Growing National Concern for the Welfare of Research Data

Data is an essential component of scholarly research. Recent advances in computation and networking tools allow for data to be mined and analyzed, verified, reproduced, and shared among researchers at any stage in the research process. Because of this, data management is becoming a critical piece of the research process.

The last decade has seen a number of data reports and initiatives come out of both the national and international governmental arenas. In addition, the media has called for the technical and organizational infrastructure to support, maintain, and preserve research data, particularly publicly-funded research data. For example:

- Revolutionizing Science and Engineering Through Cyberinfrastructure (NSF, 2003 – The Atkins Report);
- Principles and Guidelines for Access to Research Data from Public Funding (OECD, 2007);
- Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (NAS, 2009); and
- Special Issues on data in Nature: “Big Data” (Nature, 2008) and “Data Sharing” (Nature, 2009), and
- A special issue in The Economist on data, the “Data Deluge,” (The Economist, 2010).

Further, both journal publishers as well as funding agencies are explicitly addressing the need for data management as a condition or expectation for publication or funding. Most recently, the National Science Foundation released its Dissemination and Sharing of Research Results policy, which requires all proposals to include a data management plan.

In response to this global movement toward the systematic management of research data, research libraries around the nation have been adapting their long-developed skills in cataloging, discovery, licensing, copyright, and preservation to provide services to the researchers they support. For example:

- MIT offers Data Services for the Social Sciences;
- Cornell has a suite of data services for their campus research activities, including a Data Staging Repository;
- Oregon State has expanded their institutional repository to accommodate data.

The University of Massachusetts Amherst Libraries Approach

The Libraries’ primary interest in data curation stems largely from these national developments regarding the infrastructure for data-intensive research and access to research data, rather than from an explicit on-campus demand for services. Therefore the Libraries’ preliminary explorations in this area are built on a foundation of well-intentioned assumptions about faculty and their research data practices. These assumptions require local verification before the Library will be in a position to outline
and deliver solutions to perceived problems and needs surrounding research data. At this time, these problems and needs are not adequately articulated either within the Libraries or elsewhere on campus.

The University of Massachusetts Amherst claims 82 centers and institutes that are conducting research on some level, bringing in over $140 million in externally sponsored research each year. Our research output has earned the campus classification as a Research University with Very High research activity (RU/VH) by the Carnegie Foundation. This includes 10 federally-funded research centers that “reflect national priorities and provide opportunities for breakthrough research, educational innovation, and technology transfer.” (http://www.umass.edu/umhome/research.php).

Administrative recommendations to faculty regarding the curation of research data are limited to the 2006 Faculty Senate Policy on Data Ownership, Retention, and Access at the University of Massachusetts Amherst. Further, the policy lacks any specific guidelines or resources to back up its key recommendations: ensuring data quality, retaining data for a minimum of three years, and making data publicly available quickly and to the extent possible.

The UMass Amherst Libraries Data Working Group is making the following assumptions based on information about research output and past campus attention to data curation:

1. UMass Amherst research centers are producing data.
2. Those charged with managing the data of an existing project may not have the expertise to do so, nor are there any organized resources to assist them if they do not.
3. There appears to be little incentive and no support for researchers to manage data beyond the expiration of its funding cycle.

The Libraries’ mission to be a “key partner in teaching, learning, and research at UMass Amherst” and to “maintain a rich information environment, facilitate access to it, and create a hub of campus and community scholarly activity” makes it an obvious choice for providing guidance to those researchers who require support, resources, and possibly the infrastructure for managing their research data.

