OSU Libraries and Research Dataset Curation: A Beginning

Research & Innovative Services Report 4
February 12, 2009

Bonnie Avery
May Chau
Ruth Vondracek
Andrea Wirth

Table of Contents

Executive Summary

Research Statement

Identify strategies and resources which have proved successful at other libraries where programs for campus-wide dataset curation are in place. Articulate common "problems" that are encountered by the implementers of such programs. Make recommendations for further investigation and an "upgrade" or enhancement of service based on success at other libraries.

Research Sponsor

Faye Chadwell

Project Participants

Bonnie Avery, May Chau, Ruth Vondracek, Andrea Wirth

Project Scope and Assumptions

This research will:

- focus on the management and curation of born digital data sets either active or inactive, with particular emphasis on GIS data;
- identify strategies which have proved successful at other libraries where programs for campus-wide dataset curation are in place;
- address commonly articulated “problems” that are encountered by the implementers of such programs;
- include resource allocation questions when surveying other libraries, such as positions associated, time it takes, training, hardware, software, maintenance, preservation.
- discuss curation capabilities/resources currently in place at OSU Libraries
- be “informed” by prior efforts to define issues related to campus dataset curation (primarily focused on particularly spatial datasets) and by the survey of “data librarians” positions undertake by Andrea Wirth and Valery King.
- This research project will not:
- explore the curation and management of large scale datasets generated by supercomputers, such as those produced by USGS, NOAA, and NASA;
- not aimed at addressing the whole of the OSU research dataset curation issues on campus;
- not recommend whether or not to put a program in place at OSU Libraries.

Process

The research group of Bonnie Avery, May Chau, Ruth Vondracek, Andrea Wirth reviewed past communications concerning creation of a data center on the OSU campus. We conducted a literature review and web scan in order to identify libraries with data curation and management programs in place and to identify articles that dealt with data curation and management issues. We then searched the libraries' websites for further documentation. Using mind mapping techniques we defined gaps and common issues in the information found during our research and used that information to develop survey questions. We surveyed sixteen libraries.

Summary of Results

The literature and web searches, and the survey results revealed that no one existing model matches OSU Libraries' situation because these universities do not have the same infrastructure and they have greater resources. Each of these universities do provide solutions, tools, or processes that have implications for OSU Libraries, such as Oxford's planning process, Purdue's cheat sheet for librarian-researcher interview, and Cornell's overall's response to developing expertise in this area. Also responses to the survey indicate that many libraries, like us, are in the beginning stages and exploring the feasibility of establishing dataset curation and management services.

Recommendations

- The greatest impediment to dataset curation and management has been storage. Storage solutions (other than the Libraries' investing in more servers) and costs should be explored further. Terry Reese has begun this work, but a small task force could be established. A member of UO's library staff could be included on this task force.
- More investigation should be undertaken in management of dynamic data and datasets and their use in e-science.
- OSU Libraries should lead the effort with its partners to conduct a data inventory on campus, focusing on GIS data. Brian Westra, UO, has expertise in this area and should be consulted.
- OSU Libraries' should conduct a survey or interviews with OSU researchers to establish what is needed or lacking in data curation and management services on campus. This would include following up on the work done earlier by Cathy Howell's group.
- Archiving has already begun in ScholarsArchive with static datasets. What is needed at this point is the development of policies and procedures specifically related to datasets. The Digital Repository Work Group could assign a task force to work on this issue.
- The role of subject librarians in acquiring datasets and liaising with faculty should be defined. This responsibility could reside with either Collections or the Digital Repository Work Group.
- To move forward in the development of services a program of staff development and training should be developed. This would involve training subject librarians so that they

can be conversant enough to liaise with faculty and student researchers. Other selected staff would require training in how to curate and manage data.

Full Report

OSU Libraries' and Research Dataset Curation

Introduction

Sponsor: Faye Chadwell

Project participants: Bonnie Avery, May Chau, Ruth Vondracek, Andrea Wirth

Purpose of investigation:

OSU Libraries has considered the issues surrounding the collection, curation and management of born digital datasets, either dynamic or static, for some time. Questions have arisen about the level of involvement needed, personnel issues and training, and storage issues and costs. This research project undertaken by Research & Innovative Services (RIS) sought to identify strategies and resources which have proved successful at other libraries where programs for campus-wide dataset curation are in place. In addition we wanted to articulate common "problems" that are encountered by the implementers of such programs. On the local level we identified the curation capabilities and resources currently in place at OSU Libraries and reviewed the history of campus-wide discussions concerning establishing data centers. From this research we planned to make recommendations for further investigation and an "upgrade" or enhancement of service based on success at other libraries. For the purposes of this research project, we determined that curation and management of large scale datasets generated by supercomputers, such as those produced by USGS, NOAA, and NASA were outside the scope of this research.

Methodology

To understand the background on OSU Libraries' interests in dataset curation services we reviewed past communications concerning creation of a data center on the OSU campus and talked with library staff associated with ScholarsArchive. We conducted a literature review and web scan in order to identify libraries with data curation and management programs in place and to identify articles that dealt with data curation and management issues. We then searched the libraries' websites for further documentation. Using mind mapping techniques we defined gaps

and common issues in the information found during our research and used that information to develop survey questions. We surveyed 16 libraries.

Background

Historically, OSU Libraries' interests related to dataset curation services fall into four somewhat interrelated areas, all related primarily to spatial datasets:

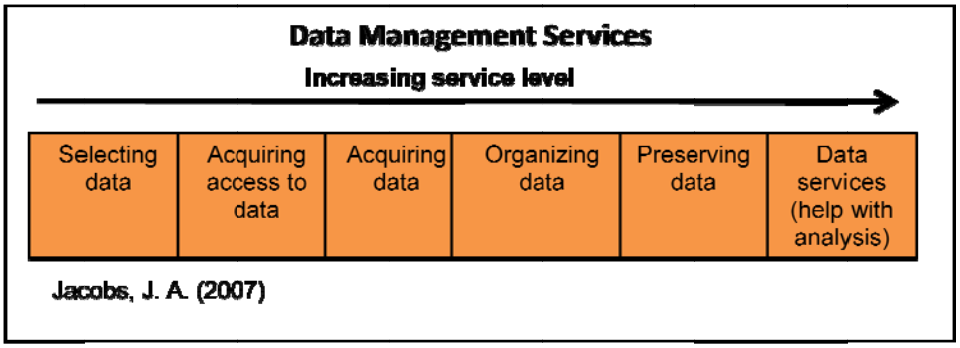
- Development of OSU Libraries Subject Librarian Expertise;
- Content: Aggregating geospatial datasets about Oregon;
- Dataset repositories and curation for Oregon Explorer partners;
- Storage and access issues, most specifically related to developing an OSU spatial dataset repository and inclusion of datasets in ScholarsArchive.

OSU Libraries Subject Librarian Expertise

Currently OSU Libraries does not have a data librarian position or anyone specifically trained in to manage or curate datasets. Sue Kunda, the Digital Production Librarian does have responsibility for adding static datasets to the ScholarsArchive. Andrea Wirth's (Geosciences Librarian) responsibilities do not include data curation or management.

