Final report

Library Pilot Data Curation Program Working Group;
Data Curation @ UCSB project

July 2014

"Bridging the gap"

1. Executive summary

This report represents the final recommendation of the Library Pilot Data Curation Program Working Group, which was formed as the final phase of the Library's two-year Data Curation @ UCSB project.

Following a whole-campus survey[1] and deeper interviews and investigations with over 50 faculty members, researchers, and Library staff, six faculty members were selected and agreed to participate in the working group. The faculty were selected based on the nature and complexity of the curation problems their research areas and research data pose; on their awareness of the problems of data curation and their interest in helping to find solutions; and, as a group, on the breadth of subject areas they represent. Joining the faculty were a researcher having significant experience in analysis of curation problems, and also data management education and outreach; and two research staff members having many years experience in data production, data management, and faculty engagement. The Library was represented by three staff members having metadata and subject liaison expertise. The working group held five face-to-face meetings in spring 2014, and also held teleconferences with the heads of research data curation units at several other university libraries.

The background of the data curation program, and the Library's proposed role in assisting in the curation of campus research data, are documented in the Data Curation @ UCSB project's interim report[2] and will not be repeated in full here. Suffice it to say that (quoting from that report):

> *...data curation is more likely to be accomplished by steps taken throughout the data's lifecycle, from inception and creation through publication and reuse, via partnerships between scientists, departmental curation efforts, and discipline-specific repositories and other external curation services. The premise of the Data Curation @ UCSB project is that, in such an environment, curation efforts would be expanded and enhanced by the presence of a cross-cutting campus unit supporting researchers in implementing data curation practices. Given its historic roles as both a collections archive and a service organization supporting all disciplines, the UCSB Library is the logical home for such a unit.*

In the working group's discussions, two gaps became apparent between the Library's proposed and desired role on the one hand, and the role it has historically played and continues to play today on the other. The first observed gap is between Library services

[1] Greg Janée and James Frew (2013). Faculty/Researcher Survey on Data Curation. http://dx.doi.org/10.5062/F4PN93K4

[2] Greg Janée and James Frew (2013). Data Curation @ UCSB: Initial findings and recommendations. http://legacy.alexandria.ucsb.edu/~gjanee/dc@ucsb/docs/initial-findings.pdf

proposed in the interim report and in interviews and surveys versus services used. Faculty members expressed great interest in obtaining help in data management and data curation from the Library, particularly in the planning and early stages of projects, beginning with help in formulating data management plans (DMPs). Note that all these services would require Library staff to be familiar with the faculty members' research. However, to date, faculty members reported using only the most traditional of the Library's services: if they had interacted with Library subject specialists at all, the interaction was limited to the Library identifying and obtaining books, journals, and other resources. The Library has already begun to offer training in data management plan creation using CDL's DMPTool[3], and this is a welcome first step in narrowing the service gap. Nevertheless, the general observation is that collaborating with faculty on research projects will constitute a significantly new role for the Library and require a new kind of relationship with faculty.

The second, more fundamental observed gap is in Library expertise. Faculty expressed a desire for help with complex, technical problems, such as working with database systems, data modeling, setting up and organizing projects, finding effective ways of identifying data to support long-term discovery, access and citation, and managing data over time. The working group found this to be true across the wide range of disciplines surveyed, from physical and computational sciences to social sciences and humanities. However, the Library staff currently in the best position to support faculty—subject specialists, whose job description already includes department and faculty outreach—lack the training to offer technical guidance at the deep level required. Moreover, the ability to offer help in technical areas is not merely a matter of education and training; rather, it is more likely to be borne out of hard-won experience, experience that UCSB's subject librarians have had no real opportunity to gain. The potential role of subject librarians in the area of data curation is further complicated by the concern, expressed by the Library side of the working group, that interest among subject specialists in taking on new data-centric roles and responsibilities, and receptivity to the required training, will vary considerably on an individual basis.

To begin to bridge the gap between services currently offered and services desired, the working group proposes that the Library undertake six pilot projects that will get the Library working with faculty and researchers more collaboratively, and working with researcher data earlier in the data lifecycle. The projects cover a range of subject areas, cover different stages of the research data lifecycle, and exercise different modes of faculty interaction. But while the projects are intentionally varied, all have the common underlying goal of education: in each case the Library is expected to gain expertise in some facet of data curation, and then be able to apply that expertise to new faculty, new data, and new projects going forward. All pilot projects work with real faculty and real data; they are precedents, not prototypes. As a result, by the conclusion of the projects the Library will be able to say that it is already and truly engaged in the curation of campus data outputs.

To bridge the expertise gap the Library will need to both train its subject specialists and department liaisons, and hire new curation expertise. In training subject specialists, it

[3] https://dmp.cdlib.org/

will likely be necessary for the Library to identify the (proper) subset of liaisons to serve as curation advisors. New curation staff will need to include both data curators and technical analysts, the latter having expertise in discipline-specific data formats and related technical issues. These new staff will need to work collaboratively with liaisons in ways they perhaps haven't before. For example, creating metadata for certain types of datasets will require both knowledge of data structuring techniques and conventions (particularly when using container-like file formats such as HDF[4]) and knowledge of how the data fits into and is best described according to discipline-specific ontologies and cataloging standards (such as the NetCDF CF Metadata Conventions[5]). This model—integrating new expertise with existing expertise—is one that university libraries comparable to UCSB's have followed to achieve success in adding data curation to their service portfolios.

The proposed pilot projects are ambitious, and will take multiple years to complete and to fully evaluate. Library staff time will be required, and success is contingent on staff having sufficient, budgeted time to work on the projects. It should be noted, however, that all projects are of a somewhat episodic nature: a project might involve holding a series of meetings with a faculty member, followed by a brief period of work; then a trial implementation phase, followed by an assessment; and perhaps another round of the same. Hence to a certain extent the projects can all be run in parallel.

The working group has three additional recommendations. The first is to bring issues related to data curation to the attention of the campus Information Technology Council (ITC). As a specific example of a project that might be appropriate for the ITC to consider, the working group discussed the possibility of offering to assign ORCIDs[6] (a type of persistent digital identifier) to all faculty and staff, as a way of better tracking researcher identity, citations, contributions, and impact. While such an effort may be led by the Library, it can also be seen as a campus enterprise effort with benefits not just to the Library, but to the Office of Research and departments (and of course to the researchers themselves). For this reason, the ITC may be the most appropriate body to address and fund such a proposal.

The second recommendation is to form an advisory group that includes campus faculty and researchers, to address the role that the Alexandria Digital Research Library (ADRL)[7] might and can play in the context of the research data lifecycle. Its current mission statement states that ADRL is "UC Santa Barbara Library's home for collections of digital research materials," implying that the repository's initial scope is solely the Library's own collections. But given that the Library is considering taking on an active role in the curation of research data products, and given that that role may require working with faculty data in stages of the data lifecycle earlier than final publication, it follows that ADRL's scope, purpose, and services may need to be expanded in the future to include, for example, self-deposit and faculty-led curation of Library-held collections.

