Research Data Management Study Group
Executive Summary

Prior work by the Research Support and Services Working Group (RSSWG) established recommendations for further activities, including collaborative partnership with campus libraries and research communities to enhance the data management of research assets. The Research Data Management Study Group (RDMSG) conducted focused interviews with representatives from a number of research communities, to assess current researcher data assets, needs, and funding situations. The interviews revealed a broad diversity in asset content and format, a large number of disparate needs, and an inadequate funding base for many researchers.

The study group proposes a one-year pilot project to address the most common, most urgent subset of these issues, involving joint funding and management by a small number of research partners along with DoIT and the Library, which will then be assessed for continuation or expansion. The functional requirements of the pilot project should focus on:

- providing easily-accessed, well-maintained storage and backup capacity for researcher data, coupled with flexible, researcher-managed access control mechanisms, and
- promotion and dissemination of specialized data management expertise to assist current efforts and build a community of practice between researchers and data management experts.

In addressing these goals, it will be important to leverage existing resources, provide for phased implementation of project deliverables, and focus on community-centered support that extends beyond purely technological solutions.

Specific pilot project activities would include:

- Partnership with between three and five campus communities in order to develop and maintain a network of distributed storage nodes, with mechanisms for automated backup and archival support of data stored on them, access to storage capacity via multiple standardized protocols, and management interfaces allowing simple, flexible, researcher-controlled assignment of access management policies. Researcher access to this distributed storage pool would be on a "pay once, use forever" basis, partially subsidized by the partners, with remaining costs covered by specific funding written into the research grants of new projects joining the pool.
- Provision of consultation services to researchers attempting to preserve existing or new assets: assistance would focus on helping researchers locate existing campus resources, determining appropriate metadata standards and resolving format compatibility issues, and helping to develop sustainable preservation workflows. The requisite curatorial and technological skills would be provided by DoIT and Library staff in partnership, under an organizational model that makes assistance easily and directly available to individual researchers.

These actions will address critical common needs of many research communities, providing support that will enhance the quality and maintainability of research efforts, and alleviating the risk of losing a valuable part of the scholarly record. An effective and immediate response will provide the foundation for further efforts at research support, and protect currently at-risk researchers and research data.

Research Data Management Study Group
Summary Report

September 15, 2008

Bruce Barton, DoIT Academic Technology,
Jan Cheetham, DoIT Academic Technology,
Doug Flee, DoIT Data Resource Management Technology (Project Sponsor),
Steve Krogull, DoIT Academic Technology (Project Sponsor),
Jim Muehlenberg, DoIT Academic Technology (Project Sponsor),
Dorothea Salo, UW Digital Collections Center,
Mike Simpson, DoIT Academic Technology (Project Manager),
Alan Wolf, DoIT Academic Technology.

While its research mission is the distinguishing feature of a university (as opposed to a college, for example), the synergy that exists between research, teaching, and service is a central part of the Wisconsin Idea, both in theory and practice ... While our university has long maintained national and international pride of place for its highly ranked departments, schools, and academic units, it is also critical that we continue to support the numerous faculty, who also have multiple or joint appointments, thus energizing faculty research and teaching and breaking down barriers between departments, disciplines, and areas of study.

Interdisciplinary research has long been and will continue to be at the creative center of the UW-Madison's mission. Confronting complex environmental, social, cultural, economic, and medical changes and challenges, whether local or global, requires the collaborative, visionary efforts of faculty and staff across multiple disciplines. An interdisciplinary campus engaged with the constant ferment of our larger world is likewise essential for attracting and retaining the best faculty, as well as for preparing students to be active thinkers, workers, and citizens.

UW-Madison Reaccreditation Self-Study 2009, Team 2:
Integrating the Processes of Discovery and Learning

Contents

1 Introduction

> Campus cyberinfrastructure is not just about the technology. We need to understand and engage the research community, bridge the cultures, enhance the collaborative relationships on campuses and between campuses, and learn from each other. What is the process by which [we] can best continue sharing and collaboration among this community? How can we best interoperate and integrate among campus and national cyberinfrastructure efforts?
>
> *Final Report: A Workshop on Effective Approaches to Campus Research Computing Cyberinfrastructure*

In the Fall of 2007, the Research Support and Services Working Group (RSSWG) submitted an Interim Report to the CIO, making several related recommendations for possible future work. These included the implementation of larger shared storage pools, further partnership with interested campus communities, and collaboration with campus libraries in pursuance of better research data management. Several of these recommendations originated in the focus group discussions conducted by the Scholarly Asset Management Initial Exploratory Group (SAMIEG), a joint project between DoIT and the UW-Madison Library. In January 2008, the RSSWG commissioned the Research Data Management Study Group (RDMSG) to conduct a set of interviews with campus researchers from a broad range of disciplines, to gather information about their data management practices and areas of need.

