Data Curation @ UCSB:
Initial findings and recommendations

Greg Janée
gjanee@eri.ucsb.edu

James Frew
frew@bren.ucsb.edu

November 2013

The Data Curation @ UCSB project[1] is a joint effort between the UCSB Library and UCSB's Earth Research Institute (ERI). The project is investigating data creation processes and curation needs within the UCSB research community. The knowledge gained will guide the Library in building a collaborative, sustainable infrastructure supporting campus-wide, cross-disciplinary curation of UCSB research datasets. This report describes the results of the project's initial investigations, recommends Library activities, and proposes the project's next steps.

[1] Funded by the Office of Research and the Executive Vice Chancellor's office.

Background: the science data curation problem

We're all aware of the Internet revolution that has taken place within the last few decades. The initial novelty of online, hyperlinked information has grown to become widespread, then commonplace, and now entirely expected. A similar and related revolution has happened in the way science data is managed. Data that was once analog has become digital, and has moved online. It is now expected that science data will be online, always available, citable, downloadable, reusable, and generally hyperlinked within the broader fabric of scholarly output.

The "onlining" of science data has led to a push to recognize the creation and preparation of datasets as intrinsically worthy of academic recognition, independently of any associated scholarly literature. Datasets are more frequently referred to as "published" and "citable," and many data publication mechanisms are emerging, ranging from the formal (new discipline-specific data journals, for example) to the less formal (data blog posts).

All these changes point to the increasing importance of **data curation**: preserving the data and ensuring its ongoing usability. The need for data curation (and our reliance on past successful data curation efforts) will only grow as our collective digital history grows longer, and as reuses of data reach increasingly further into the past.

But curating science data will not be easy, for it requires new organizations, services, and investments. Consider just three aspects of curating data: preserving the data; identifying the data and making it persistently citable; and capturing the data's provenance and maintaining its integrity. Each of these problems is difficult; each requires resources beyond the capabilities of individual researchers and their local campus units; and each operates over lifespans greater than those of projects, people, and perhaps even institutions.

- ***Data preservation.*** Physical data storage is the most immediate, visible, and easily quantified aspect of preservation. But preservation also requires an understanding of the preserved content sufficient to inform decisions about inevitable future curation-related actions. For individual researchers curating their own data this is no difficulty, but when we consider timespans greater than the individual's project, interest, career, or even lifetime, we face an almost insurmountable hurdle: transitioning from an individual labor of love to institutional practice, where roles and responsibilities are sufficiently codified that they transcend the individual actors. Preserving data is akin to delivering a package across time: the guarantee of its delivery must be a property of the organization and not a promise made by a particular deliverer. Few groups are equipped to tackle data preservation except those dedicated to that purpose.
- ***Identification and citation.*** Persistent identification and citation are critical components of data survivability. Without these, data interrelationships can be lost, and data can become "orphaned": undiscoverable, inaccessible, uninterpretable, lacking context. Persistent identification and citation have long been employed in scientific literature, but their application to science data (along with the concept of "data publishing") is relatively new. Researchers will need to

interact with global registries and publication systems, and may need to incorporate these interactions into the earliest stages of their workflows.

- ***Provenance and integrity.*** Maintaining records of data provenance, and ensuring data (and provenance) integrity over time, have emerged as important new requirements in data curation. As the historical record of digital science data grows longer, and as accesses into that record reach further into the past, these requirements will only increase in importance. The challenge here is that maintaining provenance and data integrity with any reasonable level of reliability requires fully automated, cryptographically secure tools and techniques, and this in turn requires that researchers incorporate tools just now being developed into their workflows.

Data curation clearly places newfound burdens on researchers and requires substantial support from external systems and institutions. But unless curation is proactively addressed, we are at risk of losing science data, or at least our ability to use it most effectively.

Curation and the Library

If the data curation problem is clear, the solution is not. Simply asking researchers to assume much of the additional burden of data curation is unlikely to succeed, for two reasons:

- ***Lack of resources and priority.*** Valuable as it is over the long term, curation does not necessarily directly or immediately enhance science being performed now. Time and resources allocated to curation may come at the expense of more immediate results or of pursuing the science itself.

- ***Lack of expertise and appropriate tools.*** To address curation, researchers must be aware of the problem in the first place, have access to the appropriate tools, and know how to use them. However, the tools and services provided by repositories and other curation institutions are often not intended for use by non-specialists. For example, most data repository ingest interfaces are designed and optimized for institutions doing bulk transfer. Citation systems impose metadata burdens that researchers have historically found onerous.

Nor can curation be entirely accomplished by a traditional collections acquisition model, in which an archival institution ingests finished, static objects for preservation at the conclusion of the activity that generated the objects. This approach places too great a burden on the receiving institution, at a point too late in the data lifecycle, and at a time when the researcher who generated the data has the least resources and desire to help with such a transition.

Instead, 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.