---

[4] http://www.hdfgroup.org/
[5] http://cfconventions.org/
[6] http://orcid.org/
[7] http://www.alexandria.ucsb.edu/

The working group believes the Library would benefit from receiving external input in determining the direction of future ADRL growth.

The third recommendation, also related to ADRL, is for the Library to establish some expectations about the appropriate level and uniformity of value-added metadata for repository holdings. Traditional cataloging (e.g., MARC records) of inert items (e.g., PDF files) represents a base level of description and service. But there are new opportunities, and increasingly many opportunities, for describing data at more granular and semantic levels, particularly using linked data[8] and semantic web[9] technologies. Doing so has the potential of greatly facilitating data reuse and integration. New techniques and technologies in this area are rapidly evolving and require a fair understanding of emerging standards coming out of the semantic web effort and out of the various discipline-specific informatics communities. Researchers are unlikely to be aware of these developments, let alone be expert in them or have the time to address anything more than rudimentary metadata creation. But it would be natural for the Library to expand its existing expertise in metadata and cataloging into new realms of data description. By doing so, researchers would likely appreciate the Library adding value to their data, beyond simply preserving it, and ideally researchers will come to see the Library as a partner in the data publication process.

Finally, a cautionary note. The recommendations in this report should not be construed as being complete. The proposed pilot projects are only a series of first steps at addressing the problems of data curation, which are themselves relatively new and still evolving. There will still be curation problems to solve, and there will still be new and continuing activities that can be undertaken by the Library. There will also be curation problems that the Library will likely never be well-suited to solve. In particular, we note that data-intensive units such as the Earth Research Institute (ERI) will continue to encounter data management and curation problems that will probably require the establishment of a campus data center, staffed with data specialists well-versed in the data types of the relevant disciplines, to address. The Library, through its own data curation efforts, will certainly be able to inform and collaborate with any other curation efforts on campus. But recommendations for extra-Library efforts are outside the scope of this report.

The remainder of this report is organized as follows. Section 2 lists some of the observations and key points that arose during the working group's discussions, organized by the elements of the group's charter. Sections 3 and 4 give an overview of the pilot projects: selection criteria, as well as potential risks and obstacles. The remaining sections describe the pilot projects in more detail.

[8] http://linkeddata.org/
[9] http://www.w3.org/standards/semanticweb/

2. Working group charter

2.1. Identify prospective Library-provided data curation services during the pilot period

To address the following data curation problems:

- Faculty need education in, and hands-on assistance with, day-to-day management of their data and preparation for long-term preservation and curation of their data by others.
- Faculty curating a private collection as part of their research need help managing the collection throughout the collection's lifetime, not just at the end of a project or career.
- Faculty need help transitioning collections to repository storage, whether the destination be the Library's repository or another.
- Faculty curating small collections are often unsupported, and as a result may leave data curation concerns entirely unaddressed.
- Students need education in data management and, more broadly, data literacy.
- Departments and schools need a means to archive student work products.
- Researchers encountering difficult data management and curation problems lack an on-campus community to turn to for help.
- Data curation is not amenable to one-size-fits-all solutions. The technical skills and infrastructure required vary considerably between academic disciplines (and even between individual projects within those disciplines).
- Humanities and social science departments have less technical and infrastructural support (and their affiliated organized research units (ORUs) *provide* less technical and infrastructural support) than departments and ORUs that more frequently engage in large-scale, grant-funded efforts, such as those in the earth sciences. As a consequence, humanities and social science researchers are more in need of help.

The Library, through its pilot projects, will:

- Create and maintain a website, organized by subject area, of data curation-related reference materials: data repositories, metadata standards, best practices, citation methods, management tools, etc.
- Educate (a subset of) its subject specialists in data management and data curation issues, technologies, and tools.
- Gain proficiency in managing humanities research data types.
- Explore, and gain proficiency in the use of, common software platforms (e.g., WordPress, Omeka) to hold working versions of repository materials, particularly materials used in the humanities and social sciences.
- Outreach to selected faculty and researchers, directly assisting them with their data curation needs.

- Assist faculty with using external repository services.
- Assist faculty in managing data for future ingest into the Library's repository.
- Develop mechanisms by which the Library's repository will support direct and/or mediated ingest by external entities.
- Develop a data management curriculum for graduate students, and offer it as part of a more comprehensive curriculum on data literacy.
- Organize, host, and moderate a community of campus data managers.
- Assess the effectiveness and costs of doing the above.
- Develop a curation unit, led by data curator(s), to 1) serve as a point of contact for faculty and researchers, and 2) coordinate the pilot projects and, eventually, all the Library's curation services and curation-related outreach activities.
- Continue to reach out to campus faculty in search of new data curation use cases, particularly use cases in those areas underrepresented in the interviews to date, such as engineering and the media arts.

2.2. Identify potential initiators or precipitators for services intervention

The principal initiator for Library-faculty collaboration is the data management plan (DMP). The working group expressed much interest in DMPs, both help in preparing them and in reviewing already-prepared DMPs prior to proposal submission. It goes without saying that the help desired in creating a plan is not so much in writing the plan, but in identifying the tools, services, repositories, and other components that make up the plan, and in making decisions and establishing any necessary agreements regarding the ultimate disposition of the proposed project's data.

The Library's new relationship with the Office of Research, in which the latter forwards notifications of all new grant awards to the Library, so that the Library can make contact with the researchers, partially addresses the idea of focusing on DMPs as an initiator. The responsibility in the Library for making those contacts currently resides with an individual, but at some point will need to be moved to an identified staff role or responsibility. This nascent program might be expanded in the future to engage principal investigators at an earlier point, when grants are being prepared. Although doing so wouldn't cover all cases, the possibility of getting information from the Office of Research regarding selected proposals to limited submission opportunities (when the campus is limited to one proposal and the selection of that one candidate is done internally) should be pursued. Such an additional initiator would give the Library a chance to interact with a substantial number of researchers prior to proposals being submitted, and gradually raise awareness of (and perceived value of) using Library services while preparing proposals.

Faculty also expressed interest in obtaining help early in the project lifecycle to, for example, establish metadata requirements and be made aware of potential curation issues. To this end the Library will need to establish a well-defined point for accessing curation services (website homepage, help desk, contact person, etc.).

2.3. Identify the modes and characteristics of interaction necessary to use (from the data user perspective) and deliver (from the service provider perspective) the service

A common, if not universal observation in Data Curation @ UCSB faculty interviews is that faculty and researchers:

- are focused on their research area;
- are amazingly resourceful in bringing new tools and techniques into their research;
- but are not necessarily experts in using those tools, and in particular, not necessarily expert in understanding the curatorial issues associated with complex tools such as relational databases and GIS;
- are not experts in data management;
- are resource-constrained (mostly time-limited); and
- view data management as important, but of secondary importance to their research and to all the demands associated with their research (project activities, deadlines, proposal writing, etc.).

This observation is the motivation for the fundamental recommendation made by this report—a recommendation that is inherent and woven into the proposed pilot projects—namely, that the Library outreach to faculty, that it support faculty research projects, and that the Library even directly collaborate with faculty on their research projects in the areas of data management and curation.

