io: http://animation.library.ucla.edu/AboutOverview.html		
Watson, Raymond L., 1926- (406)				
Poster, Mark (230) California, Division of Mines				
and Geology (167)	This	rnet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine] More Like		
<u>Everett, Hugh</u> (152) California Agricultural	More Info	o: http://archive.org/		
Experiment Station (92) Biolotti, Erodorio T. (Erodorio				
<u>Theodore), 1865-1939</u> (61)				
<u>California State Mining</u> <u>Bureau</u> (59)	[Inter	rnet Archive Frequently Asked Questions] More Like This		
California. Division of Mines	More Info	o: http://archive.org/about/faqs.php		
(33) <u>Vogt, G. A., photographer</u> (51)				
<u>Cruess, W. V. (William Vere),</u> <u>1886-1968</u> (45)	[] [Inter	rnet Archive: Contact] More Like This		
Putnam & Valentine (American, active 1880s-	More Info	o: http://archive.org/about/contact.php		
x				2

Figure 6: Solr Index of both Harvested and Crawled Metadata Records

As is, the combined index includes both the ability to use facets with the harvested content and searching across all content. Figure 7 is a screenshot of the results for searching for the phrase "gold rush" and Figure 8 is the display of one of those items.

🕑 Solritas - Mozilla Firefox			- 2 2	K	
Elle Edit View History Bookmarks Tools Help					
♦ ♦ ec2-50-18-28-83.us-w	est-1.compute.amazonaws.com:8081/solr/collection2/browse?q="gold+rush"&fq=content%3Awater	☆ マ C Soogle	🔎 🏦 🤻 י	*	
🔅 Solritas	+				
Apache			3	^	
Examples: Simple Spatial Group B	(
F	ind: "gold rush"	Submit Query Reset			
> content:wate	□Boost by Price #				
Field Facets	76 results found in 93 ms Page 1 of 8				
Query Facets	IRelated Links] More Like This More Info: http://bancroft.berkeley.edu/reference/links.html				
<u>content:preservation</u> (38) <u>content:water</u> (76)					
Range Facets	Range Facets Western Americana] More Like This More Info: http://bancroft.berkeley.edu/collections/westernamericana.html				
Pivot Facets					
Clusters	Special Collections, UC Santa Barbara, Online Archive of California] More Like This More Info: http://www.oac.cdlib.org/institutions/UC+Santa+Barbara::Special+Collections				
Run Solr with java -Dsolr.clustering.enabled=true -jar start.jar to see results					
	Browse Collections (E). Online Archive of California] More Like This				
	More Info: http://www.oac.cdlib.org/titles/e.html				
	CHAPTER IJ More Like This				
	More Info: http://websites.lib.ucr.edu/agnic/webber/Vol1/Chapter1.htm				
× Find: gold ru	Next 👚 Previous 🖌 Highlight all 🔲 Match case			V	
х			6		

Figure 7: Searching Across All Content

Figure 8: Object View of an Item from a Result Set

🕘 CHAP	'ER I - Mozilla Firefox		
<u>File</u> <u>E</u> dit	<u>Vi</u> ew Hi <u>s</u> tory <u>B</u> ookmarks <u>T</u> ools <u>H</u> elp		
÷)	🛞 websites.lib.ucr.edu/agnic/webber/Vol1/Chapter1.htm	☆ マ C 🚼 マ Google 🔎 🎓 🤫	≋ ▼
	+ 19		

Some of the trees in this old orchard lived on for many years, but the last of them died of age and lack of care in the early years of the twentieth century.

Early California Orange Groves.—While there are accounts of fruit trees such as pears, peaches, nectarines, apples, and pomegranates being cultivated on lands around Los Angeles, citrus orchards apparently did not become established outside the missions until after secularization in 1833. The mission fathers seem to have prized citrus highly. In 1877, Jose del Carmen Lugo (1950) of the San Bernardino rancho recalled that owners of fields could not obtain seeds of oranges and lemons from the missions, because the padres "refused to allow these fruits to be raised elsewhere than at their missions." In 1834, Jean Louis Vignes, a Frenchman, procured from Mission San Gabriel thirty-five large seedling sweet orange trees which he transplanted to his place on Aliso Street in Los Angeles, and this

number was gradually increased until he was the possessor of a sizable grove. This was the second orange orchard to be planted in California (Spalding, 1885, p. 7).

According to Lelong (1902, p. 17), in the same year, 1834, a small planting was also made in Los Angeles by Manuel Requena.¹⁶ By this time the planting of a few citrus trees in gardens and courts for home use had become a common practice, but no groves had been planted to provide fruit for sale.

Next came the famous planting of William Wolfskill, a Kentucky trapper of German descent who came overland to Los Angeles in 1831. His first trees, seeding sweet oranges, were obtained from Mission San Gabriel and were set out in 1841 on a small tract on the site occupied later by the old Arcade passenger station of the Southern Pacific Railway. Spalding (1885, p. 8) said that the orchard of William Wolfskill was no doubt the first to be planted in California with an eye to profit. Since fruit from the Vignes grove was grown for home use and distribution among friends, Wolfskill's biographer, Iris Higbee Wilson, considers her subject to have been the founder of California's citrus industry (Wilson, 1965, p. 13).

Wolfskill's neighbors ridiculed him and his idea of growing oranges for sale, but with German tenacity and patience he maintained his trees and gradually extended his plantings. It was a hard struggle in the early years, but contrary to expectation the project paid, and the grove was gradually increased in size to twenty-eight acres and finally to approximately seventy acres. In later life Wolfskill reaped a rich reward. The last crop to be disposed of in his lifetime, from about twenty-eight acres of his grove, sold on the trees for \$25,000.

