Steps Toward Implementation of Data Curation Services at Oregon State University Libraries

Michael Boock
Faye Chadwell

Report Presented to the Library Administration, Management and Planning July 2010

I. Summary

At the 2010 Library Administration retreat, Faye proposed that the Libraries “become a strategic partner in the research endeavors of OSU, especially related to e-science.” Karyle later charged Faye and Michael with investigating appropriate roles for the Libraries relating to data curation of the university’s research data, building on the Research & Innovative Services report published in February 2009: OSU Libraries and Research Dataset Curation. In this white paper, we describe work that Faye and Michael did in the last fiscal year, provide brief background information, touch on current data curation efforts at OSU and elsewhere and provide 4 broad recommendations for ways that OSU Libraries can support research and researchers at OSU by engaging in data curation-related activities in the coming academic year:

1. Investigate library capacity for data curation and develop local infrastructure and policies.
   a. Establish policies for the types of data that the Libraries will support and the levels of support that the Libraries will offer.
   b. Work with at least one college or academic unit to curate student research data vis-à-vis the final theses and dissertations in ScholarsArchive@OSU.
   c. Work with additional faculty to curate research data associated with their final research publications in ScholarsArchive@OSU.
   d. Investigate library management and assignment of DOIs to faculty datasets.

2. Assign data curation responsibilities to the new Center for Digital Scholarship and build staff capabilities to work with faculty on their data.
   a. Hire Data Services Librarian to report to new Center for Digital Scholarship department.
   b. Continue to work with GWLA to make professional development available to OSU Libraries staff.
c. Add data curation activities to charge of the Digital Repositories Work Group.
d. Build data curation workflows and train library faculty and staff to undertake data curation responsibilities.

3. **Identify methods for ensuring broadest possible access to university research data.**
   a. Continue to develop data dictionary for data sets in order to describe data so that it is understandable, harvestable, searchable and easily renderable using geospatial and other visualization tools.
b. Evaluate capabilities of current digital infrastructure to support curation of research data.
c. Ensure that appropriate Oregon related datasets captured in ScholarsArchive@OSU can be searched and/or rendered in Oregon Explorer.

4. **Continue to partner with others, both on and off campus, in order to meet the challenges of data curation for the university.**
   a. Continue to seek grant funding for tool creation in support of data curation.
b. Participate in Greater Western Library Alliance, MetaArchive West and Orbis Cascade Alliance data curation activities.
c. Terry, Faye and Michael meet with Associate Deans of colleges to describe data curation work that is being done in the Libraries and identify additional areas where we might assist them with research data produced within their colleges.
d. Identify opportunities to collaborate with other units on campus interested in data curation including the Graduate School, Institute for Natural Resources, Research Office, NACSE, College of Science, College of Forestry, HMSC.
e. Partner with the University of Oregon in the creation and use of data dictionaries, metadata standards, policies, needs assessments, faculty surveys, staff training and best practices related to data curation.

**II. Introduction**

**Why Data? And Why Libraries?**

Traditionally, academic libraries have acquired, cataloged, and preserved research after it is published, taking responsibility for cataloging and making the intellectual property available at the very end of the research cycle. Over the past couple of years, academic librarians have increasingly become involved with our users and the research they produce earlier in the research cycle, prior to publication. For example, faculty and students increasingly look to librarians for help retaining appropriate rights to their research, for help retaining copies of their research in
institutional repositories, and increasingly, for help with data curation (i.e. active management of research data for access and preservation).

There are a number of reasons for faculty to want to curate their research data. Funding agencies, publishers and institutions are increasingly going to require that data be managed, retained and made accessible. Data curation can expose data to the world and ensure its longevity and reuse in a world where “computational capacity and tools... are giving rise to new modes of conducting research.” Curation of research data allows research to be validated, taught and re-used for the creation of new science. Unique observational data that is impossible to recreate is retained. Librarian skills in organizing, describing, managing and preserving collections can be useful to the campus in managing research data. Libraries themselves may be able to serve as a central service point for helping faculty to manage research data.