From one perspective, this recommendation represents a complete reversal of the Library's traditional role. Instead of leaving faculty to their own devices for years and even decades, and then beginning to curate data only at the conclusion of a project or a career, the proposal is that the Library work collaboratively with faculty in managing and ultimately curating their data, throughout the data's lifetime, ideally starting with initial project planning. But from another perspective this recommendation is entirely aligned with traditional Library services. The Library has a long history of embedding subject specialists in departments and offering consultation to faculty and students on finding resources and using other Library services. So the proposed Library role may rightly be seen more as an expansion of the Library's lauded service role beyond information finding and into the realm of data literacy.

The faculty member can expect the Library to offer consulting services in the areas of creating DMPs and in data management and curation. Faculty will be made aware of the Library's services by virtue of the training of subject librarians and department liaisons, and by DMP classes taught by the Library. Additionally, the Library *may* offer to "embed" librarians on large projects that require a long-term commitment and dedicated time to make a meaningful contribution. Part of the evaluation of the pilot phase is to determine the costs associated with different outreach activities, and from there to develop costing models to determine what the Library can support. It is noteworthy that UC San Diego began its research data curation program by embarking on a number of pilot programs in which librarians were embedded in faculty projects, but found such a

deep level of support to be unsustainable. UCSD is now evaluating what level of service *can* be offered sustainably, and what services offered must be charged for.

From the Library side, outreach services will involve data curators, subject librarians, and other Library staff working with faculty and with each other. An overall leader of the Library's curation service activities will be required to keep track of faculty interactions.

2.4. Identify transactional mechanisms necessary to document and manage end-of-transition or conclusion of service

No particular end-of-project transactional mechanisms were identified by the working group. However, it should be noted that simply by virtue of the Library working more closely with faculty, the Library *may* be more aware of faculty research and activities. Ideally, the regrettable situation of a faculty member contacting the Library for the first time only thirty days before retiring after a 30-year career, and asking for the Library's help in archiving an office full of mixed materials, with the Library reduced to calculating the number of cardboard boxes and linear feet of shelf space required (a situation that was witnessed by the Data Curation @ UCSB project) will be avoided.

2.5. Identify requirements for staffing and training necessary to implement and deliver the services

At research data curation programs at comparable university libraries, the common staffing model is a partnership between dedicated curation staff (data curators, technical analysts, and others) and existing subject librarians, department liaisons, and metadata specialists. This partnership is expressed different ways but present in each case. At UC San Diego, curation staff take the lead in working with faculty, but bring in subject and metadata specialists as needed. At Purdue, subject specialists make the initial contact with researchers (via their HUBzero[10] platform), but then bring issues to the attention of curators. At the University of Minnesota, training courses are co-taught by data curators with general data management experience and by liaisons with subject matter expertise. The UCSB Library's plan to hire two additional data curators in addition to the already-hired geospatial data curator fits well with the partnership model.

Staffing levels at these other institutions are as follows:

- UC San Diego: A program leader; a researcher liaison; a technical analyst (who is not a programmer); and a digital preservation expert. Including the fractional time of subject and metadata specialists, the program adds up to 8 FTE total.
- Purdue: A project director (0.5 FTE); "technologists" (3.85); a HUBzero liaison (0.35); metadata specialists (0.2); a digital archivist (0.25); and a digital data repository specialist.
- University of Minnesota: a program leader; a repository developer; a preservation analyst; a metadata specialist; and several CLIR fellows. They are considering hiring graduate students to work with specialized data formats.

Note that at each of the above institutions the staffing levels do *not* include the staff needed to run, maintain, and further develop the institution's underlying repository

[10] https://hubzero.org/

system. These extra staff are organized into a separate group for reporting purposes, and are called upon as needed by the curation group. The implication for the UCSB Library is that additional staff will be required to manage ADRL itself, and to support ADRL as a tool that is used within the Library by Library curators.

It should be expected that the staffing transition (and what might even be called the *cultural* transition) involved in bringing data curation services into the Library will not be an easy one. Library members of the working group expressed concern about the willingness and enthusiasm of subject librarians to become more data literate and to participate in data-centric Library services. These concerns echo problems encountered at other institutions. At UC San Diego, the relationship and the communication mechanisms between data curators and its subject and metadata specialists are still being established. At the University of Minnesota, it was difficult "getting everybody on board" with their new services. In any case, data curators and subject librarians will need to develop a close working relationship.

2.6. Identify and recommend relevant campus partnerships and collaborations of potential value in implementing and sustaining the data curation service

The partnership with the Office of Research is seen as being very profitable.

The working group suggested that the Library also partner with academic departments, and to include Library-run data management education training as part of the departments' general professional development for graduate students (e.g., as part of research methods courses). In this way, the Library's education and outreach efforts to students could get a jump-start by building upon existing infrastructure for professional development already offered by departments.

Additionally, the proposed data management community, of which the Library would be the hub, could potentially open up many new partnerships the Library would otherwise be unaware of.

2.7. Recommend a mechanism to assess the pilot program and extend the outcomes to improving and/ or expanding the service post-pilot

Each pilot project includes its own assessment component.

Regarding post-pilot activities, the working group was intrigued by the Purdue program, which is centered around a whole-lifecycle data management platform (the aforementioned HUBzero) that also serves as a mediation mechanism between faculty and library curators. It would of course take the UCSB Library many years to achieve the same level of service and infrastructure as Purdue, given Purdue's early investment in this area, though it should be noted that UCSB may be able take advantage of Purdue's trailblazing because the HUBzero source code is open source. Even if the Library does not pursue HUBzero itself, its functions and the ideas behind it can serve as useful goals for determining the future direction of ADRL.

3. Pilot project overview

The principal recommendation from the working group is for the Library to engage in a series of data curation pilot projects. In this way the Library can gain some experience in offering curation-related services, and can build up expertise and capacity, before committing to offering the services to the campus at large. Six recommended projects are described below. These projects are ambitious, and will take time and effort to accomplish, but doing so will ultimately position the Library to be able to address many (though not all) of the data curation issues that arise on campus.

The table below describes the pilot projects by:

- **Subject area:** the subject area(s) covered by the project;
- **Works with data:** whether the project works with a specific data collection or not;
- **ADRL dependency:** whether the project has a dependency on ADRL, and if so, what the nature of the dependency is;
- **Data lifecycle stage:** what stage(s) of the research data lifecycle the project addresses; and most importantly,
- **Outreach/educational component:** what the educational value is to the Library (i.e., what expertise the Library can be expected to acquire), and in turn, what education the Library can be expected to pass on to campus faculty and researchers.