The group's primary goals were:

- to assess current and future needs for data management within a variety of research communities;
- to gather functional requirements for a potential infrastructure to address those needs;
- to establish partnerships with representative campus providers of research data management services to inform further work; and
- to determine if there is interest in collaboration with DoIT and the Library in developing data management systems that are scalable, broadly applicable, and support customization for specific needs.

This report details the information gathering process conducted by the study group; presents an assessment of existing research data assets, researcher needs, and funding; and proposes activities that could be undertaken to provide support to the research mission of the university.

2 Methodology

> Enormous changes are taking place in the information landscape that are transforming teaching and learning, scholarly communication and the role of traditional research library services. Many of these changes have been brought about by technology and the explosion of electronic content made possible by electronic publishing, mass digitization projects, and the internet ... The implications of a shift from the library as a physical space to the library as virtual digital environment are immense and truly disruptive.
>
> *Information Behavior of the Researcher of the Future*

2.1 Interview Process

The study group gathered needs and requirements through interviews conducted with individuals representing a broad spectrum of research disciplines on the UW-Madison campus. These individuals were identified through recommendation by team members and project sponsors, based on information about their data management interests, disciplinary areas, and availability, amongst other criteria.

Interviewees were contacted by email and/or telephone; interviews were scheduled to try to pair several members of each community with two study group members, to keep discussions focused on the researchers and the specific details of their community. The study group encouraged candidates to invite additional colleagues or technical support staff, or to suggest alternative interview candidates if they themselves were unable to participate.

The group conducted eight interviews between March and May of 2008. Information from an earlier interview conducted in August 2007 is also included in this study. A total of twenty-one individuals, from earth and space sciences, geographic information science, medical and biomedical sciences, life science museums and departments, arts, humanities, and social sciences were interviewed. The interviewees comprised eight faculty, eight staff scientists, three information processing consultants, one special librarian, and one graduate student.

Appendix A contains detailed descriptions of the locations, dates, and participants for each interview.

2.2 Interview Protocol

The team developed a detailed interview protocol, beginning with very general questions about state ("What do you have?"), behavior ("What are you doing with it?"), funding ("How are you paying for it?"), and the future ("What would you like to be doing?"). These general questions were then subdivided into slightly more specific topics ("How do you take care of your data?", "Do you use your data in teaching/learning?") which in turn were broken down into very detailed inquiries ("What technical metadata standards are you required to use for deposit of your research data into disciplinary repositories?").

The study group used the protocol as a general guide for the interview process, trying to encourage the researchers to spend time and fill in details on the sections they felt were most important, using the specific prompts from the protocol to stimulate further discussion as necessary.

Appendix B reproduces the complete interview protocol in outline form.

2.3 Analysis of Results

Study group members recorded each interview session, with permission from the participants. The team engaged a private transcription service to convert the audio recordings into text format. The completed transcripts were archived in a collaborative team website for ease of reference, along with a number of references to related resources, including reports from similar studies conducted at other institutions. Team members conducted an analysis of the transcript contents, focusing on assessments of existing assets, current and future researcher needs, and research funding.

Additionally, the group invited several representatives of existing research service and support operations to act as advisors and offer feedback during the analysis process and the drafting of the summary report. These advisors included Andy Arnold (Associate Director, Social Science Computing Cooperative), Rob Kohlhepp (Director, Computer-Aided Engineering Center), and Richard Kunert (Information Systems Manager, University of Wisconsin Biotechnology Center).

Appendix C contains a selected bibliography of related content.

3 Assessment

> Other research artifacts – code, datasets, simulators, reference corpora (that is, standardized datasets) – play a significant role in the work of these computer scientists. But are they archival? Right now, they aren't. Many participants confessed that they could not regenerate their published results because they had not archived intermediate datasets, datasets that were dependent on network state and other circumstantial factors (compiler parameters, e.g.). Is it possible to save these all of these artifacts? Is it necessary? This is something that must be determined by the scholars research community; the ability to reuse the data fundamentally changes the nature of the science.
>
> *From Writing and Analysis to the Repository: Taking the Scholars Perspective on Scholarly Archiving*

3.1 Data Assets

Information gathered from interviews reveals that digital research assets across communities on campus vary widely in size, in quantity, and in format (see Table 1, Appendix D).

Primary data objects used in research include medical imaging and stimuli response recordings, metering of heart rate and skin conductance, electrophysiology data, digital microscopy imaging, numeric data stored in text or binary files, audio recordings, nucleic sequencing data, high-definition video, physical specimens and their collection metadata, chromatograms, GIS data sets, digitized maps and census data. These objects are stored in a variety of formats, some standardized, some proprietary, and some custom-crafted. Even within a single discipline the use of multiple and incompatible formats can make meta-analysis across data sets infeasible.

Researchers apply a wide variety of transformative manipulations to primary data to produce secondary artifacts. These transformations include statistical and pattern analyses using both commercial and custom-developed software applications, extraction of representative subsets and other refined data products, real-time or offline simulations, and incorporation into visualization and discovery applications.