The specific goals of the Data Curation @ UCSB project are to better understand the data curation processes and curation needs on campus, and to develop and refine Library responses to these needs. Key questions to be answered by the project include:

- What services, infrastructure, and level of support must the Library provide to assume a role as campus data curation specialist?
- What are the Library's personnel requirements? What skills and training will be required? Will the Library need domain-specific curation specialists analogous to subject specialists?
- How will the Library interact with data-producing entities (individual researchers, departments, and ORUs), and at what points during the data lifecycle? Will the burden of curation be shared, and if so, how?

These are difficult questions, and there is a risk of developing programs, services, and tools that ultimately fail because they are based on an incomplete understanding of the problem space. Answers to the above questions must address these difficulties:

- Institutional support for data curation is relatively new and requirements are not well-understood. Similar efforts at other universities all appear to be in formative stages.
- Science methodology is undergoing a revolution, in which new technological capabilities (ubiquitous, online data; emerging identification and citation mechanisms; data publication) are establishing new norms for how science data is produced and consumed.
- The problems of curation are more social than technical. Addressing curation involves changes and new roles and responsibilities not just for the Library, but also for the researchers generating the data.
- Scientific communities and government agencies are also addressing data curation. Thus the problem of curation is not strictly between campus data producers and the Library, but involves various third parties as well.

Campus-wide survey

In late 2012 the Data Curation @ UCSB project surveyed UCSB campus faculty and researchers on the subject of data curation, with the goals of 1) better understanding the scope of the digital curation problem and the curation services that are needed, and 2) characterizing the role that the Library might play in supporting curation of campus research outputs. To achieve the largest possible response, the survey asked only five multiple-choice/multiple-answer questions:

- In the course of your research or teaching, do you produce digital data that merits curation? (A "no" response to this question precluded responses to subsequent questions.)

- Which parties do you believe have primary responsibility for the curation of your data, if any?
- Are you mandated to provide for (or otherwise participate in) the curation of your data, and if so, by which agencies?
- What data management activities could you use help with, if any?
- With which departments, programs, and ORUs are you affiliated?

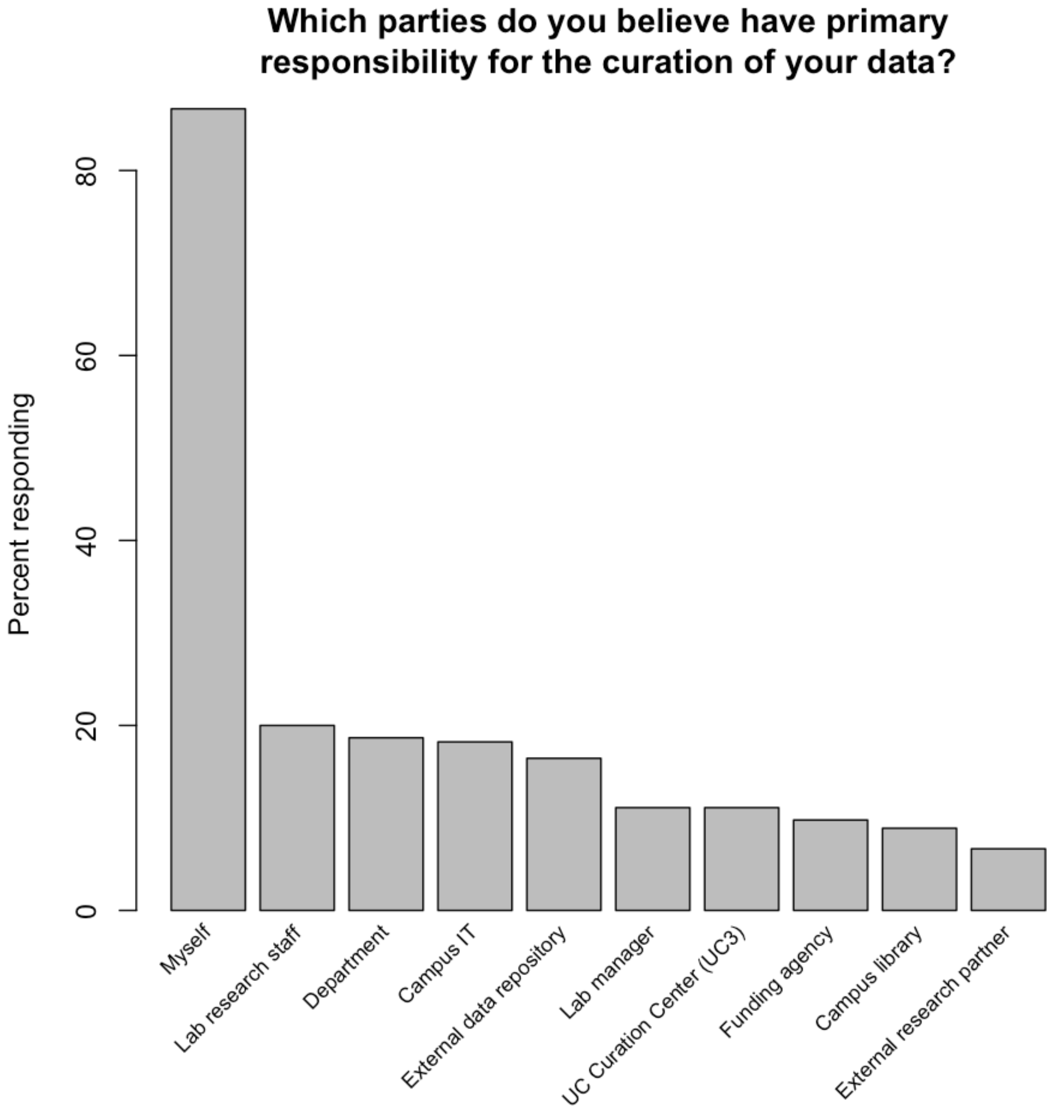
The project received responses from one-third of the estimated target audience of 900, indicating great interest in the topic and yielding statistically significant results. 77% of respondents answered "yes" to the first question of relevance of data curation; extrapolating, we conclude that digital curation is a concern for up to 60% of *all* UCSB faculty and researchers.