Pilot project	Subject area	Works with data	ADRL dependency	Data lifecycle stage	Outreach/ educational component
Sirat Bani Hilal (Reynolds)	Humanities	Yes	Yes, but may be deferred, perhaps indefinitely	Late	Yes: organizing humanities data, working with humanities platforms
Bren student projects (Frew)	Inter-disciplinary; education; GIS	Yes	Very dependent, but content is most ETD-like	Whole lifecycle	Yes: data management best practices
Maya forest (Ford)	Anthropology; archaeology; GIS	Yes	Yes, but may be deferred some	Ongoing/ late	Yes: managing researcher-curated collections
Fossil imagery (Porter)	Geology; imagery	Yes	No	Ongoing	Yes: using external data management services
Faculty outreach	Possible social science focus; or broader focus	No	-	Early	Yes: repositories, metadata standards, best practices, etc.
Data community	Unlimited, but probably earth science-focused	No	-	Whole lifecycle	People helping people

Table 1. Proposed pilot projects

Pilot projects were solicited from faculty, including both faculty on the working group and faculty identified during the Data Curation @ UCSB project's surveys and interviews. To be recommended, projects had to meet the following requirements:

- There must be willing and interested faculty and/or staff member(s) associated with the project. The goal of these projects is to enhance Library capacity in working with campus researchers, not in handling orphaned data.
- If the project works with data, it must work with specific, real data and real use cases; no prototypes or dry runs here. At the end of a pilot project, the Library must be able to say that it has a real accomplishment to its credit.
- The project must involve the active participation of Library staff, so that staff gain the necessary expertise.
- The project must afford the Library the opportunity to learn something new. The experience and expertise gained must be of a nature that staff can then offer that expertise to new faculty, to new projects, etc. The focus of the pilot projects is really on the expertise and capacity the Library will acquire, not on the data curated (though data may in fact be archived in the course of the pilot project).

In addition to these requirements, pilot projects were selected to cover different subject areas, different points in the data lifecycle, and different modes of faculty interaction. Thus the projects listed above and described in greater detail below cover the humanities to earth sciences to the social sciences; data management planning to end-of-project archival; and reference-desk-like consultation to collaboration and even partnership. But while the pilot projects are varied, if there is an overall theme to them, it is that they focus on integrating the Library earlier into the research process, and make the Library more of a collaborative partner in research as opposed to continuing the Library's traditional role as an archivist of finished products. Because the pilot projects have the Library working with real researchers on real data, by the completion of the projects the Library will be able to say that it is already doing data curation. In this way the projects are precedent.

4. Risks and obstacles

The working group identified a number of risks and obstacles to implementing these pilot projects:

- **Excessive project size (breadth).** In the course of working on a pilot project, it may be determined that there are not sufficient resources (time, staffing, other) to complete the project in its entirety. For example, the resources required to process one item in a collection may not scale to the entire collection. In this case the Library may need to scale the pilot project back. If possible, this should be done in such a way that the Library can continue to work on the remainder of the collection as time and resources permit, to avoid the pilot project becoming a prototype.
- **Excessive project size (depth).** A second form of excessive project size is the complexity involved. For example, it may be determined that the optimal curation strategy for a certain type of existing content requires that the content be

re-encoded in some way, but doing so is not affordable. In this case the Library may need to adopt less-than-optimal preservation strategies.

- **Over-obligation and over-promising.** As the Library begins to offer new services in the context of pilot projects, there is a risk that faculty (both those participating in pilot projects and those hearing about pilot projects) may come to expect that such services are henceforth supported and freely available, and can be sustained by the Library. Library staff (both on and outside the working group) have expressed a strong desire to proceed incrementally and cautiously, and to offer only those services that can be supported to a degree that the service operates successfully and can be sustained. In this way, the Library will build on its successes, and preserve the reputation it already has among faculty as an organization offering the highest-quality service. The risks of obligation and over-promising can be mitigated by carefully managing expectations among all parties, and communicating those expectations. Additionally, sustainability analysis and planning should be incorporated into the pilot projects.

And there are some obstacles, related to Library capacity:

- **Lack of data curation staff.** The Library's expertise in data curation will ultimately reside predominantly in data curation staff, specifically, data curators, but also in technical analysts. To a fair extent, the pilot projects depend on the Library having such staff already in place, to absorb and embody the lessons learned. The Library recently hired a geospatial data curator, but two other proposed data curator positions (in the humanities and in the "general sciences") remain only conceptual at the time of this writing.
- **Lack of data management expertise.** In a kind of chicken-and-egg problem, Library staff need experience to gain expertise, but much of the experience requires dispensing expertise. It may be necessary to invest in data management training before embarking on certain pilot projects in earnest.
- **Poor receptiveness to training.** Data curation represents an entirely new service area for the Library and, in turn, an entirely new set of skills to be learned. It is to be expected that some staff may be more receptive to branching out to this new area than others. It may be necessary to identify those staff in the course of pilot projects, and to focus on their training and contributions.

And there are several obstacles related to ADRL, to the extent that ADRL plays a role in pilot projects:

- **Lack of policy and definition.** As a brand-new repository, the basic policies governing ADRL's use are, at the time of this writing, just beginning to be formulated. Fundamental questions, such as the repository's scope and purpose, whether it will serve as an institutional repository, exactly how it will be used and by whom, etc., are still being answered. This lack of definition may force pilot projects to be deferred until the repository's role as a *campus* resource is better defined.
- **Lack of repository functionality.** ADRL currently supports only one ingest interface, for ingesting electronic theses and dissertations (ETDs) obtained from the Graduate Division; and one, default, generic access interface that allows those

ETDs to be downloaded. It is not clear what other ingest and access interfaces will be supported, by what date they will be developed, or who will develop them. Some pilot projects propose to interact with ADRL in ways that are not yet supported.

- **Lack of support staff.** The initial ADRL development is still being carried out by an external contractor. It is not clear that the Library has capacity at this time (in terms of either the number of staff or their expertise) to develop repository components itself, or to support curators (i.e., non-ADRL developers) in using the repository.

Eventually all these obstacles will be overcome, of course. But due to their presence in the short term, the Library will need to carefully evaluate the selection and staging of pilot projects. In particular, it may be necessary to focus, at least initially, on education and consultation only. And it may be necessary to use "bridge" solutions, such as other repository systems and tools, in place of ADRL until ADRL functionality is sufficiently robust. Additionally, regarding the evolution of ADRL, it may be beneficial for the Library to, a year after the pilot projects have begun, re-evaluate its infrastructure investments in light of evolving thoughts from the campus Information Technology Council (ITC) and Information Technology Board (ITB) about where the campus as a whole is going in regard to cloud services, policies regulating on/offshoring of data, etc.

5. Project: Sirat Bani Hilal

Rationale for inclusion: the humanities departments have less technical and infrastructural support than other departments, such as those in the earth sciences, that more frequently engage in large-scale, grant-funded efforts. As a consequence, humanities researchers are more in need of Library help. To be able to provide that help, the Library must gain more expertise in the subject.

Associated faculty member: Dwight Reynolds, Religious Studies

The goals of this pilot project are for the Library to gain expertise in working with humanities data; to be able to assist researchers in setting up and managing humanities projects and collections; and to easily ingest humanities collections into the Library's repository.

This project will take an existing, at-risk humanities collection that is currently being managed solely by a faculty member, and move it from its current *ad hoc* platform to a more standard humanities platform, and from there into ADRL.

For a specific collection, the first choice would be the Sirat Bani Hilal collection[11] of Arabic recordings, transcriptions, translations, images, field notes, and other related materials. This collection is proposed for several reasons:

- **Manageable**: the collection is relatively small and, at this point, largely static. It consists of all original source materials created by the faculty member.