In many cases, researchers collect ancillary data that extends and contextualizes the primary research outputs: survey data and patient questionnaires, lab notebooks, field notes, spreadsheets, data dictionaries, annotations and lecture notes, to name a few. Much of this ancillary data (and some of the

primary data as well) is not currently digitized, e.g. lab notebooks, handwritten specimen labels. Some of this contextualizing data is maintained; much of it is lost, either because it is difficult or impossible to locate, or because it is simply deleted or destroyed. Analytical work done in support of theses and dissertations can also be lost, or simply not collected and catalogued in usable form.

The data itself can be subject to serious access control concerns, i.e. human subject information, student-produced work with FERPA and copyright issues, and collection location information for endangered species. Some data is intended to be shared as widely as possible; some data is sensitive enough to be kept on removable disk drives, locked inside offices or server rooms. These different classes of research data are sometimes mixed together during the collection process, making it difficult in later processing stages to disassociate data that can be shared from that which must be kept private or secure.

Much of the analysis and related work done with these data sets is done by researchers on personal desktops or laptops. Researchers share data with collaborators and colleagues via many different routes: through email, FTP sites, websites, network shares, and physically-mailed CDs, DVDs, and hard disks.

Across the interview participants, the total amount of collected digital assets generally ranged from a few gigabytes to tens of terabytes (see Table 2, Appendix D). None of the researchers interviewed reported gathering petabyte-level data sets currently, although at least one group anticipates that advances in data collection sensors will produce data sets in that range within six years.

3.2 Researcher Needs

Although we encountered a breadth of different needs overall, the interviewees expressed related needs in several areas (see Table 3, Appendix D, for additional details).

3.2.1 Data Storage and Backup

Several researchers report the need for an affordable, dependable means of storing and backing up their research assets. This need was seen principally among individual faculty investigators and museum directors in biomedical and life sciences, as well as faculty in the arts and humanities. Many of these faculty are in departments or units where IT staff, busy with administrative support, are not able to assist with data storage, backup, and sharing. In some cases, inadequate storage capacity is leading to loss of data: forcing some researchers to discard data from past experiments in order to make room for current ones or to avoid certain types of experiments and research altogether:

- "We're constantly manipulating the data and saving only what we think we have space for. I mean, there's a lot of primary data which we think we don't need and so far, we haven't needed to go back but there are so many things that as science and engineering keep moving on that you might want to look back at in the future. It's definitely a problem."
- "We could generate much larger data sets but we don't because what would we do with them?"
- "The lack of data space is confining us a bit. Collaborations in research are being held back. I've actually known people who have not acquired things on the computer because they're worried about filling up the hard drive. You really want the freedom to be able to acquire whatever you want and to sift and roam through that unfettered and not have the science stymied by lack of space."

Some interviewees indicated that the availability of a centralized storage and backup solution would free up time for them to focus on their research and enable sharing of data in ways currently not possible:

- "Certainly, I can maintain a subset or a working in-progress database but if I had off-site storage with access through another system like DoIT that would be a huge load off my mind and everybody else, and certainly we wouldn't have to worry about maintaining an IT person which at this point we just cannot afford."
- "And then, eventually, when that particular machine gets full, which happens quite regularly on my workstation, I have to do an archive where I basically create a hard drive copy here, a hard drive copy of my home studio, and also two DVD copies. And that's all done with USB

> external hard drives which are pretty slow. So it's a sneakernet and that's really the only way I have to communicate between my machines here and my machine in my studio as well, which is a real pain because I'm carrying around these hard drives which are fragile and I have to bring it if we come here. I'm only giving you the basics, so I'd say for my use and for students use, our main requirement isn't for a complex system categorizing our database or web space, we just need reliable access to large amounts of storage from multiple locations."

Of the interviewees who maintain storage servers and networks locally, few directly expressed a need for centralized storage services; most have the means of meeting their current needs and/or prefer to maintain their own local systems for storage. However, some are less certain about the funding and the feasibility of scaling up their current storage systems to provide capacity for data collection from more sensitive instruments, such as microscopes and satellite sensors, that are likely to be in use in the next 5 years or so.

3.2.2 Automated Distribution and Access Mechanisms

A majority of the researchers we interviewed who routinely share data with specific audiences described challenges related to managing access to their data sets. One described her current distribution system – burning CDs and DVDs – for government data licensed through her unit to students, faculty, and staff on campus as unsustainable. Another researcher envisioned a day in the future when methods will be needed to allow colleagues to directly access highly sensitive human subjects data online, rather than through going through an individual, as currently occurs. Another researcher who manages human subjects data spoke of a need for automatic methods for sorting data, based on access levels. Even researchers in a well-funded data management center report that their current methods for distributing satellite data to users still involves emailing requests and attachments back and forth. Another researcher indicated that his current methods for visualized geospatial data for specific audiences are inadequate: he would like to collaborate with an information technology specialist to develop automated methods for transforming data online.