A summary of the survey's findings:

- Curation of digital data is a concern for a significant proportion of UCSB faculty and researchers.
- Curation of digital data is a concern for almost every department on campus.
- Researchers almost universally view themselves as personally responsible for the curation of their data.
- Curation is viewed as a collaborative activity and collective responsibility.
- Departments exhibit different characteristics with respect to curation, and as a consequence may require different amounts and types of campus support.
- There are many sources of curation mandates, and researchers are increasingly under mandate to curate their data.
- Researchers under curation mandate are more likely to collaborate with other parties in curating their data, including with their local labs and departments.
- Researchers desire help with all data management activities related to curation, predominantly storage.
- Researchers may be underestimating the need for help using archival storage systems and dealing with attendant metadata issues.
- Researchers under curation mandate request more help with all curation-related activities; put another way, curation mandates are an effective means of raising curation awareness.

The complete results of the survey are available separately[2]. In this report we present some select findings in more detail.

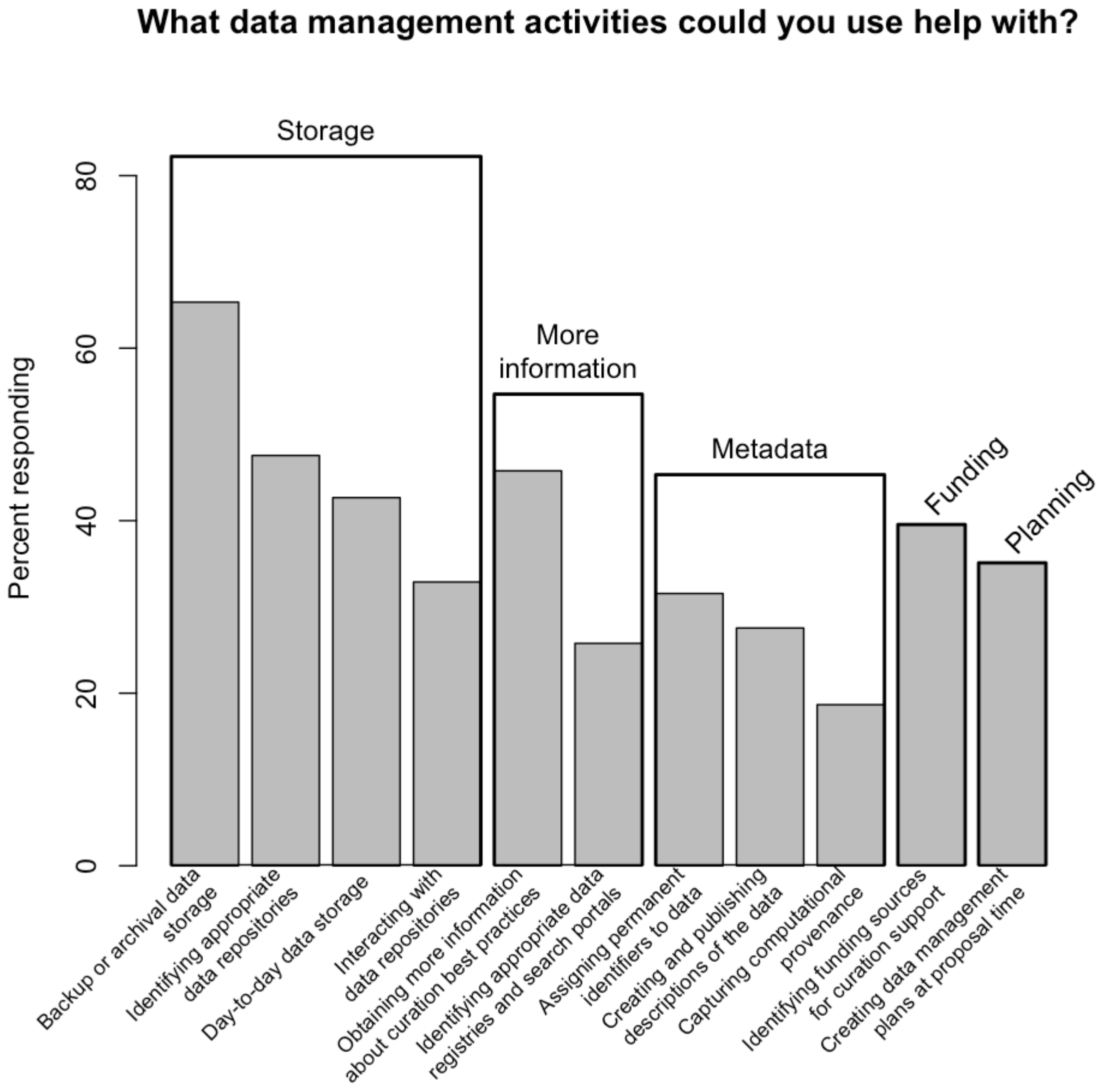
[2] http://dx.doi.org/10.5062/F4PN93K4



Almost all respondents identified themselves as being personally responsible for the curation of their data. At the same time, every other party mentioned by the survey was selected by a non-trivial fraction of respondents, and in the comments respondents supplied additional parties, notably journal publishers and professional societies. The large number of parties identified, and the *de facto* lack of formally recognized roles and divisions of responsibility in many cases, yields a complex landscape.

It may appear that responses to the question of responsibility are bifurcated between "Myself" and all other parties combined. But in fact, respondents who identified themselves as being responsible were more likely than not to identify additional parties that share that responsibility. Thus, curatorial responsibility is seen as a collaborative effort.

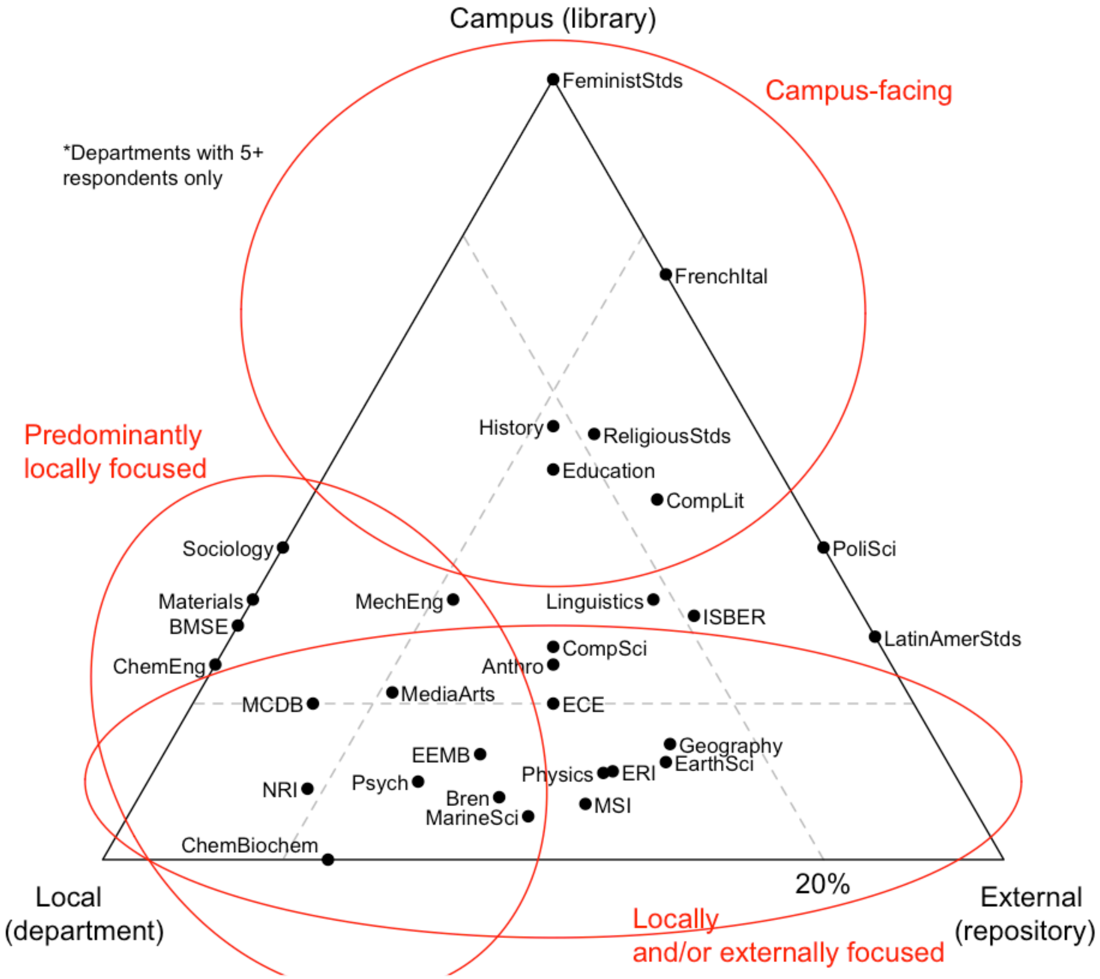
Those few respondents who excluded "Myself" as a choice indicated a greater reliance on external data repositories.