Faculty Meetings

In March 2010, Faye and Michael invited select faculty from a variety of departments on campus and the library to two lunch meetings (see Appendix A for invitation and list of questions asked of faculty attendees). At the meetings, 15 attendees were asked a series of questions about their data and the libraries’ potential role in relation to that data (see Appendix B for list of questions). The questions were based on questions that Purdue University Libraries asks their faculty.

Overwhelmingly, attendees recognized the value of preserving their data and in some cases making it more widely available. Faculty mentioned an interest in knowing what data is being captured in other departments so that it can be repurposed, built on for new science, and not recreated. There was a general recognition that data is increasing in size and complexity, that there is no central place on campus for faculty to seek assistance in managing the data that is being produced, and that there needs to be such a place.

Several attendees were in the process of reformatting data on print cards to excel spreadsheets and had questions about how they set up columns and rows so that the data is understandable to other researchers. Other attendees mentioned that they need help figuring out what metadata standards and ontologies to apply to their data.

Concerns were raised about making data accessible to the world too soon. Faculty said that in some cases embargos would need to be put in place in order to assure that the research isn’t “stolen” prior to article publication. Others suggested that some research data should be made widely available as soon as possible, even before publication, and should have no restrictions to access placed on it at all. Faculty noted that there is often confusion about who owns data, especially data generated or collected as part of national grants and data collected by students.
Currently, some agencies and publishers don't allow citing of unpublished data and require a DOI for the data to be cited. Faculty expressed an interest in being able to assign DOIs to data and databases that faculty make available. There was agreement that faculty need to have a common understanding of publishing requirements and data citation rules and that the Libraries could help with this. Faculty also expressed a need to begin depositing data in repositories based on anticipated and existing publisher and grant agency requirements.

Attendees saw a need to store not only their own unique observational research data which cannot be duplicated but also that of their students conducting research for their theses, dissertations and masters research papers. All expressed a desire for the libraries to begin to do something and agreed that the libraries adopt an iterative approach to identify and select some important datasets that should be curated and work to preserve them and make them more widely available.

Terry Reese has had meetings with a number of faculty on campus and at HMSC regarding their data needs. His discussions with deans in particular indicate that at least some colleges are comfortable with their infrastructure for handling research data but look to the libraries for help with metadata and identifying other sources of data on campus.

In the faculty meetings that Faye and Michael initiated, research data associated with theses and dissertations was suggested as a good starting place for the libraries. This data can serve as a way for the libraries to learn how to do the work involved in curating data including: selection and appraisal; metadata creation and advising; conversion to archival formats; establishment and maintenance of authenticity and provenance; annotation and linkage to related resources; provisioning for secure and redundant storage; transformation, migration and emulation as needed over time; discoverability in contemporary search systems; creation of meaningful access mechanisms; and recontextualization.

**Data Curation Webinars**

One of Faye and Michael’s goals this year was to provide subject librarians and other staff with continuing education relating to data curation. We considered inviting Michael Witt from Purdue. Instead, in order to provide library faculty with more background, from a wider variety of people with a diversity of experience, we elected to offer a series of three data curation webinars, using the University of Utah’s Wimba webinar software. Michael received an opportunity grant from GWLA to pay honorariums to some of the top data experts in the country, including D. Scott Brandt from Purdue, Sayeed Choudhury from Johns Hopkins, Chuck Humphrey from Alberta, Amy Stout and Anne Graham from MIT and several presenters from GWLA libraries (see Appendix C for the webinar series announcement, full list of presenters and their qualifications and archived versions of the online seminars).
At the first webinar, Chuck Humphrey from Alberta and Scott Brandt from Purdue provided a big picture overview of data curation activities and libraries. Chuck provided an introduction to basic data concepts relevant to librarians such as how to differentiate research data from everything else that is digital, how lifecycle data management helps us better deal with data as a resource, how collection considerations remain important in managing data and how different levels of service can be defined for data. Scott Brandt described opportunities for librarians, in their liaison roles, to engage with faculty about their data management needs within the larger scholarly communication context (copyright, open access, publication and repository services).