At about the same time, in the early 1840's, the planting was begun of a grove that later became famous as the Don Benito Wilson grove, situated several miles north of Mission San Gabriel and at one time a part of the mission property.

While citrus culture was very slowly extending, the great advances being made in the development of the United States were paving the way for the introduction of the orange into commerce. After the secularization of the missions in the 1830's, the early fruit industry began to decline, and when General Frémont visited California in 1846 he wrote that "little remains of the orchards that were kept in fign cultivation at the Missions." Some of the early settlers who had foresight enough secured certain of the mission orchards and maintained them and were thus enabled later to reap a rich reward. Influence of Americanization and "Gold Rush."—After the planting of the Mission San Gabriel grove, the next great stimulus to the citrus industry came with the ceding of California to the United States. This transaction followed the Mexican War (1846) are the conquest of California by the United States, and was ratified by treaty in 1848. Almost immediately came the discovery of gold in the new territory, followed by the great "gold rush" of 1849, which swelled the population of California. The fruit could be shipped by ocean freight to San Francisco and thence by water up the Sacramento, American, and Feather rivers to points near the mines. San Francisco became the great market for the industry and remained so for a period of three decades.

Apparently, however, there was no great haste to take advantage of this opportunity. It required too long to produce a bearing grove and the persons who were living in California were not commercially minded. The people coming in by the thousands were seeking wealth through the metal gold and not through the golden fruits. The reaction to the new conditions was therefore slow, but by providing potential market the stimulation that was needed to expand the industry was created.

In 1852, Don Benito Wilson purchased the small orchard that thereafter was known as the "Don Benito Wilson grove" and began to extend the plantings. In 1853, Mathew Keller began the planting of an orchard opposite the Wolfskill grove with seedlings grown from fruit obtained from Central America and Hawaii

In 1857, a small orchard was planted by L. Van Luven at Old San Bernardino with seedlings grown by himself and with other seedlings from Los Angeles. In 1869, L. F. Cram set out a small grove of some two hundred seedling trees near Highland (Brown and Boyd, 1922, vol. 1, p. 73). These last two orchards were important as marking the extension of experimental culture into the interior valleys, unless one accepts the curiously meager testimony of Evans (1874) that orange trees were grown at Old San Bernardino in the mission period.

The Cram orchard remained for many years a show place, but gradually the trees became badly infested with scaly bark and were largely dug out in 1928. One of the oldest trees in this planting remained until after World War II, but gradually deteriorated thereafter and was finally removed in 1961. As far as can be learned, it was the oldest and largest citrus tree in California when removed. Careful measurements of this tree made by the senior author in December, 1926, gave the following dimensions: height, 33 feet 5 inches, circumference, one foot above the ground, 59 1/2 inches, spread of branches, 30 feet (fig. 1-11). When felled in 1961, the tree measured 39 feet in height. The stump remaining in 1965 measured 25 inches in diameter two inches above ground level.

In 1865, Myron H. Crafts set out a small orchard of some two hundred trees at Crafton. By that time it may be said that the citrus industry in the interior valleys had become reasonably well established. Four years later, F. A. Kimball planted a small orchard of orange, lemon, and lime trees at National City in San Diego County, apparently the first grove planting in that section of the state. According to Kimball (1897), there were at that time a few isolated trees near San Diego and in the San Luis Rey Valley.

In 1867, according to the United States Department of Agriculture, there were in California 17,000 orange trees and 3,700 lemon trees, of which 15,000 orange trees and 2,300 lemon trees were in the Los Angeles region.

Introduction of the Washington Navel Orange Stimulated Citrus Planting.—Riverside, a community that became famous in the early history of citrus culture because of the Washington navel orange,

Effort Estimate

Because the discrete harvesting and crawling tasks were in and of themselves steps in producing the combined final set of indexed records, the effort involved in each is considered as part of the overall effort required to create the final system. Note that the calculations below reflect the amount of effort expended just for LT3C's experimentation, and, as discussed later, is substantially less than what would be required for a production system.

Task	Expended Person Days
System provisioning	3
Setting up JOAI harvester	2
Identifying OAI-PMH URLs	1
Harvesting/reharvesting	1
Indexing harvested content in Solr	1
Setting up the crawler: Nutch + Hadoop	3
(Note: Brian Tingle had undertaken significant	
advance work in this area; otherwise the time	
frame would have been much longer.)	
Crawling	1
Indexing crawled content in Solr	1
Combining crawled and harvested records	2
Refining display (e.g. titles, adding in campus	2
affiliation)	
Total	19

Table 1: Effort Expended for the LT3C Experiment

The LT3C team spent a total of almost three person weeks on this relatively restricted experiment. Not accounted for in this calculation was time spent by other people on tasks that we were able to leverage, for instance the assemblage of the list of collections by LT3A or the amount of time spent on a previous projects learning how to use Nutch and Hadoop to conduct large scale crawling. Those two tasks alone account for at least four person weeks of work. Also not included here is provisioning of a production service environment (which includes the provisioning of development and staging environments), QA/testing, and web interface refinement all of which would also add at least three person-months, if not more, for a very basic service. A very gross time requirements estimate then is that a basic harvesting/crawling system, along the lines of what LT3C experimented with, would take approximately six person months to get up and running. Additional time would have to be factored in for continued maintenance (adding in new sources of content, bug fixing, etc.) and customer support.

Results

The goal of LT3C's experimentation was to determine to what degree essential UCLDC discovery and display requirements could be relatively quickly met by leveraging harvesting and crawling technologies. Table 2 below captures the team's assessment of what we learned.