Respondents requested help with every activity the survey mentioned, and suggested several others in the comments. Help with backup or archival storage dominated the responses, and over 80% of respondents requested some type of storage-related help. However, even help with capturing computational provenance, a relatively specialized task with narrow applicability, was requested by almost 20% of respondents. Additional areas of help requested included digitization and developing data access tools.

Slightly fewer than 50% of respondents requested help related to metadata, a somewhat surprising result given that a common complaint made by researchers who use archival storage systems is that the systems are difficult to use, particularly because of the burdens they impose on metadata generation and object identification. This may represent a lack of awareness by researchers of the practical difficulties of using archival storage systems.

Distribution of departments* with respect to responsibility spheres



Ignoring the "Myself" choice of responsibility, we clustered the survey's choices for curation responsibility into three "responsibility spheres":

- "local" (comprising lab manager, lab research staff, and department);
- "campus" (comprising campus library and campus IT); and
- "external" (comprising external data repository, external research partner, funding agency, and the UC Curation Center).

Each department can then be plotted on a tri-plot, with the position of a department being determined by the average of its respondents' answers. For example, all responses from FeministStds (Feminist Studies) were in the campus sphere, and thus in the diagram above it is positioned directly at that vertex. If a vertex represents a 100% share of responsibility, then the dashed line opposite a vertex represents a depreciation of that share to 20%. For example, only 20% of ECE's (Electrical and Computer Engineering's) responses were in the campus sphere, while the remaining 80% of responses were evenly split between the local and external spheres; thus it is positioned at the 20% line opposite the campus sphere and midway between the local and external spheres.

Such a plot reveals that departments exhibit different characteristics with respect to curatorial responsibility, and look to different types of curation solutions. Departments can be loosely grouped as follows:

- ***Locally and/or externally focused departments.*** These departments look almost exclusively to external repositories or locally provided solutions. To the extent these solutions work, the departments may need little help from the campus.
- ***Predominantly locally focused departments.*** These departments look exclusively within themselves for curation. However, redundancy and fallback are key requirements of any preservation solution, and with no external or other support, locally focused departments may be at risk. Such departments may benefit from establishing backup and/or contingency relationships with a campus curation unit.
- ***Campus-facing departments.*** Lacking local or external solutions, these departments may be most in need of campus-provided solutions.

Results of additional investigations

Following the campus-wide survey, sixty researchers expressed an interest in participating in the project as a case study. Of those sixty, twenty responded to requests for additional information, and of those twenty, half were investigated to varying depths, ranging from analysis of the researcher's publications, data, and software, to email discussions, to in-depth, sit-down interviews.

We also examined existing support structures on campus that might serve as models for curation solutions, or even be co-opted to serve curatorial functions in addition to their nominal functions.

Interview results

Researchers are interested in the subject of data curation, and they appreciate its importance. At the same time, it is abundantly clear that researchers are primarily interested in their individual research areas, and are almost exclusively driven by the demands of their research and their professional careers. Therefore, any change in a researcher's behavior to support data curation—specifically, the assumption of any additional burden—will require motivation of some kind. That motivation may be (and, for externally-funded research, increasingly will be) funder-mandated data curation requirements. Otherwise, the motivation to (for example) use a new tool or service will have to be one of compelling value or convenience. Given researchers' focus on their research, we believe that simply asking a researcher to change their behavior out of a sense of duty is unlikely to be effective.

Beyond motivation, getting researchers to use new tools or change anything in the way they go about their work will likely require outreach (to raise awareness of new tools in the first place), direct encouragement, and, at least initially, personalized support.

Researchers (and the data they generate) generally fall into two broad camps. These camps are neither exhaustive nor mutually exclusive, but they are useful for discussion in that they lead to distinct curation problems and solutions.

Small-scale data producer. A prototypical example of a small-scale data producer is a researcher who takes some measurements, performs analysis on those measurements, and publishes the results in a journal article. The data is typically (though not necessarily) small in terms of bytes, and is generally created by simple workflows. The small-scale data producer typically has no dedicated staff to help them with their data management (though the researcher may have graduate students). There's a natural and obvious correlation between being small-scale and being unsupported.