[11] http://www.siratbanihilal.ucsb.edu/

- **Complex**: at the same time, the collection exhibits a number of data management and curation challenges, including: general lack of metadata; having a complex structure that is encoded only in the containing website, and outside of any archival objects; and being multi-lingual and using non-Roman character sets. Additionally, there is a desire by the faculty member to establish a community-based, but moderated translation facility if possible.
- **Prototypical**: the collection appears to be broadly similar to other types of humanities collections, in that it contains macro assemblages, each of which comprises multiple, multi-faceted, related components (in the Sirat Ban Hilal case, an audio recording together with a transcription and a translation). Any solution for this collection is likely to be applicable to other, similar collections.
- **Well-supported**: the collection's creator, Dwight Reynolds, has both expressed and displayed interest in helping the Library.
- **Obligated**: the Library has already made a commitment to house the collection.

An alternative collection would be the English Broadside Ballad collection[12]. While this collection is not in as immediate need of curation as Sirat Bani Hilal, since the Ballad project is still ongoing, it demonstrates many of the same characteristics. Regardless of the collection chosen, it would be beneficial to examine issues of more than one humanities collection to avoid developing solutions that are overly specific to any one collection.

In the first phase, the pilot project would identify an appropriate platform that is commonly used in the humanities, and get the content moved and encoded into that platform, while attending to metadata, identification, and organizational issues along the way. Possible platforms include a generic content management tool such as WordPress[13], or more humanities-specific tools such as Omeka[14] or Perseus[15].

In conjunction with the above tasks the project would examine text-encoding standards for content. The majority of text-like documents and data *should* be conformant to the TEI standard[16] or the equivalent disciplinary encoding standard (implemented in XML), at least at a gross level (with highly granular text-encoding provided by project developers and not by the repository itself). One can envision that one of the thresholds required to be met by a project being considered for inclusion in ADRL or other repository is that it must have minimal conformance to the relevant encoding standard, hence the importance of considering the issue of encoding.

A second phase would look at how a collection stored and managed in a such a common humanities platform can be placed under Library curatorial control. Possible solutions include complete transfer into ADRL; direct Library hosting of the humanities platform; or some type of hybrid approach.

---

[12] http://ebba.english.ucsb.edu/
[13] http://wordpress.org/
[14] http://omeka.org/
[15] http://www.perseus.tufts.edu/
[16] http://www.tei-c.org/

In addition to working on the collection's content, the project would analyze the tasks and level of effort required, and the project lifecycle stages at which those tasks can and might have been performed, so as to ultimately answer the question: What consultation offered at the beginning of such a humanities project would have resulted in a more easily curated collection by the end of that project?

Library personnel involved would include:

- a data curator, preferably a humanities data curator, to oversee the work, and to ensure that data curation concerns are satisfied;
- humanities subject librarians, to address content issues;
- metadata specialists, to address metadata presence, generation, and quality;
- technical specialists, to address platform issues;
- ADRL developer support, to support ingest into ADRL; and
- staff to evaluate the project.

The pilot project would require a commitment of time from the faculty member responsible for the selected humanities collection.

By the conclusion of the pilot project, we can expect that the Library will be well-positioned to:

- advise humanities researchers on content management issues;
- help humanities researchers set up projects and collections using a platform commonly used in that domain; and
- be able to more easily ingest humanities collections.

6. Project: Bren student projects

Rationale for inclusion: as the operator of a repository for campus-produced content, the Library must support ingest interfaces that can be used by other campus entities, directly and/or mediated. In such situations, the Library must be able to communicate and educate about metadata requirements and content organization best practices. More broadly, the Library should provide education in data management and curation to students and young researchers. This project addresses all these issues at once.

Associated faculty member: James Frew, Bren School of Environmental Science & Management

As part of obtaining a Master's degree, each student in the Bren School of Environmental Science & Management participates in a group project; as a whole the school runs about 20 group projects per year. Each group project produces multiple digital artifacts, including a final report, a poster and brochure, and supporting data.

As records of student achievement which the University has an interest in preserving, group projects fit under the general category of electronic theses and dissertations (ETDs); however, there are some key differences between Bren School group projects and ETDs the Library already receives from other schools and departments:

- Bren School group projects are performed collaboratively with, and for the benefit of, external clients: governments, corporations, non-governmental organizations (NGOs), etc. The projects are typically key components of, or rationales for, policy decisions undertaken by the clients. Thus, in addition to being records of student achievement, group projects are also publications of high value to the school and to others.
- Some supporting data is acquired, other supporting data is newly created or processed from acquired data. In any case, supporting data is often significant in its own right, and desirable as independent data publications outside the context of the containing project. This dual role (independently searchable, accessible, and reusable; but dependent and contextual) is a new use case for ADRL.
- Occasionally, supporting data is subject to non-disclosure agreements: the data can be analyzed and used to create derived products, but cannot be revealed publicly in its original form. This type of access restriction is a new use case for ADRL.
- Supporting data is frequently geospatial, but is not part of a more typical collection or series of homogeneous geospatial objects. This has ramifications on how the data can be found and accessed within ADRL. More broadly, a group project's supporting data may comprise datasets from a wide range of disciplines, with each dataset requiring type-specific access mechanisms.

In addition to these differences from ETDs, the model by which group projects will likely be acquired by the Library is very different from that for ETDs. The Library's current ingest stream for ETDs uniquely involves: a collaborative arrangement with the Graduate Division; the key participation of a private company (ProQuest); and the relatively simple transformation of complete, well-structured metadata. By contrast, ingesting Bren School group projects will be an acquisition use case that is both more challenging and more common. It will require at least a basic level of metadata preparation and evaluation by the Library, and will likely involve multiple metadata and subject specialists from different fields to evaluate each project's heterogeneous data types.

The goal of this pilot project is to catalog and ingest into ADRL the current archive of Bren School group projects, while addressing the many issues raised by this content. Ultimately, the project will develop an ingest stream in which Library staff work with Bren School staff, faculty, and students. This ingest stream will be a model for future interactions with other departments and campus entities.

Additionally, the pilot project will develop curriculum components that will be incorporated in the Bren School's data management course, which is slated to begin in the 2014–2015 school year. These components will address: the Library's metadata, packaging, and identification requirements; pointers to appropriate repositories for external data; data publication mechanisms and practices; and best practices in citation, source referencing, provenance tracking, etc. This education will benefit both students (learning good data management practices) and the Library (lessening the curation burden).

Library personnel involved would include:

- a lead data curator, to oversee the work;

- metadata specialists, to address metadata issues from multiple disciplines;
- subject specialists;
- ADRL development staff, to support ingest into ADRL;
- staff to develop and deliver the data management curriculum; and
- staff to evaluate the project.

An assessment phase should evaluate the comparative costs and quality of Library cataloging versus Library evaluation of student cataloging, and from there the efficacy of the data management curriculum.

By the conclusion of the pilot project, we can expect that the Library will be well-positioned to:

- continue to archive Bren School group projects with minimal cataloging and ingest effort;
- more easily establish new data ingest streams and relationships;
- educate students in basic data management practices; and
- understand the costs of adding new ingest streams.