Table 2: Comparison of Prioritized Requirements and Harvesting/Crawling Scoping Experiment Results

Prioritized Requirement	Degree of Achievement
Search	
Basic search : Every page should include a single text box for simple keyword searches that may include single or multiple search terms. When a keyword search is submitted, the following fields will be searched: title, subject, description, contributor, date, format, rights.	Complete. LT3C's essentially "out of the box" Solr implementation includes type of support for basic search. Additional fields can be added in to the simple keyword search.
Scope : By default, searches should be conducted across all collections with the option of limiting to a specific collection.	Incomplete. Collection level metadata needs to be supplied by collection owners and added into the system. Once that is done (for example via the Registry described below), it could be relatively easily added into the basic interface.
Multilingual search : Search should accommodate multiple languages. Unicode support.	Complete. Solr supports UTF-8 by default.
Search Results	
Item level information : Each item in a result set should be accompanied by the following primary metadata: title, subject, description, contributor, date, format, rights.	Partial. Much of this information could be available in harvested sites, but will not be available in crawled sites, since only titles and keywords can be extracted.
Facets : Facets should serve to refine or expand search results and should be made available for the following primary metadata: title, subject, description, contributor, date, format, rights.	Partial. Facets are easily created in Solr, so could be created for those fields captured by harvesting or through the use of the Registry. LT3C notes that some of the proposed fields do not make sense for facets, specifically title and description
Sorting: Default sorting of search results should be by relevance; users should have option to sort by additional sorting criteria: collection, author, title, date.	Partial. Solr can be customized to provide sorting on any number or combination of fields. LT3C did not have sufficient time to add this functionality in.
Pagination : Result sets should be paginated with users able to navigate back / forth through pages of results.	Complete. Pagination is included with Solr.

Object View	LT3C believes that a crawler/harvester solution will almost always direct users to the original contributing site for interactions with a specific piece of content, therefore this section does not apply to a crawler/harvester based solution.
Context : Objects should be displayed in a view that provides UC Libraries Digital Collection, UC campus, and potentially collection-branding.	Incomplete (see explanation above).
Thumbnails : Images should be represented by thumbnails that when clicked open to a full view of the image within an image viewer.	Incomplete (see explanation above).
Object level citation : All objects should have an object-level citation. A "Citation" link or icon should be available that when clicked will display citation information.	Incomplete (see explanation above).
Social media : A link or icon should be available that when clicked will allow the user to send objects to social media targets (e.g., Facebook, Delicious, Pinterest).	Incomplete (see explanation above).
Attribution	
UC Libraries: The UC Libraries attribution/brand should always be present; all pages should have a branding area at the top that will include at minimum the UC Libraries brand.	Partial. Consistent UC Libraries (e.g. UC Library Digital Collection) branding can be easily be added to a basic Solr instance.
UC campus : UC campus attribution/branding should be present on all pages associated with that campus.	Incomplete. Campus level branding that appeared with associated content would require the use of the Registry described later in this report.
Contributing institution : Objects contributed by or associated with a given entity will be identified on the object level page in the area containing associated primary metadata.	Incomplete. Contributing institution level branding that appeared with associated content would require the use of the Registry described later in this report.

Feedback / Communication / Inquiries	
Help / feedback: A link or icon should be available from all pages that when clicked provides a feedback form for submitting comments and questions to the UC Libraries Digital Collection staff.	Partial. A significant amount of support infrastructure is implied here, especially since many of these requests are likely to be for the content owners, not for the UCLDC system itself. For LT3C purposes, an email address may be sufficient, as a placeholder until this larger structure is established.
Contributor / Collection Information	
Contributing institutions : A full alphabetical list of contributing institutions should be made available, with each entry linked to a customized landing page including full contact information. The right to perform administrative activities relative to the landing page (e.g., institution contact information) should be granted to the contributing institution.	Incomplete. A listing of contributing institutions and landing pages for each is dependent upon a component such as the proposed Registry. Granting of various levels of permissions could be built out in conjunction with a Registry implementation.
Collection description: A document describing the collections included in the UCL Digital Collection should be available on the site.	Incomplete. Descriptions of collections can only be provided through the use of a system like the proposed Registry, which would gather those descriptions as part of initial establishment of the collection's record in the UCLDC.
Documentation	
User guides: A document describing how to use the features the UCL Digital Collection should be available on the site.	Incomplete. User guides must be manually created, but once developed, could be easily linked to a harvester/crawler based system.
Contributor guide: A document providing guidance for how to contribute to the UCL Digital Collection should be available on the site.	Incomplete. Contributor guides must be manually created, but once developed, could be easily linked to a harvester/crawler based system.
Technical documentation: A high level description of the components driving the UCL Digital Collection should be available on the site.	Incomplete. Technical documentation must be manually created, but once developed, could be easily linked to a harvester/crawler based system.
General	

Identifiers: Each object should have a unique,	Incomplete. Each object indexed in Solr has a
permanent identifier.	unique identifier, but this is not likely to be the
	identifier scheme preferred for the UCLDC (e.g.
	ARKs or handles or DOIs). Part of the ingest
	process for harvested and crawled content
	would have to include pre-processing to assign
	an identifier from the scheme of choice.
Search engines: Content should be optimized	Partial. Solr generated pages are easily
for and discoverable via search engines.	discoverable by Google, but since in a
	crawler/harvester based solution object pages
	would be on local sites, it is not clear how much
	search engine discovery would be improved.

Task 2: Develop an appropriate metadata schema that will accommodate discovery and display needs and SEO goals

The second task assigned to LT3C was to investigate a metadata schema that would accommodate the specified discovery and display requirements and that would also adequately achieve search engine optimization (SEO) goals. LT3C had two responses to this.

First, we feel that the metadata schema is really a secondary issue. What is most important is ensuring that there is quality metadata that is assigned in as uniform way as possible to ensure the most meaningful experience for users as they work across collections within this single system. A variety of metadata schemas exist, any number of which would be viable options for representing a given record on a website.