The second webinar included examples of extant data curation efforts from subject librarian and digital librarian practitioners. Anne Graham and Amy Stout gave a brief background summary of what the MIT Libraries are doing and discussed issues surrounding starting a data management program. They also provided an overview of what libraries need to know before starting a data management program. They discussed the following topics from a subject librarian perspective:

- How does your library system operate?
- Are you better off starting your program from the top-down or using a grassroots approach?
- Making contact with your faculty and finding out what they need
- Learning about data and e-science
- Developing relationships across departments that will facilitate your offering of services to people outside the library system

Sayeed Choudhury discussed Johns Hopkins’ early experiences related to the Data Conservancy, one of two current awards they’ve received through NSF’s DataNet program. Choudhury spoke specifically about the types of data being considered for the early prototype development, the initial technical architecture, and the new library duties or skill sets that are being developed as a result of these activities.

At the third and final webinar on July 1, librarians from GWLA institutions provided information about the kinds of data curation activities with which their institutions are involved. Brian Westra from the UO talked about the faculty survey and needs assessment he developed. Terry Reese discussed the three grants he has been involved with relating to data, in particular an NSF grant with MIT and other institutions that would develop the DSpace software to better manage fluid datasets and better enable faculty to understand the kinds of data that is being compiled across campus. Other presenters - Daureen Nesdill at Utah, Holly Mercer at Texas A&M and DeeAnn Allison at Nebraska – discussed their library’s very early efforts at beginning to work with faculty to capture their data.

The third webinar also included a discussion of possible GWLA roles in relation to data curation. Ideas included:
1. the development of a web clearinghouse that might include the UO needs assessment, information and links pertaining to each institution's policies, standards, and emerging and best practices;
2. continuing education including GWLA hosting Purdue University workshop on data curation profiles;
3. development of a cloud-based consortial storage solution and/or hiring consultants to investigate options for shared storage solution.

III. Environmental Scan

The RIS report provides background on OSU Libraries’ interest in and involvement with dataset curation services, including work to aggregate geospatial natural resources datasets in Oregon Explorer and the use of ScholarsArchive@OSU as a data repository. As of July 7, 2010, ScholarsArchive@OSU contains 28 items identified as “datasets” or “GIS Vector Data”. This is either research data associated with theses and dissertations or data gathered by Geography graduate students relating to Oregon natural resources. Dawn Wright in Geosciences now requires that her graduate students submit their research data associated with their theses to ScholarsArchive@OSU. In addition, there are 194 items in a COAS Data Report Series and 49 items in a Marine Economics Data Sheets collection. These data in these collections are in PDF format so are not easily manipulable, however they have been downloaded thousands of times this year alone.

As we learned from the data webinars, a perusal of the literature, and discussions with librarians at ARL institutions known to be involved in data curation activities (Johns Hopkins, University of Minnesota, Purdue, Illinois), there are no truly mature data curation efforts happening at academic libraries in this country. Wendy Lougee, University Librarian and McKnight Presidential Professor at the University of Minnesota–Twin Cities, said in an email that “the [ARL] survey data didn’t reveal that anyone is truly engaged in curation. We, similar to a number of research libraries, are at an early stage wherein we have begun to engage in consultation/education for researchers related to good data practices.”

Some libraries are working on deposit of data to digital repositories but are not engaged in full data curation. Purdue has engaged in coordinated and relatively intense needs assessment. Illinois and Cornell have developed some infrastructure and have data deposit programs. Johns Hopkins has two NSF projects in the works currently: Data Conservancy and DataOne. The Data Conservancy will curate and provide services that support deposition of data, retrieval, description, and conservancy of data. DataOne “represents a new virtual organization that will enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it.”

Each of these institutions engages in some level of consultation services with faculty around data management.

IV. Data Curation Issues
Depending on the level of support our library chooses to offer, data curation can be an expensive enterprise. Variables to account for in costing out data curation services include the cost of acquiring, ingesting, assigning metadata, reformatting, storing and disposing of data. While storage costs have decreased, they are not inconsequential, particularly for large datasets. Streamed data storage costs can be particularly high because copies of multiple iterations may require storage. How do we make data curation efforts within the libraries sustainable even without additional funding? One way is to identify what faculty and graduate students most need and determine the libraries’ capacity to provide these services.