7. Project: Maya forest

Rationale for inclusion: some collections spend much of their life being curated external to the Library, specifically, by a faculty member. The Library needs to be involved in the curation of such collections much earlier than end-of-project or end-of-career, ideally from collection inception. Doing so will require that the Library find new ways of working collaboratively with faculty.

Associated faculty member: Anabel Ford, MesoAmerican Research Center

The Maya Forest GIS collection is a fairly large (1TB), thematic collection of materials related to the archaeological and anthropological study of Mayan culture. It includes a number of types of data, including GIS data (raster and vector, at different scales), digital imagery, spreadsheets, relational data, and reports. Some of the materials (e.g., soils data) are original and unique; others (e.g., maps) were acquired, but many of the acquired materials have been annotated and otherwise modified in service of the research they are supporting, and in that sense they, too, are unique and thematically tied with the rest of the collection. The collection is continually updated as research progresses. The principal researcher has one staff member and a revolving assortment of graduate students and post-docs to help with data management issues.

This collection is a good example of what might be called an "organic" collection: what starts out as a small set of items supporting a research area slowly grows over time. By virtue of its completeness relative to its strong thematic focus, and in the case of a spatial collection like Maya Forest, its coverage of important spatial sites at different scales using different layers, the set of items achieves value as a collection in its own right, independent of the research it was originally curated for. As such, the collection becomes of interest to the Library as worthy of long-term preservation and access.

The curatorship of such a collection parallels its organic growth: initially the researcher is the sole curator, and ultimately a library (presumably, but not necessarily, the UCSB Library) will end up stewarding the collection indefinitely into the future. In between is a transitional period during which the Library begins to assume and assert control.

Unfortunately, for many collections, the transition does not occur smoothly, but is delayed until the end of the project or, worse yet, the end of the investigator's career, by which time less-than-ideal data management practices can yield a collection in too poor a shape to properly curate. Approximately 10 years ago the Library ingested a subset of the Maya Forest collection into the (now legacy) Alexandria Digital Library. The effort took several months of personnel time, mostly in addressing metadata issues (missing metadata, metadata lacking sufficient context, metadata conversion, etc.). Given the level of effort required for what was only a snapshot of the collection, the Library decided at that time that such ingest efforts would not scale, and therefore would not be replicated.

This pilot project will address the curation and ultimate disposition of collections such as Maya Forest in two phases. In the first phase, Library staff will examine the collection's metadata and its data identification and organizational practices, and make recommendations on necessary fixups to Maya Forest research staff. Librarians will monitor the execution and quality of the changes, and the time and resources required to implement the changes. To the extent necessary and supportable, Librarians *may* work on the collection directly, but the goals here are really for the researcher and her attendant staff to do the work, and for the Library to identify deficiencies, to communicate better (if not best) practices, and to monitor improvements and evaluate progress. In essence, the Library's proposed role is that of a collaborative, expert consultant. After this phase, we can expect that the collection (or at least subsets thereof) will be in suitable shape for archival.

In a second phase, the Library will examine mechanisms by which collections still being updated and curated by faculty, such as is the case with Maya Forest, can be gradually transitioned into Library control. Such mechanisms may include one or more of the following strategies:

- library curatorship of the collection *in situ*;
- initial and subsequent periodic snapshotting with explicit versioning of collection items;
- incremental uploads with change tracking to automatically detect changes; or
- full transition into ADRL while retaining faculty edit access rights.

Library personnel involved would include:

- a geospatial data curator, to examine the GIS content and lead the project;
- additional data curator(s), to examine non-GIS elements of the collection;
- metadata specialists, to address issues of metadata deficiency;
- ADRL development staff, to support ingest into ADRL; and
- staff to evaluate the project.

By the conclusion of the pilot project, we can expect that the Library will be well-positioned to:

- examine externally curated collections;
- offer guidance on data management and curation issues;
- direct and monitor external metadata preparation efforts; and
- effectively handle the transition of externally-curated collections into Library control.

The assessment of the pilot project should include the cost of such intervention and the efficacy of the advice offered.

8. Project: fossil imagery

Rationale for inclusion: small-scale data producers are often completely unsupported, and as a result their data can be left uncurated. The Library can make a big impact on the campus data curation problem by supporting these producers, even indirectly.

Associated faculty member: Susannah Porter, Earth Sciences

The goals of this project are for the Library to gain expertise in self-deposit and other data management services, and to be able to assist researchers in using these tools in ways that support their research and data workflows.

The project will work with a selected faculty member and his or her data. There are a number of faculty who generate relatively small amounts of data suitable for self-deposit services, but for this pilot project the first choice would be the Porter collection of fossil images because it demonstrates that a surprising number of complex curatorial problems can be generated by even a small set of images.

As background, Porter collects geological samples from around the world. These samples are dissolved in acid and the insoluble residues mounted on "stubs" and imaged with a scanning electron microscope. For each stub, an overview image is created that will serve over time as a kind of map of the stub. The map is stored as an Adobe Photoshop document. As Porter works with the stub and locates fossils of interest, they are identified as "specimens" and assigned specimen identifiers (e.g., "AK10-53-13F-7-001"), and their location on the stub is recorded as an annotation on the map (hence the use of Photoshop). Specimens themselves are imaged multiple times and from multiple angles at higher resolutions, producing TIFF images. These TIFF images contain no descriptive metadata; rather, the semantics of the images are derived largely from the image filename (which incorporates the specimen identifier) and from additional context provided by the containing folders. As Porter researches stubs and specimens, the image files get organized and reorganized in various ways (by stub, by taxon, etc.) as information comes to light and conclusions solidify.

As described thus far, this data workflow and organization is an effective means by which Porter conducts research. It's a practice that has been honed over a number of years.

Curatorial issues emerge once images are published. When an image is slated for inclusion in a publication, Porter identifies and coordinates with a geological museum to permanently archive the stub from which it was taken. At this point the museum assigns

the stub an accession number (e.g., "HUPC 62990"); a single stub may have multiple specimens each with their own accession number. Within the journal article, specimen images are identified in captions by specimen identifier and museum accession number.

The specimen images appear in the journal (technically, are embedded in the article PDF file), and if the publisher has adequately addressed the preservation of the journal, then the preservation of the specimen images follows. Nevertheless, there is value in retaining the original images since they offer considerably higher resolution, and since they are really the basis upon which Porter's conclusions rest. Equally of value is the physical stub, of course, since it is the ultimate basis of the article's conclusions. Its preservation is well-handled by traditional museum curatorial practices. Recall, though, the map that relates specimens to physical locations on the stub: without the map, no such linkage is possible. Thus all three things—specimen images, maps, and physical stubs—have curatorial value, and furthermore, the synchronization and relationships of these artifacts must be preserved along with the artifacts themselves.

Currently, specimen images and maps are stored on Porter's laptop computer, which is regularly backed up to a nearby external storage device. The journals Porter publishes in do not currently make source images available as supplemental data files. The images are not generally publicly available, and never have been. Thus, the curation of specimen images is not currently being addressed.