3.2.3 Improved Digital Curation Practices

Our interviews turned up several issues arising from what could be described as inadequate digital curation practices. Shortcomings in digital asset curation contribute to some of the problems with data distribution described above. They also hamper researchers' abilities to perform transformations and visualizations needed to conduct analyses, to deposit data in disciplinary repositories, and to conduct meta-analyses on existing data sets. Like challenges with data storage, issues arising from curation are likely to threaten the ability of researchers to conduct and disseminate research at the scale necessary in the future. Many of our interviewees described future research pursuits such as microarray experiments, high definition video, high resolution microscopy, high resolution climate measurements, and DNA sequence studies: endeavors that are likely to depend on effective digital curation methods to ensure relevance and competitiveness in their disciplinary fields.

Inadequate digital curation processes alluded to by interviewees can be distinguished as four different types:

Lack of processes for structuring and coding data as it is captured. One of our interviewees noted that data stored on their servers are not in organized structures, such as databases, which would permit them to be searched in systematic ways. A consequence of this is lost opportunities for meta-analysis of the body of data, as described in the following:

- "They started a program at the last university I was at in the late 80's where they took the data and put everything into a database and spent a lot of money on it. And they had a number of very influential meta-analyses come out of that data and it wouldn't have happened if that had not been available in the database. And I think that we've only been here for five years and we really have not gotten into that stage, but we are developing a history of studies into population groups, and I can see people wanting to do a meta-analysis and I don't know how to find all the data at this point."

Incomplete capture of critical information (metadata) providing context and provenance of experimental data. In the case of several interviewees, this is due to the diffuse methods for recording information during experimentation, which in practice often involves collecting digital data from instrumentation while recording ancillary information in analog lab notebooks. A major consequence is lost opportunities for future analysis of this data,

analyses that could show trends across larger data sets. In the words of one researcher:

- "When we're trying to track the origin of genetic data, that is becoming a huge problem. Despite the fact that GenBank and other types of genetic data bases exist, keeping track of where that data originated: if it's a blood sample, where did that blood sample come from? who collected it? when was it collected? where was it collected? All of these things are typically unknown to anyone else, other than perhaps to the student or the PI and it's going to blow up in our faces in the near future without implementation of what I would call a tracking system. You know, essentially creating the metadata to link all of these things to do a kind of a chain of custody, if you will."

Another interviewee noted that:

- "I think the scale of the problem is that on the one hand, we have this fire hydrant full of data coming off the scanner, and then we have these little bits of really crucial data that are in lab books that sometimes belong to the graduate students or post-docs that leave and take the lab book with them. It's useful when you build up a lot of data to do a meta-analysis of our subjects, or find every subject to ensure an age range with a certain type of scan. And in order to do that, you really have to have access to all of it. And that's one of the issues, its an enormous problem to go through studies over ten or five years and locate all the logbooks to find out what really happened."

Difficulties utilizing metadata standards in data management strategies. Although most interviewees appeared aware of metadata standards movements within their disciplines, several noted that overlapping efforts and the rapid pace of change in standards development in their fields are barriers to adopting them. As one researcher said:

- "So, on one hand, the approach towards standardization is to be applauded but I think the changes work at such a greater rate than can be accommodated, and then we're right back to the same case where everyone is maintaining their own databases and can only share perhaps older, more standardized data sets rather than the current one that they're actually working on, and I certainly see that in a lot of the work that I do."

Failure to adopt disciplinary metadata standards could have significant con-

sequences: several interviewees noted that use of metadata standards is increasingly a requirement for proposals for external funding and for complying with funder's expectations for data sharing.

Lack of automated processes for attaching appropriate metadata to data. An example is a researcher who noted that incompatibilities between metadata written in proprietary software and the downstream applications he uses to process data cause data files to become unlinked, leading to data loss.

3.2.4 Digitization of Existing Resources

Another area of need noted by several researchers is for resources to enable digitization of currently-physical media: existing theses and dissertations, lab notebooks, field notes, and other similar items. Individuals from two of the museums represented in our interviews cited creation of digital records for their growing collections of physical specimens as a critical but currently unfunded imperative. These museums are utilizing student volunteers for this process, but the rate of digitization lags far behind the rate of new specimen acquisition.

3.2.5 Creation of a Community of Practice

Finally, a number of interviewees cited a need for building and sustaining institutional knowledge and culture around data management practices. In the case of one researcher, inadequate staffing will lead to a lack of continuity in informatics support when a current graduate student with expertise in informatics graduates. Other researchers cite a need for more structured campus community for data management practitioners to share expertise and knowledge.

3.2.6 Further Needs

Also discussed were several needs that are beyond the scope, in terms of resources or time, of the study group as it is currently chartered. They are noted for completeness, and because the larger effort required to accomplish these goals, were it to be successfully accomplished, would greatly enhance the research mission of the university: fully electronic lab notebooks

linked to comprehensive, automated archival mechanisms; remote monitoring and actuation of sensor networks and lab equipment; and the creation of a dedicated interdisciplinary "teaching lab" combining teaching and research spaces within a single facility.