Examples of librarian activities relating to data curation could include: Initiating research data talks; Reviewing research workflows; Creating data curation profiles; Identifying metadata for discovery; Helping to organize data for dissemination; Promotion of ScholarsArchive@OSU for preserving and making research data accessible; Training on repository ingest. The use of data curation profiles can be a means of addressing the question, "Which researchers are willing to share data, when, with whom, and under what conditions?" Librarians, or other library staff involved in the conversations, will need to be familiar with how data can best be described, what contextual information is necessary for the data to be found and reused, intellectual property and access rules (who owns the data, terms of use, attribution), anticipated user support, workflow for ingest and maintenance.

Developing metadata practices relating to data curation is something that faculty have expressed an interest in. Adopting and recommending metadata practices and standards in the arena of data curation accomplishes “several important things: first, [metadata practices] must make data understandable by computers; second, they must support discovery across heterogeneous data collections; and third, they must manage all of this across data scales, from the small to the immense.”

Subject librarians working with their academic units will need to coordinate in new ways with staff in Digital Access Services and Emerging Technologies and Services. Liaison activities will need to be coordinated to ensure that appropriate IT, repository, metadata staff and subject librarians are involved in data conversations and decisions at appropriate times. The ability to work between IT, repository services, and metadata services will be important. When subject librarians talk to faculty about data they should probably be accompanied by someone with technical expertise, for example, Michael Klein, Terry Reese, Michael Boock or a new data services librarian.

While traditional librarian duties such as reference and collection development are being reassigned or redesigned to require less time, there are many new roles and responsibilities relating to digital scholarship emerging. Based on library goals, It will be important to identify what can be reasonably accomplished, the amount of staffing required, and what is expected of library faculty in light of other expectations and responsibilities.
Scientists are often concerned about misuse of their data. The libraries can explain to faculty that the validation of future uses is the responsibility of peer reviewers and not the original compiler. Other faculty are not only interested in managing and making their research data accessible but also interested in rendering it using visualization tools such as Google Maps or Oregon Explorer and providing value-added web interfaces that allow searching, browsing and geospatial rendering.

The RIS report noted that Oregon Explorer partners continue to express a need for “expert retrieval, manipulation of, if not storage of, datasets produced outside of OSU.” Clearly the library doesn’t have capacity to create and maintain web front-ends for every dataset or database on campus or from Oregon Explorer partners. The library should develop some criteria for determining when we will do this work and under what circumstances. For example, data services for Oregon Explorer partners will continue to be provided for a fee, but what about data services for data produced by OSU?

V. Conclusions

One size of data curation services does not fit all. What we do must depend on our capacity and mission. Individual libraries won’t be able to offer a full suite of data curation services on their own. Melissa Cragin from the University of Illinois advises that “depending on a library’s resources, [libraries] choose a starting point that makes sense and prepare for strategic growth based on changing needs and expectations.” Successful data curation requires buy-in. There needs to be vision for the importance of data curation at the very top, action plans incorporated in the strategic plan to move goals forward, unit plans and PD language to include data responsibilities.

We recommend that OSU Libraries begin to manage data associated with faculty final publications and theses and dissertations and to continue to render Oregon-based datasets in the Oregon Explorer as a fee-based service. Librarians should be prepared to recommend data curation options for faculty even if the Libraries aren’t managing. Librarians should be available to answer questions about intellectual property pertaining to faculty research data, explain options for describing the data so that it is findable, understandable and reusable by other researchers, and sharing information about metadata standards, ontologies and subject-based repositories that are available for preserving data.

Liaison activities will increasingly include instruction on searching for relevant controlled vocabularies and ontologies for faculty to use in describing their data, identifying and reviewing options for alternative dissemination (e.g. are there regional, disciplinary, or local repositories in which their work should appear in addition or instead of the publisher’s repository); understanding Dublin Core, and examining research workflows to identify possible metadata requirements.
This work is similar to what librarians have always done. In a liaison role, academic librarians often interface with departments to identify service needs. Librarians might attend faculty and graduate student research seminars to learn about the data associated with research projects. Librarians should also talk one-on-one with faculty, perhaps using data curation profile techniques, to understand the issues with which faculty are dealing and, for faculty who are interested, to promote better long-term storage, preservation, management and broader dissemination of their data.