Second, the system that we built as a scoping exercise did not really lend itself to exploring various metadata schemas, because we knew from the outset that the majority of the content we were working with--crawled content--would have exceptionally limited metadata, essentially titles, keywords, and URLs at best. Any work we did evaluating schemas would have been based on a set of data far removed from the quality of data that will eventually be collected in the UCLDC DAMS.

Third, and following on the above point, we felt that the schema used to express metadata in the UCLDC generated web pages would be best assessed in the context of whatever system(s) are ultimately chosen to support the DAMS and the associated access interface. Because multiple schema choices are available, it is best to consider them in the context of what can be most efficiently and reliably produced in conjunction with the other components of the system.

Task 3: Determine and test for feasibility with collection owners/managers a proposed workflow to notify UCLDC about services/content for inclusion in UCLDC

Early on, LT3C determined that some essential metadata (e.g. campus, collection name, description) could only come from collection owners, so a mechanism was needed for them to easily provide that information. This requirement overlapped significantly with LT3C's third assigned task, identifying and vetting a workflow for collection owners to notify the UCLDC about content for possible inclusion in the system. Related to this latter requirement, we realized that collections may be in differing states of readiness for taking advantage of the various services to be embedded in the UCLDC. For instance, materials in a collection might be ready to be ingested into a DAMS, but still might require metadata work before being ready for public access.

Our proposal for addressing these various needs is the creation of a Registry, which would be a component of the UCLDC that would allow collection owners to begin to participate in the UCLDC regardless of the development stage of the collection itself. The UCLDC Registry would be a straightforward tool that would allow collection owners to:

- Create and maintain a record describing each collection, thereby making the UCLDC aware of it and providing collection-level metadata.
- Indicate which UCLDC services the collection requires, from DAMS submission, to metadata editing, to access. That information could also be updated as required, for instance when access decisions change.

Reviewing the ideas from LT3A, focusing in particular on the "System functionality" subsection of the "Continual Discovery of Content" section (see <u>POT1 LT3A final report</u>, pages 18-19) revealed a high degree of similarity with that team's first three suggestions, particularly #3 which specifically called out the creation of a Registry. Items 1 and 2, which discuss harvesting and transfer of content from campuses and hosted systems to the UCLDC, are greatly facilitated by such a service. In addition, the Registry is a fundamental component of the overall model that LT1C is investigating, which provides further confirmation, from a system perspective at least, that this is a positive strategy. (See Appendix B for screenshots of the Registry concept.)

UCLDC Registry Concept Survey

LT3C created a conceptual prototype of the Registry, and included screenshots of it in a survey that was sent to potential users (see Appendix B for a reproduction of the survey). 16 individuals were sent a message asking them to complete a brief survey about the concept, and inviting them to send the survey message and link to others who would be appropriate respondents. 12 people participated, each of them answering all three questions. Overall, as detailed in the analysis below, respondents considered the Registry to be a positive and reasonable approach.

Responses to Questions

Questions 1 and 2 probed respondents on essentially the same question--would they actually use a Registry--but from two different vantage points.

Survey Question 1

my collections included in the UCLDC.				
Answer Options	Response Percent	Response Count		
Strongly Agree	33.3%	4		
Agree	50.0%	6		
Undecided	8.3%	1		
Disagree	0.0%	0		
Strongly Disagree	0.0%	0		
Comments (optional)	8.3%	1		
answ	rered question	12		
skij	oped question	0		

Survey Question 2

I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC.			
Answer Options	Response Percent	Response Count	
Strongly Agree	33.3%	4	
Agree	50.0%	6	
Undecided	8.3%	1	
Disagree	0.0%	0	
Strongly Disagree	0.0%	0	
Comments (optional)	8.3%	1	
answ	rered question	12	
skij	oped question	0	

The first question attempted to address how burdensome or, put more positively, how easy the idea of a Registry would be for content owners to use as a tool for getting their content into the UCLDC. 10 out of 12 people felt the Registry would be an easy mechanism, while one was unsure and another was concerned about the initial input. On this latter point, LT3C team members have envisioned that there will have to be initial support to assist with the mass uploading of existing collection information in order to ensure participation, an opinion supported by this respondent's comment. The results were exactly the same for Question 2, although in this question, the single comment referred to the need for more library-based discussion before being able to decide either way about participation. This respondent also said that the idea of the Registry was positive.

The third question attempted to uncover if there were preferable ways for collection owners to get their content into the UCLDC.

Survey Question 3

I would prefer to notify the UCLDC about my collection(s) by:			
Answer Options	Response Percent	Response Count	
Using a Registry similar to the examples previously	66.7%	8	
Talking to a librarian on my campus	0.0%	0	
Contacting the UCLDC directly through email, the	8.3%	1	
Other (please describe)	25.0%	3	
answ	ered question	12	
skip	oped question	0	

9 out of the 12 respondents (including one person who chose to express this in the "Other" comment area) indicated that the Registry was a good option. Two respondents wanted to use a batch upload process for getting content into the UCLDC, which is not a mutually exclusive option. If the UCLDC has a Registry, any collections that are included in it, regardless of the mechanism for that inclusion, will have a Registry record for management purposes and also for metadata that can only be gathered from the collection owner. The significance of these comments then is that a solid batch ingest process needs to be supported and that it must additionally populate the Registry to the greatest degree possible.

Only one person wanted to reach the UCLDC via another, more individualized means (in this case, through emailing the UCLDC), an avenue that will no doubt need to be generally provided through a customer support component of the UCLDC. Ultimately any email conversion would eventually develop into an ingest process, whether it be a batch, harvest or manual process, so this comment really reflects the need to support individualized support and consultation.