Therefore, in an effort to determine what, if any, role the Libraries can play in supporting data-intensive research and the curation of research data, the Libraries recently established a Data Working Group (DWG) to articulate a vision and make explicit recommendations to the Libraries’ regarding research data. The Data Working Group has outlined the following, exploratory activities:

- Education on the issues involved with data curation
- Understanding of the University’s current research environment and data outputs
- Evaluation of current Library practice for supporting active and archival data
- Exploration of partnerships for data curation both in and outside of the University
- Propose interim steps to assist the University in meeting its own policies
- Create a vision of data curation for the Libraries
- Make recommendations for future activities related to data

Understanding the University’s Current Research Environment

Understanding the University’s current research environment and data output is crucial in building strong foundations for a vision of data curation and support for the Libraries. In order to develop this
understanding, a dialog with those faculty and graduate students who are securing grants for and conducting data-driven research on campus must be established. To this end, the DWG conducted informal interviews with faculty during August and September, and held a focus group for graduate students during Open Access Week in October 2010. These discussions have helped us begin to appreciate the diversity of issues that are inherent in data management and curation activities.

**Faculty Interviews**

During August and September 2010, the DWG invited a handful of faculty from various disciplines to participate in short, casual interviews regarding their data management practices. The interviews were carefully constructed as an “exploration” (see Appendix A) and were limited to a small number of faculty with whom liaisons or DWG members have a relationship or are perceived to be supporters of the Libraries. The DWG felt strongly that care should be taken with this initial outreach to faculty for the following reasons:

- Given the diversity of research domains, their traditions, and approaches to data management, acquiring a holistic view of campus-wide data practices requires diplomacy and a non-judgmental approach;
- Any implications for data sharing ought to be minimized to avoid misunderstandings between faculty and the Libraries regarding the nature of our interest in their data management practices;

Seven of thirteen invited faculty participated in the interviews, from the following disciplines:

- Computer Science
- Polymer Science
- Natural Resources/Public Policy
- Civil and Environmental Engineering
- Psychology (x2)
- Geology

The interviews took an average of 45 minutes to complete and were conducted by two DWG members at the convenience of the faculty member being interviewed. Like the initial solicitation, the interview questions were crafted to be non-presumptive with regard to faculty practices, expectations, or disciplinary standards (see Appendix B).

In general, each faculty member has a very unique story to tell about their research data and how they collect and store it. At the same time—though these interviews should not be considered representative of the entire research body on campus—they represent some of the larger issues surrounding data management as a whole (summarized in Table 1).

The DWG found that many faculty strategies for managing research data are project-specific, organic procedures that do not take long-term preservation, data re-use—local or more broadly-conceived—or the application of metadata standards into account. At the same time, faculty are actively generating large quantities (where the definition of “large” varies between disciplines) of complex data, much of which is unique and difficult or impossible to reproduce.
Many faculty interviewed employ ad hoc mechanisms for data storage and back up and utilize the infrastructure that is available to them through their departments or OIT on a piecemeal basis in so far as those systems will meet their research needs. In those circumstances where local infrastructure is or is perceived to be inadequate, faculty will employ external solutions to conduct their research.

Faculty rely on their graduate students to curate and maintain their data. However, few faculty members interviewed provide formal training or guidelines for their researchers to follow regarding the creation, collection, documentation, or storage of their data. In cases where researchers are asked or required, for example in patent applications, to reproduce their findings, the lack of local guidelines and documentation has been problematic.

Those projects that do employ formal guidelines and procedural documents for data management are projects required to comply with Institutional Review Board (IRB) standards for the ethical conduct of research. Where external pressures such as this are not immediate, guidelines and procedures are much less formalized if they exist at all.

Sharing data for collaborative research projects was noted as being a concern for some, particularly with regard to the challenges of providing cross-institutional access to secure, institutional servers for research collaborators.

Apart from project-specific, collaborative activities, perspectives on public data sharing range from the very conservative—where data is not shared at all outside of formal publication—to the fairly liberal—where there is intent to share data after it has been used for publication. Some of the biggest constraints to data sharing are privacy for sensitive data and intellectual property for technologies that might yield patents or other profitable licenses. Where these issues are not relevant, faculty express an interest in sharing data for legacy purposes.