The curation challenges include:

1. ***Archival storage of images and maps.*** Storage on a researcher's laptop, even if backed up, is insufficient. An image repository must be identified.
2. ***Linkage between images as they appear in the journal article and image files.*** While an image is identified by a specimen identifier, there is no direct linkage to the source image file. Porter indicated a willingness to create such linkages. It remains to be answered, however, where and how such linkages would be recorded.
3. ***Linkage between specimen images, maps, and map locations.*** Filenames and identifiers do not follow rigorously enforced naming rules, and slightly different naming forms are used in different places, with relationships inferable in many cases but ultimately known only by Porter. For example, specimen "AK10-53-13F-7-001" might be marked as "specimen 7" on the relevant map; presumably the "-7" near the end of the specimen identifier corresponds to the identifier that is annotated on the map.
4. ***Additional images.*** Porter creates many images, and for obvious reasons it is not possible to include them all in the journal article. Yet the non-published images have value because they offer different views and perspectives of the same specimen, thus adding to the characterization of the specimen. While there would be no direct linkage from a journal article to these additional images, the relationships between images of the same specimen would be needed to be recorded.

5. ***Intellectual privacy.*** Porter would not be comfortable making public any images that are being used in any current research, because they represent sources of discovery and species naming.
6. ***Copyright.*** Porter receives royalties for inclusion of certain images in textbooks. Image licensing has not been an issue to date, because the images have never been made independently publicly available, but making the images public may force an examination of licensing issues.
7. ***Metadata and organization.*** Archiving files, particularly image files, without attendant metadata is generally unsupportable. The available metadata, the means by which it can be gathered, and the means by which it can be inserted or otherwise associated with the image files, remains to be determined. Also unknown is to what extent the organization of the image files should be preserved in the archival system.
8. ***File formats.*** The use of Photoshop for archiving and access is not ideal; TIFF or JPEG would be preferred. A conversion mechanism as part of the archival process would need to be implemented.
9. ***Workflow.*** A technical means by which Porter can archive specimen images must be created. A natural trigger for archival storage would be the publication of an image, but another trigger would be the creation of an image. The latter approach may require that images be embargoed for a period of time. How updates are handled would need to be addressed.
10. ***Public access.*** Repositories generally have a vested interest in making holdings publicly available. How these images would be made discoverable, searchable, and accessible remains to be determined.

The pilot project will proceed as follows:

- The Library will research available self-deposit archival systems such as FigShare[17], DataShare/Dash[18], and Zenodo[19], coming to understand their features and limitations.
- The Library will meet with the researcher, come to fully understand the researcher's workflow, needs, etc., and look for the best fit.
- The Library will set the researcher up with the selected tool, including enumerating the procedures for *how* the tool is to be used, and the best practices for doing so.

After a suitable period (several months at minimum), the Library will reconvene with the research to evaluate the effectiveness of the solution, and possibly refine the solution or selective an alternative and repeat the process as necessary.

Library personnel involved would include:

- a data curator, to lead the effort and assess the project;

[17] http://figshare.com/
[18] http://www.cdlib.org/cdlinfo/2014/06/20/uc3-dash-service-update-may-2014/
[19] http://zenodo.org/

- one or more subject specialists, to consult with the faculty member; and
- staff to evaluate the project.

By the conclusion of the project, we can expect that the Library will be well-positioned to:

- have a good understanding of external data management and repository services;
- be able to recommend their use in meaningful ways.

The assessment of the project should include the cost of the consultation, and the effectiveness of the solution.

9. Project: faculty outreach

Rationale for inclusion: the Library needs some initial, concrete tasks that will provide its staff with an opportunity to gain education in data literacy, and that will allow the Library to start advertising curation services to faculty and their departments.

The goal of this pilot project is to educate Library staff (principally subject specialists, but perhaps others) in data management and curation.

In a first step, the Library will select the staff to participate in the project. The Library members of the working group identified two possible courses of action in this regard. One possibility is to use the scholarly communication committee, which has the advantage of covering all research areas on campus and having among its membership the most experienced of the Library's departmental liaisons. A second possibility is to focus on the social sciences, where the experience levels among the subject librarians is decidedly more mixed, but the overall task is much more constrained. (The following description will proceed on the assumption that the scholarly communication committee route is chosen.)

Whatever the composition of the selected staff, the staff will first gather resources related to data management and curation, including information on relevant data repositories, metadata standards, citation practices, identification systems and methods, data formats, best practices, and the like. Due to the great differences in resources and methods between subject areas, resources will need to be gathered separately by broad subject area (humanities, social sciences, etc.) and, within the sciences, perhaps by discipline. A prototypical example of a resource the Library might locate in this phase and link to, and that also serves as model for guides the Library may create itself, is Michigan State University's "Research Data Management Fundamentals" booklet[20]. The ultimate goal of this resource gathering is to answer the question: What does a researcher need to know regarding data to get started working in a particular field?

The Library will organize the gathered resources, create a website for them, and continue to maintain the website in the same way that it maintains the existing guides it has created.

[20] http://img.lib.msu.edu/rdmg/RDMFundamentalsBooklet.pdf

Following this resource-gathering and website-building phase, selected staff will receive training in using the DMPTool, and will then offer informal classes and/or one-on-one consultation with faculty in their respective departments, and in this way advertise the Library's new services.

By the conclusion of the pilot project, the Library will:

- have gained substantial expertise in data management and curation; and
- be in a position to offer guidance in data management planning.

Assessment of this project should include the breadth of the faculty outreach that was achieved; the cost of training; and identification of gaps in breadth or expertise.

10. Project: data community

Rationale for inclusion: Difficult curation problems are being left unaddressed on campus. As the campus's neutral, inter-disciplinary unit, the Library may be uniquely positioned to facilitate their solution.

Researchers involved in large-scale, data-intensive projects on campus are encountering difficult data management and curation problems. This is particularly true in the natural sciences, where long-running (even decades-long) projects work with continually and even continuously updated data, and which must accommodate ongoing algorithm development, retrospective reprocessing, and general advances in scientific modeling. In such environments researchers are grappling with fundamental questions of how to organize and identify the data, how to implement versioning, how to track provenance, and how to archive software associated with the data and correlate software versions with data versions. But difficult questions are not confined to the natural sciences: humanities and social science researchers are grappling with, for example, how to curate projects that do not fit into any traditional repository model, such as those involving large-scale social media and those in which interactive experience is an essential component.

Researchers are attempting to address these data problems, to the extent that they are aware of them and that they are able, but the challenges remain. Help is unlikely to be found within the Library itself, as such projects and datasets are steeped in the language and theories of their disciplines, can involve highly complex processing steps, and undergo continual improvement as projects (and the science) evolve. If help is to be found, it is most likely to be found in colleagues who are working in similar areas and have encountered similar problems.

This pilot project proposes that the Library organize, and serve as the physical and virtual hub of, the campus's nascent "data community," with the mission of facilitating finding solutions to difficult curation problems.