This research project is informed by prior efforts to define issues related to campus dataset curation, primarily focused on particularly spatial datasets and by the reports on of data librarianship positions undertake by Andrea Wirth and Valery King. The geosciences/maps librarian position description at OSU Libraries has evolved overtime to acknowledge the importance of GIS/spatial datasets and tools. Likewise the government documents librarian has some responsibility for datasets in CD format from the US Census Bureau and other agencies. Prior to their general distribution via the web, OSU Libraries developed the Government Information Sharing Project (GovInfo), a web site that ran from 1995 through 2003. The site provided access to census, population, housing, agriculture, economics, and education resources. OSU Libraries terminated the website once the government began to distribute the datasets on the web.

Until recently, the extent to which GIS knowledge was distributed across campus mitigated the need for the library to assume a central role in locally produced dataset curation. In the spring of 2007 Andrea Wirth and Valery King prepared a report for the library on the range of services a Data Services Librarian might offer based on their investigations of peer institutions and the literature (Appendix A). In their report they summarized this range shown below.



Wirth (Appendix A: pg. 20) lists the known units using GIS on the OSU campus. To date no inventory has been completed with these units.

Content: Aggregating geospatial datasets about Oregon

This section includes detailed information on OSU Libraries interactions with its main OSU departmental partners and contacts in attempting to institute a data center on the OSU Campus. While grant funding has not been secured this background helps to elucidate OSU Libraries' relationships and its work to date and concerns using DSpace and LibraryFind for dataset curation and management.

OSU Libraries participated in planning and establishing for "Virtual Oregon", a data coordination center at Oregon State University (OSU) between 2000 and 2004. It was established in order to: "(1) archive environmental and other place-based data on Oregon and associated areas; (2) make those data accessible to a broad spectrum of agencies and individuals via innovative web interfaces; (3) identify key data sets that are not yet available and encourage their collection and dissemination; and (4) facilitate development of statewide standards for archiving, documenting, and disseminating data."[1]

Virtual Oregon functioned as a portal using a "distributed architecture that occupies multiple locations…". The four 'nodes' that comprised Virtual Oregon the Department of Geosciences (College of Science), the Forestry Sciences Laboratory (USDA Forest Service and College of Forestry), the Northwest Alliance for Computational Science and Engineering (NACSE), and the Valley Library. NACSE (Northwest Alliance for Computational Science and Engineering) discontinued this first generation Virtual Oregon site due to lack of funding. Tim Fiez developed a second generation site that was the precursor to Oregon Explorer that had a much modified form and a more limited partnership. The Virtual Oregon site itself is now defunct.

Content: Dataset repositories and curation for Oregon Explorer partners

In addition to the needs identified in the Virtual Oregon discussions, OSU Libraries encounters dataset curation services as a need among its Oregon Explorer partners, agencies and community groups, particularly the need for novice and/or expert retrieval, manipulation of, if not storage of,

datasets produced outside of OSU. Accessibility of OSU-produced datasets continues to be an issue as well.

Storage and access: OSU spatial dataset repository and ScholarsArchive

In the summer of 2006, Kathy Howell, then Co-chair of the Faculty Senate Computing Resources Committee, initiated a meeting with key campus staff and faculty to discuss the feasibility of establishing a spatial data repository for OSU and invited the library to participate in the hopes that the libraries' DSpace digital repository might prove a suitable platform. (Kathy Howell email 06/08/06)

By the time the meeting took place the agenda had changed to:

1) Identify the broad needs of the campus related to spatial data, especially as relates to how it is discovered, stored, accessed, and organized (our "audience" is the internal OSU campus community, as statewide/government agency needs being met by efforts of INR, Library's Explorer Series, in collaboration with State Geospatial Enterprise Office).

2) Develop a general strategy for addressing the identified campus needs. Our efforts will ultimately be successful if a strategic plan is developed which the university enacts upon.

OSU Libraries' attendees included: Bonnie Avery, Tim Fiez, and Jeremy Frumkin. Others attending: Todd Jarvis, Institute for Water & Watersheds; Jimmy Kagan and Kuuipo Walsh, Institute of Natural Resources (INR); Cherri Pancake and Dylan Keon, NACSE (Northwest Alliance for Computational Science and Engineering); Sheila Slevin, College of Agriculture (COA)/Crop&Soils; Chris Romsos and George Taylor, College of Oceanic and Atmospheric Sciences (COAS); Theresa Valentine and Terralyn Vandetta, College of Forestry (COF); Dawn Wright, College of Science (COS) Geosciences Department.

After a second meeting the group decided to host three "pilot projects" in the hopes these would better define overall scope and cost for other units around the university and make a more realistic case to the university or to funding agencies for long-term support.

Case studies included Forestry, Geosciences, and the Natural Heritage Program. Forestry's pilot project, with Theresa Valentine acting as data librarian, was to figure out the best way to map drives to campus networks, to advertise them and provide web accessible data. Geosciences, with a graduate student serving as data librarian, focused on trying to make ScholarsArchive work for spatial data. The Natural Heritage program wanted to find ways to distribute effectively its archived data as well as non-archived. Kuuipo Walsh served as data librarian.

The group agreed that the "Characteristics of successful pilot projects" would be:

Consultancy with Jeremy Frumkin, Tim Fiez and Terry Reese in order to articulate needs and identify specialized solutions, such as ScholarsArchive or Geospatial One-Stop.

Well documented. All procedures and potential pitfalls would be documented in a 'cookbook' to inform protocols and procedures for the future.

- Create a campus standard for minimum, abstracted metadata format. That is agreement that everyone must write metadata, everybody must provide metadata in this simple form, regardless of other methods in use such as FGDC, Dublin Core, etc.
- Provide case study scenarios to know how people will use the data. As an example, the Geosciences pilot tried their solutions out on instructors and students in courses to see if these courses could access data for classroom use.
- Create a methodology that others can use.
- Would be funded or has implications for future funding.

At a later meeting the OSU Libraries team shared DSpace's limitations and why it would not be the ideal storage repository for large, active datasets. Library representatives shared the attributes and limitations of DSpace as they related to archived datasets. "For example, DSpace archives files or set of files for download only; it can provide a persistent link for the file; it could use any metadata scheme – Dublin core, FGDC, although FGDC was not set up as a default; there would need to be some work to set up forms in order to handle entry of data and search on fields; it is good for large files size (2 GB or larger hard to do via web interface); not good for linking to other places via the metadata, it could for certain types of formats, but that would take further development. D-Space will cover MS-Word, but not GIS, although this would need to be policy for future migration. Storage is ultimately an issue with files other than documents; would require special development and programming for migration or to add full description and adding forms for retrieval (one time cost.)"
At this time the limitations of the open source LibraryFind tool related to GIS referencing seemed to place Geospatial One Stop (GOS) Arc9 toolkit in a better position to meet the immediate need for these case studies. Even with a system that delivers the workflow, archiving, and persistence components there remains the question about what other features and functions are needed and where do they reside.

The discussion of dataset audiences and access issues:

- Researchers: Faculty/staff/graduate students need a stable base data layers, 10-m DEM, that include watershed boundaries, road networks, stream networks, specialized land use, and hydrogeomorphic boundaries. They would also need a tool to help people automatically connect to the data and import/export/scale/ it.
- Teaching faculty need to locate data and tools. LibraryFind was developed for broad searches for undergraduates rather than as a deep searching tool for faculty.
- Building a bridge from LibraryFind to GOS might be possible in the future.