3.3 Research Funding

The funding situation across research communities on campus is best described as uneven across project, departmental, and disciplinary boundaries, as well as across time.

Several research communities on campus already have established, well-funded support operations providing for the majority of their researchers' needs. These existing support groups are a valuable resource, both in terms of their ability to quickly respond to unmet service needs in neighboring communities, and in terms of their historical expertise and experience in delivering research support services.

Outside of groups served by these existing support operations, the funding situation is increasingly precarious. Many funding agencies such as the NEA, NIH, and NSF have begun to require archival deposit and/or broadened access to primary research data; the availability or deposit of ancillary or contextualizing data and metadata as described above is also becoming mandatory. Researchers on campus see many of these new requirements as unfunded mandates: the expectation of major funders is that data management practices should be available as part of the host institution's IT infrastructure, developed and maintained through local funding. Smaller research projects, unable to secure adequate funding for locally-maintained data management infrastructure, and unable to afford current enterprise-level service offerings, are particularly vulnerable.

A number of the staff scientists and the librarian we interviewed have data management as a primary responsibility of their position. However, one of the faculty researchers we interviewed pointed out that his department lacks staff with this type of expertise and indicated the desperate need for the creation of permanent informatics positions within departments and research communities. In increasingly technology-centric disciplinary research fields, much of the support for data management and participation is dependent either upon IT-savvy doctoral students or work taken out of teaching and research time by professors. This participant noted:

- "So many times, so many things get started, and then that student is done, and they all just wither away ... I find it very difficult to compete nationally in the research sector because I'm competing with people that don't teach at all, or teach one course every four years. I teach five courses a year."

In addition to uneven distribution of funding across communities, all communities are vulnerable to unevenness of funding across time. Interviewees observed that across-the-board budget cuts by funding agencies continue year by year, with fewer grants awarded, and less money given in each grant. They expressed a general uncertainty about all future funding in the light of this "drying up" of central agency support. Lump sums are sometimes available for specific work or to purchase equipment, but these are one-time awards that do not necessarily provide for continued maintenance and upgrades. Several researchers expressed frustration with the inability to maintain research efforts and outputs beyond the short lifetimes of specific grants, e.g.:

- "The money lasts just long enough to get stuff going, turning into something interesting, almost to the point where you can share it ... and then it's done."

One issue tied to funding that was expressed by even the more well-supported researchers was the absence of an ultimate exit strategy, the option of last resort when and if all funding for a project disappears. Much data management infrastructure is currently tied to grants: true long-term archival preservation of research data is difficult under this model, as any project-specific archival infrastructure loses funding at the same time as the research efforts that are being served by it.

4 Proposal

> The digital age has presented the research community with new opportunities. Research findings in digital form can be easily moved around, duplicated, handed to others, worked on with new tools, merged with other data, divided up in new ways, stored in vast volumes and manipulated by supercomputers if their nature so demands. There is now widespread recognition that data are a valuable long-term resource and that sharing them and making them publicly-available is essential if their potential value is to be realized ... Research funders and institutions should cooperate in seeking to ensure that long-term and sustainable arrangements are in place to preserve and make accessible the data that they deem to be of long-term value, and that such arrangements are not put at risk by short-term funding pressures.
>
> *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*

4.1 Functional Requirements

In responding to the data management needs and issues expressed by the researchers who participated in the interviews, the study group suggests a focus on three basic areas:

- **simple preservation of data**: the creation and operational maintenance of a large, affordable, easily-accessed storage pool available for use by researchers to store primary or other data objects, with associated backup capacity for archival preservation and disaster recovery;
- **controlled sharing of data**: the development of interfaces and/or tools that allow simple, flexible customization of access control policies for specific content archived in the storage pool; and
- **continuity of knowledge**: the creation of new ways of sharing and disseminating specialized data management expertise, possibly including building communities of practice and/or developing methods for individual consultations between researchers and data management experts.

4.2 Underlying Principles

In addressing these functional requirements, it will be important to keep the following principles in mind:

- **leveraging existing resources**: if specific support needs can be met by an existing departmental organization or operation, the proper role of a central research support program is to act as a liaison to help facilitate collaboration between the researcher and the support group, not as a gatekeeper or competitor to those operations;
- **phased implementation**: initial efforts will be scoped to two or three collaboration partners, rather than trying to meet all community needs immediately, so that success in a smaller pilot effort will increase the confidence and willingness of other researchers to participate;
- **community-centered support**: support efforts should be directed towards assisting research communities to meet their own needs, with expertise from within the community, controlled by the community, and with only very general services (as described above) provided centrally;
- **support beyond technology**: solutions comprised solely of expensive technology will fail, because of the underlying need to establish long-lasting cultural stability within and between the research, library, and IT communities on campus.