Summary and Findings

LT3C conducted a narrowly scoped experiment to determine how much content and essential discovery and display functionality could be supported with a minimal expenditure of effort. Approximately three person weeks of time resulted in a very basic discovery system for just under 58,000 content items; a more complete, production quality system is expected to take approximately six person-months to build and would include a substantially greater amount of content. This estimate does not account for the financial resources required (e.g. purchasing server space), nor does it include the effort required to build the proposed Registry, which LT3C sees as an essential component of even the earliest version of the UCLDC. Although in our estimation the experimental system received only a "Partial" or "Incomplete" rating for the majority of the prioritized requirements, many of them would be met more fully through the introduction of the Registry, which is likely to take at least an additional three person months to develop. The LT3C survey of content owners indicated that such a service would be well-used and would therefore be an effective means for not only beginning the process of including content in the UCLDC, but would provide a natural conduit for getting the collection level information required to augment a simple harvester/crawler based system.

Specific Findings

- The metadata schema appropriate for the UCLDC is best determined by the team designing and implementing the production system.
- OAI-PMH harvested content is not the same everywhere and is not "plug and play." Variations and problems have to be anticipated with every collection.
- Harvesting and crawling are insufficient on their own or in combination and need to be complemented by the provision of collection metadata provided by collection owners.
- Collection owners are comfortable with the concept of a Registry as a way to manage and include their collections in the UCLDC.
- While the Registry concept is viable, it will have to include support to assist with the mass uploading of existing collection information in order to ensure participation, an opinion supported by this respondent's comment.
- The Registry must work in concert with a solid batch ingest process, a process that must populate the Registry to the greatest degree possible.
- Individualized support and consultation will be required to help with initial interaction with the UCLDC, in addition to addressing any complications related to getting content included in it.

Recommendation

In order to maintain motivation for the fully -realized UCLDC and to provide improved discoverability of collections across the UC system, LT3C recommends developing an initial access system composed of:

- harvesting and crawling technologies for acquiring content metadata
- a Registry that supports the initial recording of collections to be included in the UCLDC along with the collection level metadata required for the display and management of the related content.

Because this initial system would be a component of the larger UCLDC, it should be developed in such a way as to facilitate its integration with whatever DAMS solution is ultimately adopted. This connection is critical, since even though the harvester/crawler based system described above would be a productive initial step towards the development of the complete UCLDC, it would not in any way fulfill the need for a systemwide DAMS designed to support robust stewardship of UC digital materials, which is one of the fundamental goals of the UCLDC.

Appendix A: Prioritized Discovery and Display Requirements

LT3C was charged with scoping out harvesting and crawler based solutions for a first iteration access (discovery and display) system to the UC Library Digital Collection (UCLDC) that could be made available in advance of the DAMS component. That scoping work is dependent upon two inputs from additional POT1 Lightning Teams: 1) the general UCLDC basic discovery and display requirements from LT1A and 2) the collections on the campuses that are currently ready and interested in greater exposure, as identified by LT 3A.

Because the LT1A requirements describes a relatively expansive set of features designed for a more built-out system, LT3C has prioritized a subset of that list that will be used for the scoping exercise, indicated by a "Phase 1" tag at the end of a given requirement listed below. Features to come at a later date are italicized.

Requirements

Search

1. Basic search: Every page should include a single text box for simple keyword searches that may include single or multiple search terms. When a keyword search is submitted, the following fields will be searched: title, subject, description, contributor, date, format, rights. **Phase 1**

2. Advanced search: All metadata fields exposed in the search results display should be available to be searched independently or in combination from an advanced search page. The exposed metadata fields will include all available fields in a given metadata schema.

3. Scope: By default searches should be conducted across all collections with the option of limiting to a specific collection. **Phase 1**

4. Spelling correction: Search term spelling correction should be provided.

5. *RSS: Users should have ability to subscribe to RSS feeds in lieu of stored queries.*

6. Multilingual search: Search should accommodate multiple languages. Unicode support. **Phase 1**

7. *Mobile devices: Content should be discoverable and displayable via mobile devices.*

Search Results

8. Item level information: Each item in a result set should be accompanied by the following primary metadata: title, subject, description, contributor, date, format, rights **Phase 1**

9. Facets: Facets should serve to refine or expand search results and should be made available for the following primary metadata: title, subject, description, contributor, date, format, rights **Phase 1**

10. Sorting: Default sorting of search results should be by relevance; users should have option to sort by additional sorting criteria: collection, author, title, date **Phase 1**

11. *Items per page: Users should be provided option to display pre-set items per page (e.g., 10, 15, 20)*

12. Pagination: Result sets should be paginated with users able to navigate back / forth through pages of results. **Phase 1**

Object View

(NOTE: it is probable that a crawler/harvester solution will always direct users to the original home/host site in order to interact deeply with a specific piece of content).

13. Context: Objects should be displayed in a view that provides UC Libraries Digital Collection, UC campus, and potentially collection-branding. **Phase 1**

14. *PDF display: PDFs should be displayed within the branded area and not in a separate Adobe Acrobat Reader window.* (*Note: users would be taken to the hosting site for this type of view*)

15. Thumbnails: Images should be represented by thumbnails that when clicked open to a full view of the image within an image viewer. **Phase 1**

16. Image viewer: Images should be easily optimized for viewing, including zoom in/out, rotate, mirror/flip, fit image, and full size. (Note: users would be taken to the hosting site for this type of view)

17. Search terms: Search terms should be highlighted in the object view, regardless of format. (Note: available once at hosting site only.)

18. "More like this": Items similar or related to the displayed object should be linked to from the object view page allowed users to view "more like this". (Note: available once at hosting site only.)

20. Object level citation: All objects should have an object-level citation. A "Citation" link or icon should be available that when clicked will display citation information. **Phase 1**

21. Download: A link or icon should be available on all object views that when clicked will allow the user to save the selected content. (Note: available per capacity at hosting site.)