- Building in campus standards for harvesting of metadata

In February 2007 Tim Fiez, Dawn Wright, and Theresa Valentine submitted a $10,000 Proof-of-Concept Grant Proposal to the Northwest Academic Computing Consortium (NWACC) for a Spatial Data Mapping and Distribution System: Implications for Oregon "Collaboratories" which was not funded. During the spring of 2007 the University also relinquished responsibility for central distribution of software contracts. Significantly the ESRI contract oversight was assumed by Geosciences on behalf of others on campus. By the fall of 2007, the Campus Spatial Data Repository Group was on hiatus. By that time library representation on the group was limited to the technical team from the Oregon Explorer.

This group recently reconvened to determine whether the concept of an OSU Data Center could be resurrected with a similar proposal that might be a viable candidate for a grant (NSF – Math and Science Partnership (NSF-MSP) Innovation through Institutional Integration. Representatives from Linn-Benton Community College also attended this meeting. It is unclear at this time if a proposal would be appropriate for this particular grant opportunity.

ScholarsArchive is used to store some static datasets. The majority of these are associated with theses and dissertations. Dataset inclusion in ScholarsArchive is determined on a cases-by-case basis. To use the datasets users must download the data and provide their own software to manipulate them.

It should be noted that Terry Reese recently submitted a grant proposal that if accepted will impact the Libraries' involvement with dataset management and curation. This is a MIT-led proposal to the NSF Office of Cyberinfrastructure's DataNet Program on the MIT DataSpace project, federated data curation system. Included in that proposal is a request for both a research data analyst and a systems analyst.

While this background informs OSU Libraries that there is a longstanding need for a spatial, if not also non-spatial, dataset repository for the campus, it is not clear that "participating in the discussion" has been a sufficient tactic for contributing to the solution. Rather, an articulation of how the current library expertise and infrastructure can be the solution to a piece of the campus puzzle is the tact we have taken in this investigation.

For that reason we looked to cast a broad literature informed net to find libraries that seem to be a step ahead of us.

Literature Review

The rationale for our literature review was threefold.

- Find institution(s) that had dealt successfully with its need for dataset storage and curation.
- Identify additional issues concerning dataset curation we had not considered.
- Identify good candidate institutions for further investigation.

We searched LibraryLit, ACM Digital Library and Google for articles and documents. Where it seemed appropriate we used the Web of Science and Google Scholar to check the articles that were found as cited references. Once libraries were identified we searched their websites for further documentation.

The scope of this literature review is focused on identifying strategies and resources other libraries used to build campus-wide dataset curation programs. Since this concept is relatively new to the library community, literature about its development is scarce. A few -libraries in the world had established functional working groups for data curation programs, their scopes are broader than the RIS intended. These libraries collect a variety of data including text, statistics, images, objects etc, while Oregon State University Libraries (OSU Libraries) dataset curation project team is more interested in specific datasets (e.g. GIS). Nevertheless, publications produced by other libraries, including Cornell, Monash and Oxford provide valuable information; based on their experiences and the well thought out solutions they offered.

Literature Review Results

The CUL Data Working Group from the Cornell University Libraries (CUL) published a white paper *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities,* to furnish "an overview of the current landscape and issues surrounding data curation, and includes recommendations for CUL in this area"[2] This report is one of the most thorough reports available in terms of their environmental scanning, identifying issues and recommendations. It is a 'must read'; several of their recommendations seem applicable to OSU.

Monash University, Australia, communicated that the idea of the one single repository approach is not sufficient to support the entire research cycle. The report introduces the concepts of curation continuum, domains division and curation boundary[3]. University of Oxford reports a comprehensive investigative plan. Steps include stating of objectives, presenting overall approach, planned project activities, expected outcomes and analysis of stakeholder and risk.[4] This particular document may be of use as OSU Libraries pursues further development in data curation and management services.

Purdue University libraries also sees the need to change and thus created the Distributed Data Curation Center (D2C2) to "serve as a mechanism to bring researchers together to investigate ways in which optimal dataset management can be achieved at Purdue and throughout the research world".[5] In addition, a review specifically on literature focused on geospatial repository construction and design is also available.[6]

The rest of the selected literature targets many important issues such as data management, collaboration, user requirement analysis, staff development and jargon associated with data

management. In the literature, data management is a frequent topic, which includes data collections at the research level[7, 8], accessibility[4, 6, 7, 9], preservation [9], metadata standards [9, 10, 15] usability (use and reuse) of data [11, 12], policy[13, 14] and quality of data.[9]

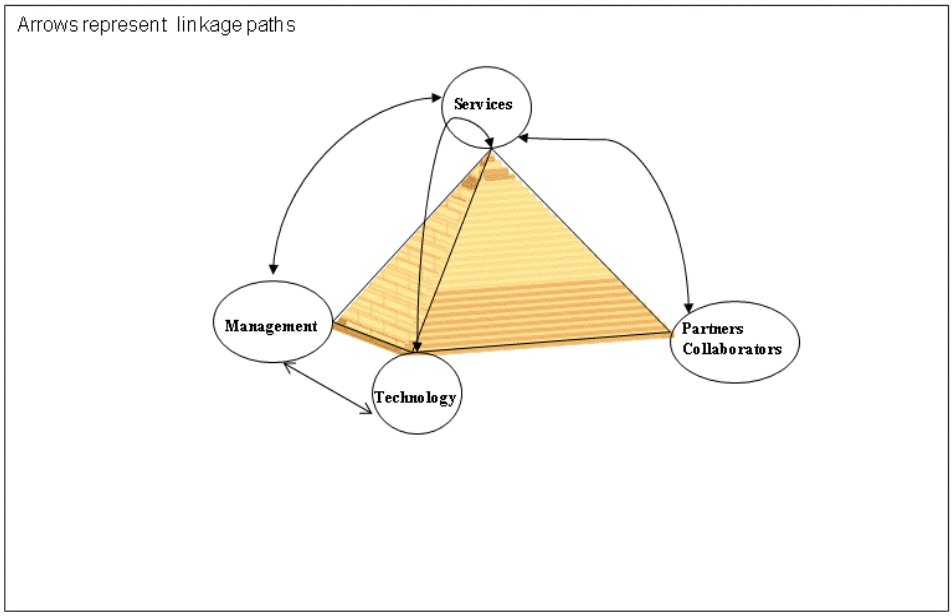
Collaboration[7] is not limited among researchers; the librarians' role [9, 15] in data curation has been discussed. For example, Purdue produced a cheat sheet for librarian-researcher interview for data curation.[16] Another piece of information needed to establish a successful data curation program is user requirement analysis. The Digital Curation Center in the United Kingdom published a user requirement analysis report, in which users are "discussed from the point of view of their roles in functional relation to research data, and also from the point of view of significant functions of organisational entities."[17]

The literature also provides insights on staff development in building data curation programs. It covers the areas of current practices and training need of research staff.[16], the need to provide "both institutional capacity and appropriately qualified individuals"[18] is also identified. Furthermore, jargons[6, 14,] in data curation programs are at times obscure and need to be further defined (e.g. archives[4], repository[6], Cyberscholarship[11]).

Survey of Issues

Informed by the literature review we brainstormed questions for further inquiry. Using mind mapping we categorized these questions and identified relational linkages (primarily, communication but funding and expertise also apply). This served as a means of organizing our survey and managing redundancy in our questions.