4.3 Proposed Activities

Two complementary activities could provide a starting point for addressing functional requirements within the context of the specified principles:

4.3.1 Distributed Storage Pool

Initially, identify between three and five campus partners willing to collaborate in developing and maintaining a network of distributed storage nodes, for use by researchers on campus (note that a number of potential partners have already expressed an interest in participating in this work). Initial capacity should be on the order of ten terabytes per node, with a modular design and sufficient expansion capacity to grow to one hundred terabytes per node within three years, if needed. Partners agree to provide support staff to maintain their local node, and to integrate each node into a grid-style storage infrastructure, available for general use. Partners also agree to collaborate on developing mechanisms for archival support of data stored on the grid, including regular backups, long-term archival backups, file integrity checking, and disaster recovery planning. Access to the pool must

be made available through several standardized protocols (i.e., HTTP, SRB, NFS, WebDAV, etc.) so that storage capacity can be accessed by specific researchers using whatever protocol is most convenient to their existing tools and workflows.

Secondly, develop mechanisms and/or management interfaces by which simple, flexible access controls can be applied to storage pool data directly by the researcher communities depositing the content. Researchers should have full control over the creation and maintenance of dynamic user groups that can then be assigned appropriate access rights to specific collections of content. Time-based access controls (i.e. embargoing) should be easy to establish and automate. These access control mechanisms allow siting of data outside departmental firewalls, where it can be fully and appropriately protected by explicit controls assigned by the data owner. The system must provide for easy sharing, re-use, and interoperability with other systems (content management, workflow processing, resource discovery, etc.) to enable collaboration and federation where desired by the researcher.

In this distributed model, the storage pool exists in the "middle ground" between individual research operations and a more fully centralized solution. Fundamental data management issues such as preservation are handled "close to the node," where automation and operational maintenance are easier to standardize and apply. Content and access control remains "close to the user," allowing quick access through familiar tools, and recognizing that the researcher knows the data, and can most appropriately apply access controls. The multiple-node architecture alleviates the need to unnecessarily transfer terabytes of data across the campus network, while offering opportunities for robust data integrity (i.e., through node-to-node mirroring, offsiting of archival content, etc.) where important. The network of nodes can be expanded incrementally, either by adding capacity to each node, or by adding additional nodes from new partners.

Because researcher funding is primarily grant-driven, access to the storage pool is on a "pay once, use forever" basis: each research project paying for initial access via funding written into the research grant. Continuing maintenance of data beyond the lifetime of any individual grant is partially subsidized by the partners, and partially funded by new projects joining the storage pool. In the words of one interviewee, no researcher "should ever have to worry that they will be forced to discard data that might have value later on."

4.3.2 Asset Management Consultation

Create a small, efficient, research-focused digitization assistance staff, or enhance an existing operation, to provide consultation to researchers attempting to preserve existing or new primary content or ancillary data. Assistance would focus on helping researchers locate existing campus resources that are available for their use in preserving their content; determining appropriate metadata standards and resolving format compatibility issues; and helping to develop sustainable preservation workflows.

Some content could be the focus of a more directed preservation effort, e.g. addressing the preservation of existing theses and dissertations, with a secondary goal of assisting the development of a formal policy on deposit and preservation of future materials in digital format. Additional (more ambitious) starting points could be assisting in the digital cataloging of museum specimen data, and helping to develop a preservation strategy for research lab notebooks, leading eventually to recommendations and/or specifications for electronic lab notebook standards.

In undertaking these activities, both Library and DoIT participation would be beneficial in supplying the requisite curatorial and technological skills. Developing an organization model that makes these skills easily and directly available to individual researchers in need of assistance would be critical.

4.4 Project Timeline

Any activity that becomes part of a pilot project should be in full operation within six months to one year of the start date at which collaborative management and joint funding is established: the ability to demonstrate immediate, direct benefits to researchers will be a driving force in attracting other partners to collaborative effort.

Progress assessments should be made at three months, six months, and one year after operational roll-out of the pilot project. At each assessment point, collaborative partners should come to a joint agreement on continuation or expansion of an activities undertaken.

5 Conclusion

> IT in the service of research can, like the research it supports, push the frontiers of knowledge. The only way to really understand researchers' needs and the value central IT can bring is to engage with them ... The stunning numbers of institutions without even a plan for research IT is a warning bell. There have just been too many other priorities. As they continue to partner with researchers and their support units, central IT organizations must focus on the research itself before the unsustainable dimensions of the budget for research IT begin to hamper institutional effectiveness and research reputation.
>
> *IT Engagement in Research: A View of Medical School Practice*

Despite the huge spectrum of assets, needs, and resources, there are several basic services that would be of great use to researchers on campus. Large-scale, well-maintained archival storage and simple access control over research data and related objects, coupled with consultation on curatorial issues surrounding digitization of current and future research content, address critical common needs of many research communities. This support is vital to the continued quality and maintainability of research being done at this institution. Ignoring these needs will endanger the availability of research funding grants, risk the loss of a valuable part of the scholarly record, and potentially damage the reputation of the university as a center of innovative research and scholarship. Infrastructural inadequacies faced by many researchers on campus should be addressed as soon as possible, in collaboration with any and all support efforts already under way, focusing on goals that can be achieved quickly and efficiently. An effective and immediate response will provide the foundation for further efforts at research support, and protect currently at-risk researchers and research data.