Despite being small scale, the researcher's data may bring up many curation issues: metadata, citation, copyright, sharing, access, etc. But the greatest curation problem confronting the small-scale data producer is bringing their data "into the fold," i.e., getting the data into a curation-supportive repository in the first place; and, as a means to that end, getting the researcher set up with a workflow pipeline that eases and even automates ingest into a repository. Without that first step, there is little hope of addressing other curation concerns. Thus **the focus on small-scale data producers should be on providing them with ingest support**.

Large-scale data operation. Prototypical examples of large-scale data operations include remote-sensing datasets and *in situ* measurements from sensor networks. The large-scale data operation generates large amounts of data, and/or generates data continuously over long periods of time, and/or produces data using complex workflows. Large-scale data operations have dedicated staff, who are either associated with the project or resourced from the researcher's unit. These staff may refer to themselves as "scientific programmers," "data librarians," or "data managers."

Large-scale data operations can bring up extremely complex and difficult curation problems: total data size; storage and replication; organization of the data into components or granules; identification; citation; provenance; versioning; periodic retrospective reprocessing; continuous algorithm development; maintenance of uniform data formats, calibration, and processing levels over the lifetime of the dataset; and more.

Despite having dedicated staff to manage the data, these curation problems can remain largely unaddressed, for many reasons:

- The problems (or at least their recognition) are relatively new, so there is not an established body of practice or community knowledge that data managers can draw from.
- Norms for data management have changed over the last few years. Tools and techniques are still evolving.
- Data managers lack the training to recognize and solve curation problems on their own. (This is discussed more in the next section.)
- Data managers lack the time and/or funding to work on anything other than producing the data, supporting the researchers and the demands of the research program paying for the data, and supporting the data's users.

- Supporting data curation requires making changes to a running system that is already undergoing other changes due to concurrent algorithm development, continuous data acquisition, retrospective reprocessing, etc. A large-scale data operation most closely resembles a running engine: data is being continuously produced, and processing software is being constantly modified by data managers and researchers.

Offering outside consultation to a large-scale data operation runs the risk of being ineffective due to the complexity involved and the domain knowledge required:

- The complexity of data production processes precludes any easy dispensing of advice. Rendering any useful advice about a curation-related problem on a large-scale data operation often requires a deep understanding of the data (and even the science behind the data) and the production processes.
- Obtaining the requisite understanding is difficult. Workflow processes and software are not always documented, and even when they are, the documentation tends to be targeted primarily at other members of the project team and secondarily at other practitioners in the field.
- The Library has little experience in large-scale data production processes, and hence is not ideally staffed to offer help.

In this situation we believe **the most effective way to incorporate curation expertise and best practices into large-scale data operations is to introduce data managers to curation experts** and to new practices and tools, and to provide a forum in which data managers can network amongst themselves and share solutions to similarly-encountered problems.

Co-opting or emulating existing support mechanisms

Four existing support mechanisms were identified: campus mailing lists; campus centers; campus system administrators; and a class of staff we will refer to as "data managers."

Campus-wide mailing list. The CSF ("computer support forum") mailing list is jointly operated by the Office of Information Technology (OIT) and Letters & Science Information Technology (LSIT). Probably every system administrator on campus is on the list, which provides a forum for:

- asking questions related to system administration (and frequently getting them answered);
- announcing the availability of used equipment;
- requesting equipment;
- raising awareness of significant or upcoming events;
- issuing security alerts; and
- advertising classes and workshops on tools and systems.

The list also performs a general social function (e.g., organizing an annual beer bash). Per Kevin Schmidt, OIT network manager, "CSF is a collaborative affair in all respects, and has never been an ‘official’ university group or effort.” Nevertheless, the longevity of the list (~30 years) is a testament to its usefulness, and the list is a compelling example of how something as simple as a mailing list can bring campus employees together who would otherwise have no opportunity for interaction. Importantly, usage of the list does not require that system administrators cooperate or coordinate together, a critical consideration given the reality that, as one system administrator noted and others have echoed, “There isn’t really a cohesive IT management scheme at UCSB.”

Campus center. The Center for Spatial Studies (better known as “Spatial@UCSB”) was formed in 2007 to serve—and, in a way, to define—the geospatial community on campus. The center:

- provides a help desk, staffed by Geography department graduate students having knowledge of spatial theory and practical expertise in geospatial services and tools;
- presents talks of various kinds, drawing in both campus and external speakers;
- hosts annual meetings;
- hosts hands-on workshops oriented at practitioners;
- offers resources through its website; and
- runs limited services.

Such a center is an obvious model for a similar campus-wide unit focused on digital curation, particularly in the ways it cuts across traditional University boundaries and offers support for both experienced practitioners and novices.

Campus system administrators. It may seem that campus system administrators—as a class of staff—would be a good choice for providing curation support. After all, every researcher is serviced by system administrator(s) in every department and unit they work in. System administrators can also play key roles in providing file storage and performing backups.