22. *Print:* A link or icon should be available on all object views that when clicked will allow the user to print the selected content. ((Note: available per capacity at hosting site.)

23. Purchase: A link or icon should be available on all object views that when clicked will provide users with the contributing institution's contact information. ((Note: available per capacity at hosting site.)

24. Item / book bag: Users should be able to click a link associated with each object to add a citation and actionable URL to a session-based item / book bag page.

25. *Email item / book bag: Users should be able to email to themselves or others the objects saved to a session-based item / book bag page.*

26. Social media: A link or icon should be available that when clicked will allow the user to send objects to social media targets (e.g., Facebook, Delicious, Pinterest). **Phase 1**

Attribution

27. UC Libraries: The UC Libraries attribution/brand should always be present; all pages should have a branding area at the top that will include at minimum the UC Libraries brand. **Phase 1**

28. UC campus: UC campus attribution/branding should be present on all pages associated with that campus. **Phase 1**

29. Contributing institution: Objects contributed by or associated with a given entity will be identified on the object level page in the area containing associated primary metadata. **Phase 1**

Feedback / Communication / Inquiries

30. Help / feedback: A link or icon should be available from all pages that when clicked provides a feedback form for submitting comments and questions to the UC Libraries Digital Collection staff. Phase 1 (NOTE--a significant amount of support infrastructure is implied here, especially since many of these requests are likely to be for the content owners, not for the UCLDC system itself. For 3C purposes, an email address may be sufficient, as a placeholder until this larger structure is established.)

Contributor / Collection Information

31. Contributing institutions: A full alphabetical list of contributing institutions should be made available, with each entry linked to a customized landing page including full contact information. The right to perform administrative activities relative to the landing page (e.g., institution contact information) should be granted to the contributing institution. **Phase 1 (needs to be driven by a registry)**

32. Collection description: A document describing the collections included in the UCL Digital Collection should be available on the site. **Phase 1 (needs to be driven by a registry)**

33. User guides: A document describing how to use the features the UCL Digital Collection should be available on the site. **Phase 1**

34. Contributor guide: A document providing guidance for how to contribute to the UCL Digital Collection should be available on the site. **Phase 1**

35. Technical documentation: A high level description of the components driving the UCL Digital Collection should be available on the site. **Phase 1** (**33-35 are essential, but would have to be generated by a human being and added to any harvester/crawler populated system.**)

General

36. Each object should have a unique, permanent identifier. Phase 1

37 Search engines: Content should be optimized for and discoverable via search engines. **Phase 1 (NOTE: was #7 in "Search" section).**

Appendix B: Registry Concept Survey

The survey below was sent to 16 collection owners, who were invited to pass the survey URL along to any other appropriate individuals:

Screen 1: Overview



Screen 2: Registry Concept Screenshots

The five screenshots below are an illustration of the proposed UCLDC Registry concept. Please review them before answering the survey questions.

The second secon			
a did-dev.cdb.org/colection registry/UCR		Q.4	e 🔄 - coop
fost Visited Of Yahoo! Mail: The best 🛐 Welcome t	to Facebook 🗌 eschol-dev.escholarsh 🗍 SUBE: Av	valable branc 💿 eScholarship: Universi 🗌 My Journals 🏑	vos2do - Your Dashbo Dashboard - Pivotal Ir
Isable• 🛓 Cookies• 🧨 CS5• 🔯 Forms• 🖾 Image	s• 👔 Information• 📒 Miscellaneous• 🥖 Outline• 🥖	Resize+ 👷 Tools+ 🔳 View Source+ 🔝 Options+	
ttp://dsd-dev.odib.oollection_registry/UCR +			
Digital Collections UCB UCD	UCI UCLA UCM UCR UC	CSB UCSC UCSD UCSF	Sign In •
2. View existin collections.	ng campus	1. Login with your campus	this would link to a UC Trust WAYF page
		credentialsno new	University of California Libraries
3.0 G trytes need DAMS solution History of Subtropical fruits and nuts in	3. Add a new collection.	Avery E. Field collection	Citrus Experiment Station Centennial
talinated an entrance of the state of the state of the	same meeters below a	Participation of the second se	
California (1963)			Exhibition
California (1963) Citrus Fruits and their Culture in Wangtung Province, South China(circa-	George Fujimoto diaries	Mexican Revolution photographs	Exhibition Oral History Interviews
California (1963) Citrus Fruits and their Culture in Avangtung Province, South China(circa- 1936)	George Fujimoto diaries Oswald Jonas memorial collection	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR)	Exhibition Oral History Interviews Riverside, California photographs
California (1963) Citrus Fruits and their Culture in Kwangtung Province, South China(circa 1936) Sabino Osuna papers	George Fujimoto diaries Oswald Jonas memorial collection The B. Traven Collections at UC Riverside Libraries	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR) The Citrus Industry (1967-1989)	Exhibition Oral History Intensiews Riverside, California photographs The Highlander student newspaper
California (1963) Citrus Fruits and their Culture in Wangtung Province, South China(circa- 1936) Sabino Osuna papers	George Fujimoto diaries Oswald Jonas memorial collection The B. Traven Collections at UC Riverside Libraries	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR) The Citrus Industry (1967-1989) Tom's Rivera Archive	Exhibition Oral History Interviews Riverside, California photographs The Highlander student newspaper Tuskgee Airmen Archive
California (1963) Citrus Fruits and their Culture in Kwangtung Province, South China(circa 1936) Sabino Osuna papers UCR Libraries Author Series video webcasts	George Fujimoto diaries Oswald Jonas memorial collection The B. Traven Collections at UC Riverside Libraries	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR) The Citrus Industry (1967-1989) Tom s Rivera Archive University Archives photographs	Exhibition Oral History Interviews Riverside, California photographs The Highlander student newspaper Tuskgee Airmen Archive WRCA ContentDM
California (1963) Citrus Fruits and their Culture in Kwangtung Province, South China(circa- 1936) Sabino Osuna papers UCR Libraries Author Series video webcasts	George Fujimoto diaries Oswald Jonas memorial collection The B. Traven Collections at UC Riverside Libraries UCR faculty publications UCR faculty publications WRCA Lippincott Collection [Lee Collection & other published non-circulating materials]	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR) The Citrus Industry (1967-1989) Tom's Rivera Archive University Archives photographs WRCA Lippincott Collection [Los Angeles Aqueduct]	Exhibition Oral History Interviews Riverside, California photographs The Highlander student newspaper Tuskgee Airmen Archive WRCA ContentDM WRCA Web Archiving Service California Water Districts
California (1963) Citrus Fruits and their Culture in Kwangtung Province, South China(circa- 1936) Sabino Osuna papers UCR Libraries Author Series video webcasts	George Fujimoto diaries Oswald Jonas memorial collection The B. Traven Collections at UC Riverside Libraries UCR faculty publications UCR faculty publications WRCA Lippincott Collection [Lee Collection & other published non-circulating materials]	Mexican Revolution photographs Pre-1600 Single Leaf Registry (SLR) The Citrus Industry (1967-1989) Tom s Rivera Archive University Archives photographs WRCA Lippincott Collection [Los Angeles Aqueduct] WRCA Web Archiving Service Sacramento-San Joaquin Deta	Exhibition Oral History Interviews Riverside, California photographs The Highlander student newspaper Tuskgee Airmen Archive WRCA ContentDM WRCA Web Archiving Service California Water Districts WRCA eScholarship