Data Curation Issues



We inquired about four aspects of dataset curation/preservation activities. These were:

- User services
- Partners and collaborators
- Technology
- Staffing levels and organization/management for this activity

We elected to use SurveyMonkey for our survey tool. The survey questions can be found in Appendix B.

Survey participant list:

Prior to culling, we had identified a list of 39 institutions. That list is available in Appendix C. To cull this list, we investigated their websites looking for:

- The presence of an institutional repository;
- Evidence of archived datasets and/or the inclusion of datasets in their IR guidelines for collection scope; and
- Other evidence to indicate that planning for dataset acquisition was in process.

This resulted in the following list of 16 candidate institutions for which we then located contact information. In November, 2008 we sent an email to each contact that included a link to the survey.

- Australian National University
- Cornell University
- MIT
- Purdue
- Rutgers
- Stanford
- University of Illinois
- University of Kansas
- University of Oregon
- University of Utah
- Utah [statewide] Digital Repository
- University of Washington
- California Institute of Technology
- Monash University Library (Australia)
- Scripps Institute of Oceanography
- Woods Hole Oceanographic Institute

Two other requests for information appeared on the SPARC listserv related to this topic as we were finalizing our survey. One was an open ended survey from Gabrielle Wong of the Hong Kong University of Science and Technology Library. We communicated with Ms. Wong and agreed to share our results. Then on November 13, 2008 SPARC served as the vehicle for a survey request from the Digital Curation Centre seeking input for a major international survey on digital preservation.

Whether these "competing" requests for information had the effect of adding survey fatigue for those presented with our survey is unknown. For whatever reason, we did not receive the volume of results we'd hoped, though we can benefit from some of the comments make in response to the survey from the Hong Kong University of Science and Technology Librarian. To date replies to this query have been received from MIT, UI/UC, Optical Society of America, and the Smithsonian.

Survey results

We sent invitations to complete our survey to 16 institutions. Eight of these initiated the survey and three completed it. Of the five minimal responders, four indicated they were in the "planning stages" while one respondent characterized the extent of data preservation was

"extensive" but provided no further detail. We received phone calls and emails from staff at three institutions, who felt that they were not far enough along to answer the questions; these individuals were asked to complete what they could.

University of Illinois and Cornell provided the most information and indicated that follow-up inquiries were fine. The other respondents indicated no interest in providing follow-up information and/or provided no contact information. Still since the University of Illinois respondent indicates that they are "in the planning stages" and Cornell, that they are "well underway", their responses proved useful. To this we elected to add the response from MIT to the SPARC LISTserv/Hong Kong University of Science and Technology request for feedback on collection research datasets.

Respondents indicate that they are offering their curation/preservation services to the entire campus and Cornell indicates that they are also working with a "handful" of state and local agencies. The services offered are long term storage, metadata assignment (assistance with identifying what metadata is needed), and in the case of Cornell: location aids for users, dataset manipulation software, collaborative online workspace (via Wiki, and DataStaR).

Guidelines for dataset collecting scope are most developed at Cornell and include:

- eCommons (quotes from survey)
- GIS Data Repository
- DataStaR

Issues related to rights management include further investigation into who actually owns the dataset (a case of this nature was mentioned by the UI respondent) and concerns researchers may related to "misuse" of their data (providing sample rights statements for them to modify).

Cornell's dataset repository efforts began with an NSF funded pilot project and continue with collaboration between the library and the Center for Advanced Computing on their campus.

Dataset deposition has been voluntary in all cases. Early adopters on campus were in plant science, anthropology and animal science at UI. They credit some of the interest in "sharing" with the fact that UI is a land grant institution with an Extension service. Cornell notes that early adopters tend to be from one of two communities:

- Those who rely on data from a shared facility or instrument (e.g. physics and astronomy).
- Those who work in a field where comparative or long-term studies make shared data valuable (e.g. ecology, genomics).

MIT noted that, “The researchers here haven't been shy about sharing their data in many cases. Those who are can talk to us about limiting access or setting an embargo period (that goes against our standard policy, but we can make exceptions if they're justified). Luckily copyright hasn't really been an issue since in the U.S. you can't copyright data, which does lead some researchers to want to keep their data locked up so that's a policy issue you'll have to work through for yourselves.”

While the institutional repository serves as a home for datasets for all respondents, some datasets are also stored on departmental servers. In some cases researchers are using commercial services as well. Datasets associated with electronic theses and dissertations (ETDs) and/or current student/faculty research are collected by all respondents and include:

- archival (closed) datasets,
- versioned datasets, and
- active/dynamic datasets (Cornell).

While UI integrates these services under its IDEALS rubric, Cornell indicates that integration of non-IR aspects of dataset curation are somewhat linked to the expertise and promotion by the librarian. To date, respondents indicate that depositing of datasets is most often done by library staff, though the dataset provider is authorized to do this task.

Cornell relates that problems of interoperability they noted include use of acceptable file formats and metadata. Their data librarian works with the researcher to reformat datasets, correct metadata, and be knowledgeable as to best practices.

Google-like searching and the institutional repository descriptions remain the primary means of third party discovery of datasets. Cornell experimented with extracting GIS metadata and converting it to MARC records but discontinued as it required too much work for the impact. They found that the catalog was not a path for dataset discovery.

The role of the library is quite varied:

- UI indicates that librarians work with faculty to get appropriate sets of material to deposit into the IR including the data, provenance and protocol information, and other publications that have resulted from the data. They indicate that it is not clear what is happening in other units on campus related to dataset curation.
- Cornell indicates that neither the library nor other units have formal responsibility for dataset curation but that the library is active in this area as are some other units.

IT staff have responsibility for dataset storage while dataset description and user services area is a collaborative effort between the library and dataset provider. Areas of needed expertise noted are: subject knowledge, knowledge of metadata standards, knowledge of the repository set up and management, current best practices for dataset preservation, and research computing. Consultation on what datasets are important to share and expertise in specific metadata standards were noted as on-going areas of need.

Current levels of staffing also vary. At UI the IR staff can also depend on "portions of subject liaisons based on FTE." At Cornell, three librarians devote a significant amount of their time to curation efforts. In addition, six to eight IT staff members spend a portion of their time supporting dataset curation services (including IR services). In both instances, old positions were "repurposed" to accommodate the need for staff. This repurposing includes a redefining of subject librarians' positions as well as folding in "datasets" within the collecting scope of the IR and its staff.

Respondents consider their efforts to be "successful" and that dataset curation has been well received by faculty. However, they indicate that it remains to be seen if their efforts are scalable and sustainable.

Conclusion

Cornell Libraries' white paper describes reasons for pursuing data curation and management in libraries succinctly.

> There are three primary (and related) motivations for developing a robust data curation infrastructure: enabling new discoveries by exposing data for use in data-driven research, ensuring access to and preservation of scholarly output, and meeting existing or forthcoming requirements of funding agencies or institutions regarding data management, retention, and access. Libraries have demonstrated expertise in several areas that could be productively applied to the practice of data curation, and in some cases, cyberinfrastructure development.[19]

OSU Libraries' has developed strong relationships with other campus entities that may lend themselves to developing on-going partnerships if we decided to pursue developing a robust suite of data curation and management services. Our current server infrastructure does not support storage of a large volume of datasets and we do not provide tools to manipulate the data. For that reason, we have relied on a distributed data model that continues to be a viable option. Other solutions for storing datasets are available on campus, such as NACSE and externally, through commercial means, but need to be explored further.