However, system administrators are unlikely to have sufficient knowledge of datasets to be able to curate them. This is true even at ERI, where the system administrators have significant experience in relevant domains (remote sensing, oceanography, seismology, etc.), are familiar with every researcher and research program in the unit, and even directly participate in certain projects. Despite this, in many cases ERI system administrators are not familiar enough with specific data products to make curatorial decisions or take curatorial actions without guidance from attending researchers.

System administrators will necessarily play a role in providing foundational support for data curation. However, other staff will be needed to perform higher-level curatorial functions and maintain the knowledge necessary to perform those functions.

Campus data managers. There are many campus staff that perform much of the actual work of producing and processing data. These staff are typically officially classified as Computer Network Technologists, but refer to themselves by various titles such as “information manager,” “data manager,” or “data librarian.” We will use the term “data manager” in this report to refer to the entire class. Staff that refer to themselves as "scientific programmers" are frequently *de facto* data managers. These staff are associated either with a research group (the local Long Term Ecological Research (LTER) projects being the prime examples) or, more often, with specific researchers and/or projects (as is the case in ERI). The managers work very closely with researchers by:

- helping to develop, refine, and test scientific models;
- developing data processing software;
- running model software;
- managing data files and the artifacts associated with software and processing runs;
- creating and managing metadata;
- developing data access services; and
- uploading data to external repositories.

Given the key role data managers play throughout the data lifecycle, they are clearly prime candidates to (help) shoulder the responsibility of data curation. Indeed, some managers have already taken on that role; evidence of this can be seen in the Santa Barbara Coastal LTER's robust, published information management plan[3].

However, we generally observed a limitation in data managers' awareness of data curation issues and techniques, and in knowledge and use of standard software development practices that would enhance and simplify the management and integrity of data and the software that produces the data. The lack of knowledge of data curation issues is simply due to their newness, and a concomitant lack of an established body of data curation practice. By contrast, we believe the lack of use of standard software development practices is primarily due to the managers' backgrounds as discipline researchers (who happen to focus on developing software) as opposed to computer scientists or software engineers (who happen to be working in a particular scientific discipline).

Data managers typically have an educational background in a discipline related to their research area (geography, oceanography, biology, etc.); few have education in computer science or any significant experience in large-scale software engineering. (This is not surprising, given that the nature of their work and the ability to communicate with researchers requires deep domain knowledge.) This educational and experiential background may have been sufficient in the past because data production software, while mathematically quite complex and entirely steeped in the theories of the domain of research, tends to be script-like and relatively unstructured. But data curation places new demands on software and the management of workflow processes. Good data management is a precursor to data curation. Thus we see it as a problem that data managers show little experience with technologies adopted in larger-scale software development environments: source code control systems, assertions, test harnesses,

[3] http://sbc.lternet.edu/info_management/

automated build systems, and other techniques that work (and work together) to automate and guarantee integrity.

The data managers on campus are the most capable staff to be exploited in the service of data curation. Education in curation issues, practices, and tools, and in software development practices, would go far here.

Library recommendations

Our initial investigations lead us to recommend that the Library pursue three new activities:

1. Establish a curation unit. The primary purpose of the unit would be to support researchers and campus data managers in the curation of the data they produce, and not to support the curation of the Library's own collections (though of course the Library's internal curation activities would likely be informed by the expertise it offers to the rest of campus).

For campus researchers, the unit would be the focal point for the Library's curation-related services. The unit would provide a human contact point and a mechanism by which researchers can ask questions; in this regard, the unit would be analogous to the Spatial@UCSB center.

For campus data managers, the unit would form a virtual home in which they can ask questions, share tips, dispense advice, and otherwise network with each other; in this regard, the unit would play a role analogous to that played by the campus CSF mailing list with respect to campus system administrators.

Additionally, the unit would:

- staff a help desk;
- provide expertise in the latest curation-related concepts, services, tools, techniques, and other developments;
- keep up-to-date on, and be able to recommend, external repositories and services relevant to different disciplines;
- define, document, and promote data curation standards and best practices;
- create and disseminate tutorial materials;
- provide training classes on curation-related topics;
- host events, such as regular speaker series and occasional workshops;
- offer help in writing data management plans; and
- offer consulting services on campus research projects.

A curation unit would support large-scale data operations by providing a forum in which curation expertise can be shared among campus data managers, and in which new expertise can be acquired from outside practitioners. It would also support small-scale data producers by performing outreach to such researchers and their departments. It would raise awareness of data curation, and increase the uptake of curation practices on campus.

2. Develop a small object ingest service. A small object ingest service would directly support the small-scale data producers on campus who, as mentioned previously, are largely unsupported, and whose data may therefore go uncurated. The service would support ingesting one-at-a-time or one-batch-a-time small numbers of relatively small objects.