A Interview Schedule

Waisman Center, 3/4/2008

- **Terry Oakes**, Director of Image Analysis, Waisman Laboratory for Brain Imaging and Behavior.
- **John Ollinger**, Associate Scientist, Waisman Laboratory for Brain Imaging and Behavior.
- **Adrian Pederson**, Information Processing Consultant, Waisman Laboratory for Brain Imaging and Behavior.
- **Nathan Vack**, Information Processing Consultant, Waisman Laboratory for Brain Imaging and Behavior.

(Interviewers were Jim Muehlenberg and Mike Simpson.)

Animal Sciences Building, 3/12/2008

- **Kevin Eliceiri**. Director, Laboratory for Optical and Computational Instrumentation.
- **Brenda Ogle**, Assistant Professor, Department of Biomedical Engineering, College of Engineering.
- **Justin Williams**, Assistant Professor, Department of Biomedical Engineering, College of Engineering.

(Interviewers were Jan Cheetham and Jim Muehlenberg.)

Social Sciences Building, 3/17/2008

- **Robert Hauser**, Director, Center for Demography of Health and Aging, Social Science Computing Cooperative.

(Interviewers were Dorothea Salo and Mike Simpson.)

Humanities Building, 3/25/2008

- **Steve Hilyard**, Professor, Department of Art, School of Education.

(Interviewers were Jan Cheetham and Alan Wolf.)

Helen C. White Hall, 4/8/2008

- **Jon McKenzie**, Associate Professor, Department of English, College of Letters and Science.

(Interviewers were Mike Simpson and Dorothea Salo.)

Russell Labs, 4/25/2008

- **Peter DeVries**, Ph.D. student, Department of Entomology, College of Agricultural and Life Sciences.
- **Dan Young**, Director, Insect Research Collection, Department of Entomology, College of Agricultural and Life Sciences.

(Interviewers were Jan Cheetham and Mike Simpson.)

Academic Technology, 4/25/2008

- **Mark Berres**, Assistant Professor, Department of Animal Sciences, College of Agricultural and Life Sciences.
- **Ken Cameron**, Director, Wisconsin State Herbarium.
- **Mark Wetter**, Senior Academic Curator, Herbarium Library.

(Interviewers were Jan Cheetham and Alan Wolf.)

Science Hall, 5/15/2008

- **Sam Batzli**, Assistant Scientist, Environmental Remote Sensing Center, Space Science and Engineering Center.
- **James Beaudoin**, Information Processing Consultant, Applied Population Laboratory, Department of Rural Sociology, College of Agricultural and Life Sciences.
- **Jaime Stoltenberg**, Map and GLS Librarian, Robinson Map Library, Department of Geography, College of Letters and Science.

(Interviewers were Jan Cheetham and Mike Simpson.)

B Interview Protocol

Constraints

1. Interview participants will consist of:
   (a) No more than two interviewers from the Study Group.
   (b) One researcher being interviewed.
   (c) Optionally, an additional technical expert invited by the researcher.
2. An interview will last no more than sixty minutes.
3. Information to be gathered includes:
   (a) The state of current data curation efforts.
   (b) A prioritized needs assessment covering:
      i. Current situation.
      ii. Future needs.
   (c) Functional specifications for a service to meet those needs.
   (d) Possible commitment to a prototype implementation effort, including:
      i. Financial contributions.
      ii. Staffing contributions.
      iii. Constraints on participation.

Preamble ("Why are we here?")

1. DoIT interest in meeting researcher needs.
2. Library interest in data management efforts.
3. Prior work:
   (a) Research Support and Services Working Group (DoIT).
   (b) Scholarly Assets Management Initial Exploratory Group (DoIT+Library).
4. Purpose of the Research Data Management Study Group:
   (a) Identify common researcher needs across several communities.
   (b) Identify opportunities for collaboration.
   (c) Create proposal:
      i. Pilot implementation to meet needs.
      ii. Potential partnerships for implementation.

Current State ("What do you have?")

1. What kinds of datasets do you have?
   (a) What kinds of datasets do you collect?
   (b) Where do they come from?
      i. What tools created them?
      ii. What formats are they in?
         A. How quickly do the structure and format of your data change?
         B. What primarily drives those changes?