This report represents a preliminary investigation into the issues surrounding dataset curation and management and libraries involvement. We found that the most complete information about planning and implementing a program came from Cornell, Oxford and Purdue Universities.

The drawback at this point is that none of these models match OSU Libraries' situation because these universities do not have the same infrastructure and they have greater resources. Each of these universities do provide solutions, tools, or processes that have implications for OSU Libraries, such as Oxford's planning process, Purdue's cheat sheet for librarian-researcher interview, and Cornell's overall's response to developing expertise in this area. While we may not be able to follow a specific model, we could certainly adapt some of these practices in creating our own model.

Cited References

[1] Keon, D., Pancake, C., and Wright, D. 2002. Virtual Oregon: seamless access to distributed environmental information. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (Portland, Oregon, USA, July 14 - 18, 2002). JCDL '02. ACM, New York, NY, 387-387. DOI= http://doi.acm.org/10.1145/544220.544334

This article provides an excellent overview of the Virtual Oregon project and the motivation behind its development. Provides a highly technical view of the infrastructure and software used to develop and implement the site.

[2] Steinhart, G. et al. (2008). "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library" (http://hdl.handle.net/1813/10903)

[3] Treloar, Andrew and Cathrine Harboe-Ree (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". (http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf)

[4] Martinez Uribe, Luis (2008). "Scoping digital repositories services for research data management: a project of the Office of the director of IT". Oxford e-Research Centre, University of Oxford. (uk/odit/projects/digtialrepository/docs/DigRepoProjectPlan.pdf) February 27, 2008 ver. 2.2

[5] Mullins, J.L. (2007) "Enabling international access to scientific data sets: creation for the distributed data curation center". (http://www.iatul.org/doclibrary/public/Conf_Proceedings/2007/Mullins_J_full.pdf)

Discusses how Purdue University Libraries created the Distributed Data Curation Center (D2C2) in response to the needs of researchers and requirements for NSF funding.

[6] Bose, Rajendra (2006). "Geospatial repository literature review". (http://homepages.inf.ed.ac.uk/rbose/pubs/GRADE_RepReviewFinal_rkb.pdf)

As the title indicates this article is about geospatial data repositories. The purpose of the article is to communicate the information gathered during the planning phase of the EDINA GRADE (http://edina.ac.uk/projects/grade/) project. After a review of what it means to be a repository or archive, the author provides a brief synopsis of many existing geospatial repositories often including the scope of the collections and the types and sources of data. Metadata standards are also reviewed. The article is good summary of geospatial data repositories, but I think the most helpful part of the article is in the summary where the author describes how the GRADE group is looking at managing datasets - both depositing and accessing data and issues of curation and metadata assignment (curation in their case means editorial review of the deposited datasets).

[7] Palmer, Carole L., Melissa H. Cragin, P. Bryan Heidorn, and Linda C. Smith. “Data curation for the long tail of science: the case of environmental sciences”. (n.d.) (http://209.85.173.132/search?q=cache:L3Jz78dGKY4J:https://apps.lis.uiuc.edu/wiki/download/attachments/32666/Palmer_DCC2007.rtf%3Fversion%3D1+palmer+carole+data+curation+evironmental+science&hl=en&ct=clnk&cd=8&gl=u)

> The authors are developing a plan to manage long tail research data generated in an interdisciplinary field (Environmental Sciences) in order to help address how they might "unleash the market potential of these data collections and lower the barrier to access and reuse" (p. 2). They also address their project in terms of how a "coordinated or consortial" (p. 2) Institutional Repository might be a good economically viable and scalable solution to managing these smaller datasets.
>
> The remainder of the article describes the research study - how the authors are developing it at the time it was written. They provide a list of their research questions such as "how can research level collections best share and exchange data with existing resource and reference level collections" (p.3). Following that is a description of their survey and other data-gathering methods.

[8] Arms, William Y. and Ronald L. Larsen. (September 12, 2007) The future of scholarly communication: building the infrastructure of cyberscholarship. Report of a workshop held in Phoenix, Arizona, April 17-19, 2007.” NSF and the Joint Information Systems Committee. (http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf).

[9] Gold, Anne (2007)”Cyberinfrastructure, data and libraries; Part 1”. DLib Magazine September/October 2007 (http://www.dlib.org/dlib/september07/gold/09gold-pt1.html) and Gold, Anne (2007) and ”Cyberinfrastructure, data and libraries; Part 2”. DLib Magazine September/October 2007 (http://www.dlib.org/dlib/september07/gold/09gold-pt2.html)

[10] Henty, Margaret, Belinda Weaver, Stephanie Bradbury and Simon Porter. (2008) “Investigating data management practices in Australian Universities”. APSR Australian Partnership for Sustainable Repositories. July 2008. http://www.apsr.edu.au/orca/investigating_data_mangement.pdf

> The goal of this survey is to find out what the current practices and training need for research staff. Their target audience includes academic staff, postgraduate students, emeritus/adjunct and others. It tells us that what percentage the respondents are generating digital data. What are the types, sizes are being generated. How long the data will maintains its value, what is the software retention etc. In addition, equipment required in e-research is also mentioned, ownership etc. One of the interesting things they found out about training is that staff wants advices on 3 things. (1) Advice on data management plan, digitization, data exit plan for retiring faculty. They also talked about the interest in "tiers of access" with different privileges granted at different levels.

[11] Research information Network (June 2008)” To share or not to share: publication and quality assurance of research data output. Main Report.” Research Information Network in association with JISC Natural Environment Research Council(http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf)

[12] Rusbridge, Chris. 2007. “Create, curate, re-use, the expanding life course of digital research data”. EDUCAUSE Australasia 2007. (http://www.era.lib.ed.ac.uk/bitstream/1842/1731/1/Rusbridge.pdf)

[13] Soldi, Miguel (2006). “Safeguarding research data policy and implementation challenges (University of Texas System). http://connect.educause.edu/Library/Abstract/Safeguarding ResearchDataP/43896

[14] Carpenter, Leona (2004) “Taxonomy of digital curation users: part of the digital curation center user requirement analysis.” (http://www.dcc.ac.uk/docs/Taxonomy-dc-users.pdf)

Detail on digital curation users, their relationship to each other. They are categorized by (1) roles in functional relation to data, (2) significant functions of organizational entity.

[15] Pritchard, Sarah M. Smiti, Anand, and Larry Carver. (2005) “Informatics and Knowledge Management for Faculty Research Data”.
http://connect.educause.edu/Library/ECAR/InformaticsandKnowledgeMa/40109

This Educause Research Bulletin describes an investigation of the needs of faculty researchers in the areas of informatics on the UCSB campus. The study was funded by the Andrew W. Mellon Foundation.
The study primarily describes the findings from interviews with faculty members who met the criteria of being technologically innovative and had data-intensive research systems (p. 4). The areas addressed included data organization, storage, long-term preservation, tools and technique development, metadata application, digital rights management, and data sharing and system cooperation.

The authors found that lack of time or incentive to seek better methods for data organization, storage, and preservation were at the forefront. Some felt their current, if informal, systems were "good enough" (p. 6). Faculty are also heavily influence by what is required from their funding agencies for data preservation. Time was again an issue in the area of metadata use - even the faculty that were well versed in metadata standards for their discipline did not always see it as a necessary component of their research.