We use the term "develop" here in the broad programmatic sense and not in the narrow sense of software development—for example, such a service may be assembled from existing repository components and services and need not necessarily involve in-house software development. However, to be successful the service would need to be permanent, production-ready, attractive, and well-supported by the Library from the very beginning. Personalized support will probably have to be offered to early adopters. It is in this sense of commitment that the service must be locally developed, even if, in the software sense, it is acquired.

Large-scale ingest (of large and/or continuously-updated datasets) is of course a more demanding problem than small-scale ingest. But any large-scale ingest requires a programmatic solution (i.e., software development against repository APIs), and as such tends to be handled by data managers. In this way ingest for large-scale data is a solved problem in a way that small-scale ingest is not. It is the small-scale data that is often the unhandled case, and thus the activity proposed here is to develop an ingest service/tool specifically for small objects.

The focus of this activity is on the development of a *service*, not a repository system. Of course, a repository system necessarily requires some form of ingest interface, and conversely any ingest service must be connected to a repository system. Thus development of an ingest service would naturally follow the development of a local repository system, and indeed it would be natural for the service proposed here to be tied to any repository system the Library adopts or develops in the future. However, researchers would be better supported by a generic ingest service capable of being connected to any of a number of repository systems.

A small object ingest service could be modeled on the DataShare service[4], which presents a streamlined interface for uploading and managing files that is only incrementally more complex than DropBox and other commercial file-sharing services. But to be successful, we believe an ingest service needs some functionality beyond DataShare's current capabilities. To encourage adoption by researchers, the ingest service should also provide:

- a repertoire of recommended, relevant metadata standards;
- selection of metadata fields, both ad-hoc and from recognized standards;
- ingest templates;
- pre-population of metadata from templates;

[4] http://datashare.ucsf.edu/

- automated extraction of metadata from uploaded files;
- support for assigning persistent identifiers to uploaded objects and managing those identifiers over time;
- support for easily citing uploaded data objects and formatting citations;
- support for uploading files to discipline-specific repositories and journals;
- ongoing management of uploaded files; and
- synchronization of local and repository storage.

3. Develop tie-ins with the filing and execution of data management plans (DMPs). DMPs are increasingly being required by funding agencies, and for externally-funded research, DMPs represent the first, best, most well-defined point at which curation of a dataset is considered. Support for creating DMPs is already provided by existing services such as CDL's DMPTool[5], and that support would be augmented by the proposed Library curation unit's help desk. But because of DMPs' initial and central role in the lifecycle of datasets, DMPs may serve as a trigger mechanism for potential interaction between researchers and the Library.

Next project tasks

There are three tasks the Data Curation @ UCSB project can work on in its remaining time that would help further define and direct the Library's growth into data curation:

1. Determine staffing requirements for a curation unit. We propose to look at existing models, such as the Spatial@UCSB center, the UK's Digital Curation Centre[6], and other university-based centers. Note that the focus of this investigation is on staffing a unit that would provide a kind of home for existing campus data managers; this focus may differ from other university-based units that are more focused on providing local repository services.

2. Develop requirements for a small object ingest service. Actual development of an ingest service, even a prototype service, is not possible within the scope and timing of this pilot project. We can, however, gather requirements for an ingest service/tool. Specifically, we propose to look at:

- existing ingest services and components;
- metadata standards;
- metadata fields and field selection mechanisms;
- customizing/profiling/templating mechanisms;
- fields amenable to automated extraction;
- feasibility of automated extraction techniques; and

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

[6] http://www.dcc.ac.uk/

- integration into existing workflow systems and patterns.

The resulting requirements would greatly inform the design of a future service.

3. Investigate possibilities for DMP-related interactions. As mentioned previously, DMPs represent a unique opportunity to trigger interaction between researchers and the Library. However, the exact mechanism by which this might occur is unknown and requires investigation. We propose consulting with the Office of Research, ORUs, and departments that handle filing of DMPs, and answering questions such as:

- Are there privacy issues? To what extent are DMPs public?
- DMPs are typically required upon the filing of a proposal, but come into force only upon an award. What trigger mechanisms are available in each case?
- What is the protocol for interactions, particularly in light of the fact that the Library's role is necessarily consultative and not authoritative?

Appendix: sample small-scale data producer case study

Faculty member R_1 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 R_1 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 R_1 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 R_1 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, R_1 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 R_1'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 R_1's laptop computer, which is regularly backed up to a nearby external storage device. The journals R_1 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. R_1 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 R_1. 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.*** R_1 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.*** R_1 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.*** R_1 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 R_1 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.

Library implications: due to heavy workload and intense research focus, R_1 is unlikely to address curation concerns without outside prodding and help. Some outreach would be required to put a system in place that R_1 could use to easily archive work products, while addressing the above concerns. Such a system might require Library staff to invest non-trivial time in more fully understanding R_1's workflows. However, once a system is in place, it seems likely that R_1 would willingly and conscientiously support it.