<table>
<thead>
<tr>
<th>Table 1: Data Management Issues by Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data storage &amp; Infrastructure</strong></td>
</tr>
<tr>
<td>• Size</td>
</tr>
<tr>
<td>• Storage/capacity</td>
</tr>
<tr>
<td>• Computational power</td>
</tr>
<tr>
<td>• Preservation of proprietary programs/file formats</td>
</tr>
<tr>
<td>• Various infrastructure ownership scenarios</td>
</tr>
<tr>
<td><strong>Procedures &amp; Training</strong></td>
</tr>
<tr>
<td>• Workflow routinization</td>
</tr>
<tr>
<td>• Knowledge transfer</td>
</tr>
<tr>
<td><strong>Documentation &amp; Metadata</strong></td>
</tr>
<tr>
<td>• Project and discipline-specific practices</td>
</tr>
<tr>
<td>• Impact of external requirement on practices</td>
</tr>
<tr>
<td><strong>Data Reuse and Sharing</strong></td>
</tr>
<tr>
<td>• Making data public vs. making data useful</td>
</tr>
<tr>
<td>• Collaboration for publication vs. collaborative projects</td>
</tr>
<tr>
<td>• Occasional need to reproduce research</td>
</tr>
<tr>
<td><strong>IP &amp; Data Sensitivity</strong></td>
</tr>
<tr>
<td>• Privacy/IRB constraints</td>
</tr>
<tr>
<td>• USPTO rules for confidential exposure</td>
</tr>
<tr>
<td>• Emulation and Post-publication sharing</td>
</tr>
</tbody>
</table>

The overall picture of current data management on campus, as indicated by this preliminary survey, is that it ends at publication, and faculty data management strategies vary widely up until the point of publication. Post-publication data management or access seems to be a consideration only when there
is a specific request to verify or reproduce the study. Only one faculty member interviewed intended to make data publically available.

Graduate Student Focus Group

“I Can Has Data?” a focus group for graduate students hosted by the DWG, was billed as one of six events organized by the University Libraries for Open Access Week 2010. A partnership with the Graduate Office helped to disseminate material about the event and enabled us to bring together 17 participants from the following disciplines:

- Mechanical Engineering
- History
- Education
- Polymer Science
- Chemical Engineering
- Sociology

Like the faculty interviews, the Graduate Student Focus Group was intended to be a casual and non-presumptive conversation with students about their current data management practices (see Appendix C). To our surprise, the students who attended were primarily interested in learning about the services that the Library had to offer to facilitate data management.

As with the faculty, graduate student data management practices vary across disciplines. Some students maintain paper laboratory notebooks, while others have built their own hardware infrastructure and run Subversion to manage their projects. While it is clear that these students are primarily responsible for the collection, documentation, and management of data for their research projects, they report no formal training on this topic within their departments or research groups. Most learn on their own or from their colleagues and lab partners. In one discipline, it was stated that they “hope they pick well” when choosing a mentor whose data they will likely inherit.

Although graduate students indicate that they would consider public data sharing—and in one instance reported submitting data to publishers along with manuscripts as a standard, disciplinary practice—their primary concerns for data management centered on data storage and backup, versioning, and security.

Overarching Data Needs

The following items have emerged through the DWG’s exploration as the primary needs regarding research data among the faculty and graduate research population:

- Infrastructure for data storage and backup (centralized or alternatively-conceived)
- Training in best practice for data management
- Consultation Services (particularly regarding the NSF mandate)

Potential Roles for the Library
Based on the information gathered in the interviews and focus group, there are a number of services that the library could develop to facilitate and support data management practices on campus:

1. Guidance in the identification of appropriate data repositories for the archiving of large-scale data sets and associated research outputs, such as the ICPSR or Dryad, and assistance with material deposition.
2. Consultation on metadata and standards for format and content of data, policies for data sharing and accessibility, and plans for long-term access and preservation of data sets.
3. Provision of a globally accessible and widely indexed online location for faculty's research outputs, including persistent URLs and searchable metadata, through ScholarWorks, the University's Institutional Repository. ScholarWorks accounts are available to faculty at no charge and with minimal size limitations.