[16] Witt, Michael and Jake R. Carlson. “Conducting a Data Interview.” Purdue Libraries, Libraries Research Publications (2007).
(http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1092&context=lib_research)

[17] Carpenter, Leona (2004) “Taxonomy of digital curation users: part of the digital curation center user requirement analysis.” (http://www.dcc.ac.uk/docs/Taxonomy-dc-users.pdf)

[18] Henty, Margaret.(2008) “Developing the Capability and Skills to Support eResearch.”Ariadne. No.#55 April,2008. (http://www.ariadne.ac.uk/issue55/henty/)

[19] Steinhart, G. et al. (2008). “Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library” (http://hdl.handle.net/1813/10903). pg. 4.

Appendix A: King & Wirth Reports on GIS & Data Management Librarian

GIS Librarianship @ OSU
Prepared by Andrea Wirth for Ruth Vondracek
6/2007

Geographic Information Systems (GIS) Librarianship can mean many different things. Some factors that affect the role that librarians have in GIS services include:

- Library collections and technical support for GIS users
- Librarian knowledge and skills in GIS
- Marketing/making known library GIS services
- Other sources of support for GIS activities on campus

Library collections and technology support for GIS users

Data sets, data sets, data sets. All areas reviewed for this summary, including a very brief review of the published literature, a survey conducted by the libraries/Geosciences in 2000/2001, and comments from librarians, point to patrons with GIS needs typically needing the library to get help finding, accessing, or storing data sets or digital imagery and maps. Purchasing, linking to, locating, and accessing data are likely viewed from a reference or consulting services standpoint. Storing, preserving, and "cataloging" data could involve Library Technology or Technical Services on a greater scale.

A GIS Librarian would need to:

- know where to obtain geospatial data sets and be able to respond to requests for specific data
- know how to incorporate data into collection development activities
- obtain local data or direct users to data already owned or created by other members of campus (if not hosted by the library)
- solicit data for inclusion in the library's collection (if the library chooses to include this aspect)
- maintain an up-to-date collection of books, serials, maps and other media that are useful to the GIS certificate programs (professional, undergraduate, and graduate levels) and other units using GIS (see list at end of this document)

Providing technological support for GIS activities is another area that could be pursued, such as:

- providing workstations that could include the following (modeled after the University of Kansas Libraries)
    - software (ESRI products, Photoshop, and potentially statistical software)
    - digital manuals to the software
    - data sets included with ESRI products
    - local high-use data sets (including aerial photos)
- providing scanning, printing (high quality, large formats) in the library
- helping users troubleshoot data downloading
- software support and distribution

Librarian knowledge and skills in GIS

At minimum, a GIS Librarian should be trained in GIS theory and the ability to find appropriate data sets for patrons, provide instruction sessions for GIS courses (where finding resources and introducing library technology, etc. could be addressed). A librarian with more developed GIS skills could work directly as a GIS consultant and assist in the manipulation of data and production of end products.

Cooperation between the GIS librarian, map services staff, data services librarian (if separate and if the position exists), and the government information librarian is essential (as described by KU Libraries).

Reference
The GIS Librarian would need to be able to respond to a variety of types of reference questions. A list of reference questions that represent the types of questions KU Libraries received is included in Hauser's article. The Librarians here at OSU report that finding data and maps (digital and print) are the primary types of questions encountered.

If the library aims for a high level of technology support of GIS, the Librarian may also need to know how to troubleshoot software problems, distribute software, and provide assistance in data manipulation and map creation (as Scott W had previously done).

Instruction
There are some creative examples of GIS instruction at http://mapzlibrarian.blogspot.com/2006/03/kolb-learning-inventory-gis.html where the librarian at University of Texas at Arlington makes GIS instruction interesting by teaching GIS using "real world" problem solving activities. GIS instruction can also take a much more traditional approach as well and focus more on spatial data access and literacy and the software used to

work with data. There is also interest from the Geosciences for including Oregon Explorer in the GIS curriculum – both for it's use as a GIS application and for the administrative technology behind it.

Marketing/making known library GIS services

If the library begins to offer support of GIS activities on campus, there would be many opportunities for working across disciplines to define the services more accurately and find a niche for the library to fill. The OSU GIS community appears to be a fairly open and dedicated group that already works beyond unit boundaries. Adding the library to this group would not likely be difficult and would be a good inroad to further assessment of the GIS service needs.

Other sources of support for GIS activities on campus

OSU currently has some level of technical support for GIS users in specific communities. This would affect where the library needs to pool its resources.

There is currently (semi-official) ESRI software technical support on campus for the following groups:

- Forestry community (College of Forestry and Forest Service)
- College of Agriculture and Crop and Soil Science
- Geosciences, COAS, and Engineering

Many units that use GIS are not represented in these supported areas above (see list attached at the end of this document). This could be an opportunity for library services to focus on these unsupported departments, but this may take a more formal assessment to determine the need with certainty. As it is primarily the social sciences that are not known to be supported internally, this may mesh well with the discussion for data services support.

Sources consulted:

In addition to comments from OSU Librarians, a few GIS Librarians from other universities, GIS Librarian position announcements, and an informal conversation with Dawn Wright (Geosciences) the following resources were used to gain more insight into how GIS services could be designed at OSU Libraries.

*Hauser, R. (2006). Building a library GIS service from the ground up. *Library Trends*, 55(2), 315-326.

*Stoltenberg, J. & Parrish, A. (2006). Geographic information systems and libraries. *Library Trends*, 55(2), 217-360.

Virtual Oregon survey. (2002). http://virtual-oregon.nacse.org/people/index.html

*These both come from the Fall 2006 issue of *Library Trends* that was devoted to the subject of GIS and libraries.

List of *known* units using GIS on the OSU campus (provided by Dawn Wright - Geosciences):

College of Ag
- Ag and Resource Economics
- Crop and Soil Science
- Fisheries and Wildlife
- Rangeland Ecology & Mgmt

College of Forestry
- Forest Engineering
- Forest Science
- Forest Resources
- HJ Andrews

College of Science
- Botany and Plant Path.
- Environmental Sciences
- Geosciences
- Horticulture
- Statistics
- Zoology & PISCO

College of Liberal Arts
- Anthropology
- Political Science
- Sociology

College of Health & Human Sciences
- Public Health

COAS
- Oregon Climate Service and Oregon Spatial Climate Analysis Service
- Marine Geology & Geophysics

- Marine Resource Management Program
- College of Engineering
  - Biological and Ecological Engr
  - Civil Engr
  - EECS
  - NACSE
- Valley Library, especially Digital Collections and the INR
- OSU Extension
  - Columbia Plateau Conservation Research Center
  - Klamath Experiment Station
  - Northern Willamette Research & Extension Center
- HMSC

Data Services Librarianship @ OSU

Prepared by Andrea Wirth and Valery King with assistance from Michael Baird for Ruth Vondracek

6/2007

Research in a wide variety of disciplines has increasingly involved the collecting and manipulation of sets of digital data. Students and other researchers need to access, manipulate and analyze data sets, and frequently have questions about it. Our users very often need assistance with using data sets that the library receives from the state and federal government, as well as the variety of social sciences data that OSU users download from ICPSR. OSU researchers need to archive the data sets they produce, and are increasingly requesting that the library acquire more data sets from other sources. We have become ever more aware that there are some gaps in our collective knowledge and in our services that could be addressed by hiring a Data Services Librarian.