The community will be organized around an online forum, hosted and moderated by the Library, that serves as a place where all manner of questions related to data—production, structure, organization, storage, access, identification, citation—can be asked and answered. The forum is intended to be similar to the long-running and successful CSF (computer support forum) mailing list that has brought system administrators on campus into contact with one another. The Library will host and moderate this new forum, but

beyond moderation, Library staff will monitor the forum to try to ensure that questions get answered, to watch for common topics and trends, and to keep track of on-campus expertise by noting participant activity and expertise. It is expected that the Library will be able to facilitate finding answers to questions by searching for expertise within the Library, on campus, in relevant communities, or by soliciting external help. The intention is that the forum will, over time, gain a reputation as a place one may reliably turn to when faced with a data-related problem.

With an electronic forum as foundation, the data community will be further built by the Library watching for frequently asked questions and trending topics, and on that basis propose informal meetings, talks, brown-bag sessions, and other types of events, whether led by the Library, by members of the community, or by invited speakers. In this way the data community becomes more than simply the set of subscribers to the forum; it's a group of people who come to know each other, and who happen to use the forum as a means of communication.

The working group was unanimous in its opinion that such a proposed community needs a catalyst to get started. The group's favored suggestion is that the Library hold a kickoff event, inviting participants by sending out a campus-wide announcement asking all departments to send interested representatives, and also by personally contacting key campus staff (department chairs, MSOs, ORU leaders, etc.). In addition to simply meeting each other, this group would, as an initial activity, be charged to come up with a set of common questions (and answers, of course) that they often get or that they consider related to their area of expertise—in essence, using the kickoff meeting time to pre-seed the forum with content. The group could also make recommendations for how to organize the content so that the eventual Library moderator would have some ideas from the group to work with. (It would of course require some thought how to constrain this task enough so that it would seem tractable and not too much work for anyone.) The benefit of such a task would be that it would "archive" some of the group knowledge from the very start. For the community members, such a kickoff activity would give them a sense of ownership over the resource and increase their sense of membership. For the Library, it would offer an opportunity to find commonalities between disparate groups on campus. It was noted in the working group that different research areas may use the same tools and technologies but in different ways, yet nevertheless encounter some of the same problems. This kind of observation may emerge from such a community meeting, and lead to new insights.

The principal risk of this project is the same fear that a party host has ("what if nobody shows up?"), i.e., that few people will participate in or contribute to the community. As one working group member put it, "I'll come to a meeting if I can get some code out of it." The criterion that one must reap some benefit from participating in a community is surely to be a common one, but the challenge of seeing that *all* members accrue some benefit in this particular case is that the problems for which data managers seek answers may be too specific to attract the interest of others. A researcher searching for a solution for delivering gridded earth science data is unlikely to be interested in a humanities researcher's problems in text encoding and markup, and vice versa. It is for this reason that this pilot project proposes to 1) expend considerable effort at the beginning, to look for commonalities within the community, and 2) to ensure that questions are answered in

the forum, so that members find value in belonging to the community, and therefore will be more inspired to contribute to it. Nevertheless, the risk remains.

The project will perhaps require the time of Library subject specialists and data curators, to assist in monitoring the forum. But it will also likely require some dedicated time from at least one staff member who is specifically tasked to forum maintenance and community building. The experience of other community-building efforts, such as DataONE's[21], is that dedicated effort is required.

The pilot project's success can be assessed directly by community participation, in terms of membership, questions asked and answered, and events held.

[21] http://www.dataone.org/

11. Appendix: working group membership

Michael Kim, *Library* (chair)

Greg Janée, *Data Curation @ UCSB project* (report editor)

David Court, *Earth Research Institute*
Anabel Ford, *MesoAmerican Research Center*
James Frew, *Bren School of Environmental Science & Management*
Stacy Rebich Hespanha, *NCEAS*
Alan Liu, *English*
Chizu Morihara, *Library*
Norman Nelson, *Earth Research Institute*
Margaret O'Brien, *Santa Barbara Coastal LTER*
Susannah Porter, *Earth Science*
Dwight Reynolds, *Religious Studies*
Stephanie Tulley, *Library*

12. Appendix: project participants

The following UCSB faculty and staff members participated in the Data Curation @ UCSB project to varying degrees: as in-depth survey participants, as case studies, and/or as interview subjects or consultants.

Bodo Bookhagen, *Earth Research Institute*
Manuel Carlos, *Anthropology*
Connie Christensen, *MesoAmerican Research Center*
Rolf Christoffersen, *Molecular, Cellular, and Developmental Biology*
Michael Colee, *Earth Research Institute*
David Court, *Earth Research Institute*
Jeremy Douglass, *English*
Jeff Dozier, *Bren School of Environmental Science & Management*
Andrea Duda, *Library*
Jack Engle, *Marine Science Institute*
Joel Feigin, *Music*
Erik Fields, *Earth Research Institute*
Ruth Finkelstein, *Molecular, Cellular, and Developmental Biology*
Erica Fleishman, *Earth Research Institute*
Anabel Ford, *MesoAmerican Research Center*
James Frew, *Bren School of Environmental Science & Management*
Mary Gastil-Buhl, *Moorea Coral Reef LTER*
Gunther Gottschalk, *Germanic, Slavic & Semitic Studies*
Michael Glassow, *Anthropology*
David Gurba, *Instructional Development*
Ruth Hellier-Tinoco, *Music*
Scott Hodges, *Ecology, Evolution, and Marine Biology*
Gerhart Hoffmeister, *Germanic, Slavic & Semitic Studies*
Chuck Huber, *Library*
Jon Jablonski, *Library*
Matt Jones, *NCEAS*
Kalju Kahn, *Chemistry and Biochemistry*
Michael Kim, *Library*
Kristin LaBonte, *Library*
Sotiria Lampoudi, *Computer Science*
Gene Lerner, *Sociology*
Corina Logan, *Psychology*
Tim Lynch, *Molecular, Cellular, and Developmental Biology*
Stacy Rebich Hespanha, *NCEAS*
Kent Jennings, *Political Science*
Stéphane Maritorena, *Earth Research Institute*
Aaron Martin, *Earth Research Institute*
Joe McFadden, *Geography*
Janet Martorana, *Library*
Norman Nelson, *Earth Research Institute*
Catherine Nesci, *French and Italian*

Margaret O'Brien, *Santa Barbara Coastal LTER*
Pete Peterson, *Earth Research Institute*
Linda Petzold, *Computer Science*
Christopher Pilafian, *Theater and Dance*
Annie Platoff, *Library*
Steve Poole, *Molecular, Cellular, and Developmental Biology*
Susannah Porter, *Earth Science*
Josh Preston, *Library*
Eric Prieto, *French and Italian*
Karl Rittger, *Bren School of Environmental Science & Management*
Dwight Reynolds, *Religious Studies*
Eunice Schroeder, *Library*
Peter Slaughter, *NCEAS*
Stuart Smith, *Anthropology*
Jamison Steidl, *Earth Research Institute*
Heather Stoll, *Political Science*
J.D. Thomas, *ThomaStudios*
Stephanie Tulley, *Library*
Thomas Turner, *Ecology, Evolution, and Marine Biology*
Libe Washburn, *Marine Science Institute*
Tara J. Yosso, *Chicano Studies*