Librarians and staff with diverse backgrounds will be affected by the new services. Because data management is a relevant issue for numerous disciplines affecting a large campus constituent, and because there are many facets to data management, the Data Working Group envisions possible roles for liaisons, metadata experts, systems staff, and scholarly communication staff. The Data Working Group is developing a workflow and formal process for librarian and staff support.

Conclusion

As a key partner in teaching, learning, and research, the University of Massachusetts Libraries are committed to extending our services to include data management support. The Data Working Group believes that data should be recognized by UMass Amherst as part of the institution’s overall research product and that appropriate incentives should be created for the management and dissemination of data. The Libraries are a central hub on campus, have well-established connections with numerous departments, and have expertise in issues of description, organization, access, dissemination, and preservation, which makes them a logical choice for data support. While the libraries are pleased to be able to support researchers with infrastructure and training services, we will continue to explore new models and trends in data management, and may offer additional services as needs arise and if resources allow.
Appendix A: Invitation to Participate in Data Interview

Dear ___:

In response to a number of national reports and initiatives on data (see for example Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age), including the National Science Foundation's recent announcement about implementing a policy that will require proposals to include a data management plan, the library has formed a working group to explore current data management activities on campus in an effort to identify potential services that could be provided to faculty going forward. This working group is conducting short, informal data interviews with faculty this summer to begin to identify local needs.

Do you produce data as a part of your research and would you be willing to participate in one of these data interviews at a convenient time this summer? The interviews will likely take 30 - 45 minutes. We would be willing to share the survey results with you if you are interested in them.

Many thanks,
Rebecca

Rebecca Reznik-Zellen
Chair, Data Working Group
Integrated Sciences and Engineering Library
University of Massachusetts Amherst
413.577.0912
Appendix B: Faculty Interview Questions

1. Tell us about the project that you are working on.
2. Tell us about the data you’re creating/collecting. How is it generated? What guidelines do you follow, if any? Do you work with other researchers to generate your data? Are you building on existing or established data?
3. How is the data stored and what file formats are you using? (Software, hardware?)
4. How large is the data set and what is its growth rate? What is the expected lifespan of the dataset? (If lost, is it difficult to replace?)
5. How do you use the data? Do you expect to reuse or repurpose it in the future? Is the data used by other projects/researchers (IE: visualization, analysis, simulation)?
6. What are we not asking you about your data that we should know?
7. Would you mind if we contacted you again as we work through this process?
Appendix C: Graduate Student Focus Group Questions

Overview

Introductions – Name, department, status, 1 minute elevator pitch about their data/research

What is data management?

Data management describes how you plan for creating, using, possibly reusing, and storing research data. Managing data can include backing up raw data to external drives, naming data files, organizing files into directories, and even creating separate files to describe the nature of your data. Each research project has its own data management needs and strategies and we’d love to hear more about yours.

Current Data Management Practices

All of you are in the same situation in terms of creating and using research data and we’d like to start by having you share some ideas with us and each other.

1. Tell us a little bit about your data management practices, particularly in terms of what you think works well for you or your project/department.
2. Are there aspects of managing your research data that you’re struggling with?
3. What is your data management role in your project or department? (Managing own data only, responsible for shared data, responsible for an advisor’s data…)
4. What kind of training, documentation, or other support related to data management have you been given? (in the project, department, from advisor)
5. [If it hasn’t come up yet: Is using/reusing data from other sources? How big a part of the research is this data? How did you find it?]
6. [If there’s time: How do you feel about sharing your data publicly or on a case-by-case basis?]

Potential Data Management Services

1. What resources or services have you found on campus to help you manage your data? How did you learn about these services?
2. What services related to data management would you find useful? (If this doesn’t generate ideas, follow up on concerns that were stated and/or follow up with examples such as training, storage.)

Conclusion

1. Is there anything else you’d like to share with us about data management that we haven’t covered?

Thanks for participating. Next steps. Let us know if you’re willing to continue providing feedback or would like to know about any data management services that we develop.