Defining a Data Services Librarian for OSU depends on the level of support that the library wants to pursue as well as whether the duties of the position will be combined with other developing services (such as Geographic Information System services). Below is the spectrum of data services as summarized from Jim Jacobs' talk earlier this year at Willamette. This visual provides the most succinct depiction (that we have found) of what data services can mean at OSU Libraries. Though the spectrum was designed with social science data services in mind, it could apply to geographic information systems (GIS) services as well.

Selecting data	Acquiring access to data	Acquiring data	Organizing data	Preserving data	Data services (help with analysis)

Increasing service level---------->

There is often overlap between numerical data services and GIS services in library literature, position descriptions, and existing or developing library data services. Two of the position description reviewed for this summary included a major GIS component: Geospatial Data Librarian at Emory (http://web.library.emory.edu/services/hr/geospatial_data.html) and the Map and Data Services Librarian at the University of Illinois - Chicago (http://www.uic.edu/depts/lib/admin/personnel/map_lib.pdf). The announcements for the positions at the University of Southern California (http://www.usc.edu/libraries/jobs/librarians/documents/223DSocSciData.pdf) and at Notre

Dame (http://www.library.yale.edu/~llicense/ListArchives/0010/msg00015.html) also explicitly list working with the Government Documents librarian to facilitate user access to data received from the government, notably the U.S. Bureau of the Census.

One noticeable feature of these position descriptions is the need for the librarian to work in an interdisciplinary capacity. While many required skills listed in position announcements for data or digital librarians are similar to those currently required of subject librarians--collection development, instruction, reference, liaison and so on—data services librarians do not work within a single subject area. Whether the focus is numerical data or GIS services (or both), the librarian will likely assist users from across the campus and may be more of an expert in quantitative data analysis and/or GIS than an information specialist in any particular subject area.

Skills and duties of a Data Services Librarian:

- Knowledge of statistical processes and terminology including understanding how users (individually and generally) want to use data
- Knowledge of statistical software (SAS, SPSS, STATA are often listed in the position descriptions we reviewed; SAS, S-PLUS, and R are featured prominently on OSU's Statistics department website)
- Ability to work with data files (downloading data and troubleshooting format issues for successful use with statistical software)
- Education or background in the social sciences
- Collecting (either through purchase or providing access) data sets
- Creating and maintaining a user interface that provides access to data
  - Example: University of Connecticut Libraries Social Science and Geospatial Data Services (SSGDS) http://www.lib.uconn.edu/ssgs/index.cfm
  - Example: Emory University Libraries Electronic Data Center http://einstein.library.emory.edu/
- Instruction (of library staff and of students and teaching faculty) in the use of library resources for obtaining data, using/searching metadata, some software instruction.
- Marketing data services to students and faculty
- Knowledge of data use on campus (who and which departments are using numerical data) and coordinating with those departments and other campus resources currently available (working with subject librarians or directly with departments).

Data Preservation

In addition to the above, which could be lumped together as reference and consulting data services for patrons who need to *use* data, there is an opportunity for the data services librarian to act as a Data Archivist (to use Jacob's terminology) and actively seek to preserve data. This could include ensuring consistent, perpetual access to purchased or subscribed resources (similar to other electronic resources issues).

Another direction this could take is the coordination and management of a repository for data created or owned by campus researchers. This particular aspect of the Data Librarianship question is important because it has the potential to involve many other sections of the OSU Libraries, including but not limited to Library Technology and Technical Services but likely others as well. Creating a secure "place" to store the data as well as developing meaningful metadata are just two parts to the larger issue of data preservation. In addition, Jeremy Frumkin (in a brief conversation) pointed out that assessment of campus-wide needs in this area should be done before designing data preservation or storage services.

Sources Consulted:

In addition to notes from Jim Jacobs' presentation at Willamette U in March 2007, the following resources were consulted.

Bennett, T & Nicholson S. (2004). Interactions between the academic business library and research data services. *Portal*, 4(1), 105-22

Gerhan, D. (1999). When Quantitative analysis lies behind a reference question. *Reference and User Services Quarterly*, 39(2), 166-76.

Jacobs, J. & Humphrey, C. (2004). *Preserving research data*. Retrieved May 25, 2007 from http://3stages.org/jj/w/preserving_research_data.html.

****Read, E. (2007). Data services in academic libraries: Assessing needs and promoting services. *Reference and User Services Quarterly*, 46(3), 61-75.**
Comment: This article provides a very good analysis of the reference portion of data services librarianship. It is based on a survey of data use at the University of Tennessee and seems to address every *reference* issue encountered in other articles consulted for this write up on data services. It includes a "Skills and Knowledge" section. It does not address collection development or data preservation.

Steinhart, G. (2006). Libraries as distributors of geospatial data: Data management policies as tools for managing partnerships.

Position Announcements consulted:

Geospatial Data Librarian, Emory University
(http://web.library.emory.edu/services/hr/geospatial_data.html) (October 2006)
Map and Data Services Librarian, University of Illinois - Chicago
(http://www.uic.edu/depts/lib/admin/personnel/map_lib.pdf)
Social Sciences Data Librarian, University of Southern California
(http://www.usc.edu/libraries/jobs/librarians/documents/223DSocSciData.pdf)
Data Librarian, University of Notre Dame
(http://www.library.yale.edu/~llicense/ListArchives/0010/msg00015.html) (October 2000)
Social Sciences Data Librarian, Rutgers University Libraries
(http://www.libraries.rutgers.edu/rul/hr/libpersonnel/APP171.pdf) (April 2006)

Social Science Data Librarian, Yale University Library (http://data.uwindsor.ca/cgi-bin/iassist/job.cgi?jobid=7) (May 2005)

Appendix B: Dataset Management and Curation Survey

Introduction and Definitions
Thank you for agreeing to complete this survey. The Oregon State University (OSU) Libraries is in the process of articulating a dataset curation and preservation strategy for our digital repository, the ScholarsArchive@OSU. To help us guide our efforts we are requesting that you fill out a questionnaire about your institution's experience to date. Our specific interests include user and contributor services, collaboration and partnerships, technology, and impact on staffing and management.

For clarification:

Datasets of interest are those which are "born-digital" (whether spatial, numeric, etc.) and in general require some software for analysis and manipulation.
Contributors refers to those individuals or groups who are depositing datasets.
Users refers to those using the datasets once they are stored/perserved.
Audience refers to both of these groups but may in your institution have additional meanings.
Dataset Preservation refers to storage and metadata assignment needed to access datasets.
Dataset Curation refers to contribution, preservation and services needed to provide access (and

If you have questions please feel free to contact ruth.vondracek@oregonstate.edu

Section 1. How would you characterize the extent of dataset preservation/curation by your library?

- ☐ Extensive
- ☐ Well underway
- ☐ In the early stages
- ☐ In the planning stages
- ☐ No plans to preserve datasets at this time

Please comment as needed:

Section 2. User services and dataset curation/preservation.

In this section we are interested in getting an idea of the extent of datasets curation undertaken to date and services provided to users of these datasets.