Appendix A - Faculty Attendees at Faculty Data Curation Meetings, March 4 and March 18, 2010

Kathy Howell – Forestry
Christopher Romsos – Geosciences
Theresa Valentine & Suzanne Remillard – USFS
Roger Nielsen – Geosciences
Scott Baker – Fisheries and Wildlife – Dialing in from Guin via Polycom
Brian Sidaulskis – Fisheries and Wildlife
Sarah Henkel – HMSC
Mark Myers - Geosciences
Dawn Wright – Geosciences
Rich Holdren – Research
Bonnie Avery – Libraries
Michael Klein – Libraries
Janet Webster – Libraries
Philip Vue – Libraries
Michael Boock – Libraries
Faye Chadwell – Libraries
Appendix B - Faculty Data Curation Meetings

March 4 and March 18, 2010
Faye Chadwell – Associate University Librarian for Content and Collections
(Faye.chadwell@oregonstate.edu)
Michael Boock – Head of Digital Access Services
(Michael.boock@oregonstate.edu)

For the last few years the library has been considering its role in curating, describing, and disseminating research data that is gathered and maintained by faculty at OSU. Michael Boock and I are speaking to faculty with an interest in data curation to find out more about the data you collect and to hopefully discern an appropriate library role in relation to that data. Subject librarians gave your names to us.

We have scheduled two lunch meetings in the Drinkward room on the 4th FL of the Valley Library. This room has videoconferencing with Guin Library at HMSC if that is needed. We will ask the same series of questions at each meeting. If you are interested in meeting with us, please indicate which of the two meetings you plan to attend. If you are unable to attend, please feel free to forward this to a colleague and ask that they RSVP. We will have pizza, salad and soda from American Dream.

Thursday, March 4th from 11:30-1pm
Thursday, March 18th from 11:30-1pm

These are the questions we will ask:

1. What form and format are your data in?
2. What is the expected lifespan of your data?
3. How could your data be used, reused, and repurposed?
4. How large is your dataset, and what is its rate of growth?
5. Who are potential audiences for your data?
6. Who owns the data?
7. Does the dataset include any sensitive information?
8. What publications or discoveries have resulted from the data?
9. How should the data be made accessible?

If not already clear from previous responses:

10. How could the library best assist in the preservation and access to your data?
11. Are there concerns with the library taking on these responsibilities? What things need to be in place before the library should take this on?

Appendix C- GWLA Data Curation Webinar Series

Announcement and Schedule

Data Webinar Series Announcement

The Greater Western Library Alliance is providing a series of three free webinars on the topic of data curation. The webinars will be conducted using "Wimba" webinar software. No pre-registration is necessary. The webinar login address is: http://uutahlive.wimba.com/launcher.cgi?room=GWLA_DCC. In the Participant Login "Name" field type your name. Click the "Enter" button. The Wimba Classroom interface will load, and you will automatically enter the meeting room.

Test Your Computer:

Before attending, please check your system with the following link: http://uutahlive2.wimba.com/wizard/launcher.cgi. If you encounter any problems while running the Set-Up Wizard or joining the event, please contact the Wimba 24X7 Helpdesk at 866.350.4978 or http://technicalsupport@wimba.com.

Each of the webinars will be recorded. After each session, a URL will be posted to mailing lists where the series was announced.

Presentation Details:

I. Big Picture Overview of Data Curation, Thursday May 6, 12-1:30 EDT

Archived version

Over this decade, increasingly more people are viewing research data as an asset requiring proper management and long-term stewardship. This outlook is a major cultural
shift from the perspective that knowledge outputs such as journal articles and books are the sole treasures of research. One consequence is that librarians in research institutions are now having to consider how to incorporate data as a library resource. Chuck Humphrey's presentation provides an introduction to basic data concepts relevant to librarians. Topics to be discussed include how to differentiate research data from everything else that is digital, how lifecycle data management helps us better deal with data as a resource, how collections remain important in managing data and how levels of service can be defined for data.

It is important to understand data curation within the larger scholarly communication context, and then to identify opportunities and capacities where librarians can and should find a role to engage. Given a definition that ranges from managing to archiving to preserving data along the data lifecycle, there are various points where data curation services can be pursued by librarians: at a point of research initiation (articulating the problem and pursuing funding), at a point of recent or ongoing research (organization within the lab), at a point where a larger community needs to be engaged (broadening access), and at a point where time scale is important (archiving and preserving in a repository). D. Scott Brant will discuss the role of librarians in pursuing and engaging in these data curation activities with specific examples presented.