Step 1: Login, View Collections, Add a New Collection

Add collection	Django site admin - Mozilla Firefox		
e Edit Yew P	figtory Bookanarks Ipols Help		
• - • • • •	d-dev. cdb.org /collection_registry/admin/d_collections/collection/add/	😭 🔻 😋 🚺 - Google	P 🕈 * ·
Most Visited Of Vi	ahoo! Mail: The best 🛐 Welcome to Pacebook 🗌 eschol-dev.escholarsh 🔲 SUBE: Available I	xanc 🔹 eScholarship: Universi 🛄 My Journals 😡 vos2do - Your Dashbo 💽 Dashboa	erd - Pivotal Tr 🗌 Log In 🛛 🗵
Disable- 👗 Cooki	es• 🧨 CSS+ 🔯 Forms+ 🔛 Images+ 😗 Information+ 🧱 Miscellaneous+ 🥖 Outline+ 🥒 Resize	- 🄀 Tools+ 🔳 View Source+ 🔝 Options+	~ 0 0
Add collection Dja	ngo ste admin +		
ojango adm	ninistration	Welcome, dsc.	Change password / Log out
ome > DI_collectio	ans > Collections + Add collection		
Add colled	tion		
Name:	1	Provide collection level	
Slug:			
Campus	Hold down "Control", or "Command" on a Max, to select more than one. Available comman 0	.	
campus.	Q Eller		
	Davis Drvine Drvine Riverside Riverside San Francisco Santa Barbara V		
	0.0		
	Chesen Compute O	1	
Description:			
Url local:			
Url pac:			

Step 2: Add Collection Level Metadata

Step 3: Choose Appropriate UCLDC Services (can be changed later)

anne- 32 anne-	
8	
0	
Decide which UCLDC	
	Decide which UCLDC services are needed at any given point.

Step 4: Manage Your Campus' Collections

in Edit Year History Bookmarks Book 1940			
🔄 🔅 dist-dev. cdib.org/collection_registry/admin/d_colle	ctions/collection/?calipus_id_exact=6	🖆 🗹 😋 🚮 - Geogle	P 🕈 #
Most Visited 🞯 Yahool Mail: The best 🛐 Welcome to Pacebo	sk 📋 eschol-dev.escholarsh 📋 SUBE: Available branc 🧶 eSch	olarship: Universi 🛄 My Journals 📈 voo2do - Your Dashbo 🥑	Dashboard - Pivotal Tr 🗌 Log In
Disable- 👗 Cookies- 🥕 CSS- 🞑 Porms- 📖 Images- 😗 Im	formation - 🧧 Miscellaneous - 🥒 Outline - 🥒 Resize - 🗩 Tools - 🔳	Vew Source- 🔝 Options-	10
Select collection to change Diango site admin 4	Quickly locate registered		
Django administration	collections.	weic	ome, dsc. Change password / Log
tome + DI_collections + Collections			
Select collection to change			Add collection
Q Search	19 results (188 total)	Manage the collections	Filler
Action:	elected	from your campus.	By campus
Name		Extent	d Berkeley
Paul G. Pickowicz Collection of Chinese Cultur	al Revolution Posters	(None)	Davis
Korean Posters		(None)	Merced
UCSD History: Oral Histories / 50th Annivers	ary	(None)	Riverside San Diego
Missions of Alta California		(None)	San Francisco
Estado que manifiesta el que tenian las mision	es de la antigua California	(None)	Santa Cruz
Call to Arms: Communist Ephemera of Madrid	During the Spanish Civil War	(None)	UCLA
Howard E. Gulick Papers		(None)	By access restrictions
Marquis McDonald Photographs		(None)	A A
Harry Crosby Collection		(None)	Restricted
Shots of War		(None)	Some items restricted
Visual Front: Spanish Civil War Posters		(None)	By status
Scripps Images		(None)	Completed
Sylvester M. Lambert Papers (Melanesian Arc	hive)	(None)	Planning
Roger Keesing Papers (Melanesian Archive)		(None)	By need for dams
Harold Scheffler Papers (Melanesian Archive)		(None)	1A
Southworth Spanish Civil War Collection		(None)	Medium
Dr. Seuss Went to War		(None)	Low
Advertising Artwork of Dr. Seuss		(None)	By ready for surfacing
18 collections			All Yes