1. To whom do you offer data curation/preservation services?

- ☐ Faculty
- ☐ Students
- ☐ Specific Departments
- ☐ All Campus

Other (please specify):

2. What preservation/curation services are provided to contributors?

- ☐ Long term storage
- ☐ Metadata assignment
- ☐ Location aids for users
- ☐ Dataset manipulation software
- ☐ Collaborative online workspace
- ☐ Other services

Please elaborate as needed:

3. Do you provide users with copyright information and/or citation guidelines for curated datasets?

- ☐ Yes
- ☐ No

Please describe and/or provide URL if appropriate

4. What is the collection scope for curated datasets? If you can provide a URL with this information, please do so.

5. Have you encountered issues of rights managements in dataset curation/preservation? If so, how have these been resolved?

Section 3. Collaboration/Implementation and dataset curation/preservation

In this section we are interested in learning more about partnerships and collaboration between the library and other units on or off campus that led to dataset curation/preservation.

1. Did you have a pilot project with a specific group of dataset contributors?

- ☐ Yes
- ☐ No

Please comment:

2. What departments, disciplines, other groups have deposited datasets to date or seem most interested in doing so?

3. Who were the early adopters? If you have thoughts on why, please clarify.

4. Is dataset contribution mandatory or voluntary?

- ☐ Mandatory
- ☐ Voluntary
- ☐ It depends

5. Have you documented or assessed whether posting research datasets has facilitated scholarship and/or career advancement for the dataset contributor?

- ☐ Yes
- ☐ No

Please elaborate as needed:

Section 4. Technology and dataset curation/preservation

In this section we are interested in how you have handled the technology needed to preserve datasets at your institution and limitations (if any) that available technology places on this activity.

1. Where are "curated" datasets stored? (Check all that apply).

- ☐ On library servers.
- ☐ On institutional repository servers.
- ☐ On separate computing or information technology department servers.
- ☐ On commercial service servers.

- [ ] Other

Please comment for clarity:

2. What categories of datasets are accepted for preservation/curation?

- [ ] Archival (closed)datasets
- [ ] Versioned datasets
- [ ] Active/dynamic datasets
- [ ] Datasets associated with ETDs
- [ ] Datasets associated with current faculty or student research
- [ ] Other

Please comment for clarity on "other":

3. Are your dataset preservation/curation services integrated with other library services or programs (e.g. ETDs, the university IR, digital projects, etc.)?

- [ ] Yes
- [ ] No

Please describe:

4. Who actually "deposits" the datasets for curation/preservation in the designated repository?

- [ ] Library staff.
- [ ] Dataset creator/provider.
- [ ] Other.

Please comment for clarity on "other":

5. Have you encountered problems with interoperability (e.g. incompatibility with systems or formats) reported by either contributors or users?

- [ ] Yes
- [ ] No

Please describe how you have/do resolve these issues.

6. What have you done to enable/enhance third party discovery of curated datasets?

Section 5. Staffing and Management for Dataset curation/preservation

We are interested in how you have handled issues related to staffing and management for dataset preservation/curation including expertise and professional development.

1. Please describe the aspects of dataset preservation/curation for which the library is responsible.

2. Please describe the aspects of dataset preservation/curation which are the domain of other units on or off campus.

3. Who is responsible for these aspects of dataset curation?

	Library staff	IT staff	Dataset provider	Other
Storage:	☐ Library staff	☐ IT staff	☐ Dataset provider	☐ Other
Description:	☐ Library staff	☐ IT staff	☐ Dataset provider	☐ Other
Data set user services:	☐ Library staff	☐ IT staff	☐ Dataset provider	☐ Other

Please comment for clarity on "other":

4. What expertise was critical in setting up your dataset curation services?

5. In creating this service, what additional expertise (not at your organization) was needed?

6. How many positions and/or what level of staffing is involved with managing and curating datasets?

7. How did/do you accommodate staffing needs for dataset curation/preservation?

☐ New positions were/will be funded.

☐ Old positions were/will be "repurposed."

8. Has your dataset curation program been successful? How would you characterize it?

Appendix C: List of Universities

(Libraries in bold were sent survey)

Library Institution	**Institutional Repository (Y/N)**	**Datasets in IR (Y/N/?)**	**Dataset curation or preservation activity noted? (Y/N/?)**
Australian National U.	Yes	Dataset is a format "type" in the IR	Yes - survey
Baylor (GWLA)	Yes	No	No
BYU (GWLA)	Yes	No	No
Colorado State U (GWLA)	Yes (Ex-Libris)	No	No
Cornell	Yes	Yes	Yes (metadata services)
Emory	Yes -- ETDs only so far	No	Data Service Center (mostly reference services not preservation)
Iowa State U (GWLA)	No	No	No
Johns Hopkins	Yes	No	No
Kansas State U (GWLA)	Yes	No	No
MIT	Yes; DSpace	Yes	2 types of preservation; bit & functional
Monash (AU)			
yes, ARROW repository			
North Carolina State U	Scholarly publication repository	Cannot get in full text, don't have the right. articles from journals	reference service, preserve according to standard digital preservation practices
Oxford	Yes, Digital repository	Not yet	Yes, excellent planning for infrastructure, documentation on preservation and curation

Purdue	Yes, E-scholar	E-data is setup, but cannot find any data	Yes
Rice University (GWLA)	Yes	Has a separate GIS, data center	some, in Text coding under FAQ
Rutgers	Yes, community repository	Yes, also has a separate data center	Yes
Stanford	Yes	No	Yes
Texas A&M (GWLA)	yes	no	yes
Texas Tech (GWLA)	yes	no	yes (services to faculty but not necessarily dataset specific)
U. Arizona (GWLA)	yes	not yet	seems to be in planning stages this
U Colorado at Boulder(GWLA)	Not yet but noted in notes for 3/2008 meeting of faculty assembly re: NIH mandate	No	No
U. Conn	Yes	Seems no	No
U. Hawai'i(GWLA)	Yes (since 2007)	No (not yet?)	Not really
U. Illinois	yes	YES (some trials data in Excel and a SAS output file)	(Yes?)
U. Kansas (GWLA)	yes	yes (GIS Datasets in a collection)	(Yes?)
U. Melbourne	yes	not now but mentioned in collection policy	yes
U. Minn	yes	no but mention of "compilation of university data" in content guidelines	no but nice chart of preservation support provided and other data identification services provided
U. Nebraska (GWLA)	Yes	No	No
U. Notre Dame	Yes	No	No
U. Oregon (GWLA)	Yes	No	Yes?- recently closed a position announcement for a Data Services Librarian
U. So. California (GWLA)	Yes	No	No
U. Utah (GWLA)	Yes	Yes (their IR indicates	Yes if counting the mention of

		datasets are welcome - did not find any however)	adding datasets to the IR.
U. Washington (GWLA)	Yes	Yes (their IR indicates datasets are welcome - did not find any however)	Yes, in their Vision 2010 doc they mention progress towards this service
Washington State U. (GWLA)	Yes	No	No
Washington U. (St. Louis) (GWLA)	Yes	No	No
Yale	Yes, but pretty limited	No	No
Caltech	Yes	Yes (BLOB)	Yes
Scripps Institute Of Oceanography (SIO)	Yes	Yes	Yes
Woods Hole Oceanographic Institute	Yes	Yes	Yes