Speakers:
Charles (Chuck) Humphrey has been the Head of the Data Library at the University of Alberta since 1992 and began a data library service in 1980 in the University's academic computing centre while employed as a statistical consultant. In 2000, Mr. Humphrey also assumed responsibility for the implementation and management of a Statistics Canada Research Data Centre (RDC) at the University of Alberta, which is a data enclave for Statistics Canada confidential data. As the Academic Director of the RDC, he oversees the operations of this facility and serves on the RDC National Coordinating Committee.

D. Scott Brandt is a professor of library science and associate dean for research in the Purdue University Libraries. Primarily he helps guide the Libraries' research and facilitates participation in sponsored funding (e.g., NSF, IMLS, local seed grants)---since April 2005, Purdue librarians have participated in more than 70 grant applications with more than 80 faculty across campus. As acting director of the Distributed Data Curation Center, he oversees investigation into curation issues of organizing, discovery and access to, and archiving research data in complex environments. Prior to arriving at Purdue in 1993 he was associate head of the Science and Engineering Libraries at MIT, and is the author of Teaching Technology (2002) and Unix in Libraries (1991).

II. What kinds of data are libraries managing, how are they doing it and with what staff?, Tuesday June 1, 12-1:30 EDT
Archived version

After giving a brief background summary of what the MIT Libraries are doing, Anne Graham and Amy Stout will discuss issues surrounding starting a data management program and provide an overview of what libraries need to know before starting a data
management program. They will discuss the following topics from a subject librarian perspective:

- How does your library system operate? Are you better off starting your program from the top-down or using a grassroots approach?
- Making contact with your faculty and finding out what they need
- Learning about data and e-science
- Developing relationships across departments that will facilitate your offering of services to people outside the library system

**Sayeed Choudhury** will discuss early experiences related to the Data Conservancy, one of two current awards through NSF's DataNet program. Choudhury will speak specifically about the types of data being considered for the early prototype development, the initial technical architecture, and the new duties or skill sets that are being developed as a result of these activities.

**Speakers:**

**Anne Graham** is Civil and Environmental Engineering Librarian and GIS Liaison and **Amy Stout** is Computer Science Librarian at MIT Libraries. Both Anne and Amy have been working on starting data services at the MIT Libraries. This includes educating faculty, students and lab managers about "best practices" for data management as well as initiating projects that encourage faculty to look to the libraries for the stewardship of their data.

**G. Sayeed Choudhury** is the Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University. He is also the Director of Operations for the Institute of Data Intensive Engineering and Science (IDIES) based at Johns Hopkins. He is a Lecturer in the Department of Computer Science at Johns Hopkins, a Research Fellow at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign and a Senior Presidential Fellow with the Council on Library and Information Resources. He is a member of the ICPSR Council and DuraSpace Board.

**III. What is happening at GWLA Libraries and what are next steps for GWLA?, Thursday July 1, 12-1:30 EDT**

**Archived version**

Presenters from several GWLA libraries will provide 5-10 minute presentations on what is happening at their institutions, at what stage they are at with data curation, and what they see as a possible GWLA role. Presentations will be followed by discussion about possible next steps for GWLA and potential areas of collaboration among GWLA libraries regarding data curation.

**Speakers:** GWLA members involved in data curation: DeeAnn Allison (University of Nebraska); Holly Mercer (Texas A&M); Brian Westra (University of Oregon); Daureen Nesdill (University of Utah), Terry Reese (Oregon State University)
1 Cornell University, Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library. http://ecommons.cornell.edu/handle/1813/10903
ii Michael Witt and Jake Carlson, Conducting a data interview. http://docs.lib.purdue.edu/lib_research/81/
iii Clifford Lynch and Sayeed Choudhury, Initiatives from the NSF’s DataNet Program: DataONE and the Data Conservancy https://jscholarship.library.jhu.edu/handle/1774.2/33953
iv Purdue University and University of Illinois, Investigating Data Curation Profiles Across Multiple Research Disciplines. http://wiki.lib.purdue.edu/display/dcp/Purdue-UIUC+Data+Curation+Profiles+Project