Step 5: Edit a Collection's Record

×

Disable- L Cookie	s- / CS5- 🔯 Forms- 🔯 Images- 👔 Information- 🧮 Miscellaneous- 🥖 Outline- 🥖 Resize-	Y Tools- Wew Source- A Options-	101
Change collection 0	tiango site admin +		
ango adm	inistration	Welcome, dsc. Cha	
ne > Dl_collection	ns > Collections + Chinese seals		
hange co	llection	History	View on site
Name:	Chinese seals		
slug:	chinese-seals-imprints-and-drawings-c		
	Hold down "Control", or "Command" on a Mac, to select more than one.		
ampus:	Available campus @	*	
	W Filter	Update a collection's record as	
	Davis Irvine	information changes.	
	Riverside San Diego		
	San Francisco San Francisco		
	Santa Darbara Santa Cruz		
	UCLA		
	0.0		
	Chosen campus @		
	Merced		
escription:	Imprints and drawings of Chinese seals dating from the Warring States Period (5th ce	ntury BC) to the	
	20th century, contained in 14-volume set		
irl local:	http://ucmercedlibrary.info/on-display.		
n oac:			

.

Screen 3: Registry Concept Feedback Question

bit Ext
A rank a number of the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Comments (optional) Comments (option
POT1 LT3C UC Library Digital Collection Registry Concept Your Feedback is Essential The goal behind the Registry concept is to provide an efficient mechanism for gathering key information only available from content owners. Please let us know how feesible you think the Registry is for accomplishing this goal by answering the questions below. *1. Overall, the proposed UCLDC Registry would be an easy way for me to get my collections included in the UCLDC. Strangly Agree Agree Undexided Disagree \$Strangly Ragree *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strangly Agree Agree Undexided Disagree Strangly Agree Agree Comments (uptional) Comments (uptional) Comments (uptional) Comments (uptional) Comments (uptional)
Your Feedback is Essential The goal behind the Registry concept is to provide an efficient mechanism for gathening key information only available from content owners. Please let us know how feesible you think the Registry is for accomplishing this goal by answering the questions below. *1. Overall, the proposed UCLDC Registry would be an easy way for me to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (uptional)
The goal behind the Registry concept is to provide an efficient mechanism for gathering key information only available from content owners. Please let us know how feasible you think the Registry is for accomplishing this goal by answering the questions below. *1. Overall, the proposed UCLDC Registry would be an easy way for me to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Agree Undecided Strongly Agree Agree Ocomments (optional) *2.1 would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Ondecided Disagree Comments (optional) *2.1 would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Ondecided Disagree Comments (optional) *3.1 would prefer to notify the UCLDC about my collection(s) by:
 *1. Overall, the proposed UCLDC Registry would be an easy way for me to get my collections included in the UCLDC. Strongly Agree Gromments (optional) *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Strongly Agree Growthere Comments (optional) Strongly Disagree Comments (optional) Str
Agree Undecided Strongly Disagree Comments (optional) *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Indecided Disagree Strongly Disagree Comments (optional) *2.1 would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3.1 would prefer to notify the UCLDC about my collection(s) by:
Strongly Disagree Comments (optional) *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Agree Undecided Disagree Strongly Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Strongly Disagree Comments (optional) *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Comments (optional) *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
 *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
 *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Undecided Disagree Strongly Disagree Comments (optional)
 *2. I would be very likely to use the proposed UCLDC Registry to get my collections included in the UCLDC. Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional)
 Strongly Agree Agree Undecided Disagree Strongly Disagree Comments (optional)
 Agree Undecided Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
 Undecided Disagree Strongly Disagree Comments (optional)
 Disagree Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Strongly Disagree Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
Comments (optional) *3. I would prefer to notify the UCLDC about my collection(s) by:
*3. I would prefer to notify the UCLDC about my collection(s) by:
*3. I would prefer to notify the UCLDC about my collection(s) by:
*5. I would prefer to houry the OCLDC about my conection(s) by:
() Using a Deviation similar to the examples provided beyond
Cosing a registry similar to the examples previously shown Talking to a libraries on the compute
Contacting to a normalism on my campus
Desc
x 🔋

Screen 4: Thank You

🕹 [SURVEY PREVIEW MODE] POT1 LT3C UC Library Digital Collection Registry Concept Survey - Mozilla Firefox	
Ele Edit View History Bookmarks Tools Help	
🕙 www.surveymonkey.com/s.aspx?PREVIEW_MODE=D0_NOT_USE_THIS_LINK_FOR_COLLECTION8.sm=wAeHWq9VMM5jjAdm3qLyX145GW481juNY3t%2f%2bR1QC58%3d	*
POT1 LT3C UC Library Digital Collection Registry Concept	
Overview	
The UC Libraries Digital Collection (UCLDC) is in the very early planning stages. As part of that planning, POT1 LT3C has been charged w identifying a possible workflow for collection owners to notify the UCLDC about content for possible inclusion in the service. Because collect may be in differing states of readiness, we are proposing the creation of a registry that would allow collection owners to begin to participate UCLDC at any point.	/ith :tions ∋ in the
The UCLDC Registry would be a straightforward tool that would allow collection owners to:	=
Create and maintain a record describing each collection, thereby making the UCLDC aware of it and providing collection-level metadata.	
 Indicate which UCLDC services the collection requires, from DAMS submission, to metadata editing, to access. That information could a updated as required, for instance when access decisions change. 	lso be
Next	
Powered by <u>SurveyMonkey</u> Check out our <u>sample surveys</u> and create your own now!	
x	8