   (c) How many datasets do you have?
   (d) How large are your datasets?
2. Where are they stored?
   (a) Are they in one place, or spread out in several places?
      i. Do you mix your datasets with other kinds of files?
      ii. How do you keep everything tied together?
      iii. Are there related things that need to be associated manually?
         A. Historical data?
         B. Lab notebooks?
         C. Field notes?
   (b) How do you take care of your data?
      i. Are you doing backups?
      ii. Do you do any periodic checking for corruption?
      iii. Do you have a disaster recovery plan?
   (c) Do you deposit copies in national or disciplinary repositories?
      i. Are you required to deposit them?
      ii. Are there requirements for deposit?
         A. Technical metadata standards?
         B. Descriptive metadata standards?
   (d) Does your data ever need to be expired/deleted?
3. Are there restrictions on access to your data?
   (a) Data security policies?
   (b) Legal or regulatory requirements?
      i. Requirements on storage?
      ii. Limits on retention?
      iii. Specific regulations?
         A. FERPA/HIPAA
         B. Human subjects?
         C. Munitions laws?
   (c) Privacy or confidentiality?
   (d) Intellectual property?
   (e) Staged access?
      i. Embargoes?
      ii. Share-later?
4. Do you have old or unreadable data?
   (a) Proprietary formats?
   (b) Analog data?
   (c) Data migration needs?

Current Behaviors ("What are you doing with it?")

1. What kinds of processing do you do?
   (a) Do you keep all of your original data?
   (b) Are transformations stored as metadata, or as new datasets?

   (c) What other secondary artifacts do you produce?
2. Do you share data?
   (a) Are there sharing requirements imposed by funders/sponsors?
   (b) Do other people use your data?
      i. Who needs access?
      ii. How do they get to it?
   (c) Do you use other people's data?
      i. How do you get to it?
   (d) Do you participate in any groups that share data?
      i. Do you share at the institutional level?
         A. With other labs on campus?
         B. With other institutions?
      ii. Do you share at the disciplinary level?
         A. Within you discipline?
         B. Across disciplines?
3. Do you use your datasets for teaching/learning?

Current Funding ("How are you paying for this?")

1. How are you paying for what you have?
2. What requirements do your funding agencies have?
3. What is your funding situation going to be like in the future?
   (a) How much funding will you need?
   (b) How will you get it?
4. Would you interested in shared funding with other researchers on campus?
   (a) What could you contribute?
   (b) What constraints would there be on your participation?

Future Directions ("What would you like to be doing?")

1. What are you most pressing current needs?
2. What trends do you see in the future?
   (a) Number of datasets?
   (b) Size of datasets?
   (c) Rate of accumulation of new data?
3. What is your outlook on the future?
   (a) What are your dreams?
   (b) What keeps you up at night?
   (c) For what would you like to say, "My campus gives us that."
4. What kinds of collaboration would you find useful?
   (a) What specific technical expertise could you most use?
   (b) What do you have that you could offer to others?
   (c) Would you be interested in a cooperative model?

        i. Would would it look like?
        ii. Where does DoIT fit?
        iii. Where does the Library fit?
    (d) What kinds of services might be developed?
        i. Data preservation services?
        ii. Discovery services?
            A. Metadata searching?
            B. Data searching?

Postamble ("What comes next?")

1. Additional interviews followed by analysis.
2. Are you interested in a copy of the proposal?
3. If the project moves forward into implementation:
    (a) Would you be interested in participating?
    (b) Would you be able to offer any resources?

C Selected Bibliography

Harley, Diane, et al. "Assessing the Future Landscape of Scholarly Communication: An In-depth Study of Faculty Needs and Ways of Meeting Them." Center for Studies in Higher Education interim report (2008).
⟨http://cshe.berkeley.edu/publications/publications.php?id=300⟩

"Information Behavior of the Researcher of the Future." University College London (2007).
⟨http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/googlegen.aspx⟩

Kiley, Patricia J. and William J. Reese. "Team 2: Integrating the Processes of Discovery and Learning." University of Wisconsin-Madison Reaccreditation Self-Study (2009).
⟨http://www.greatu.wisc.edu/theme-teams/documents/team2_report_052908.pdf⟩

Klingenstein, Ken at al. "Final Report: A Workshop on Effective Approaches to Campus Research Computing Cyberinfrastructure." Pennsylvania State University (2006).
⟨http://middleware.internet2.edu/crcc/⟩

Marshall, Catherine C. "From Writing and Analysis to the Repository: Taking the Scholars Perspective on Scholarly Archiving." Best paper finalist, Joint Conference on Digital Libraries (2008).
⟨http://portal.acm.org/ft_gateway.cfm?id=1378930⟩

Nelson, Mark R. "IT Engagement in Research: A View of Medical School Practice." EDUCAUSE Center for Applied Research (2008).
⟨http://www.educause.edu/ERS0801/15255⟩

Sheehan, Mark. "Higher Education IT and Cyberinfrastructure: Integrating Technologies for Scholarship." EDUCAUSE Center for Applied Research (2008).
⟨http://www.educause.edu/ers0803/121168⟩

Steinhart, Gail, et al. "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library." CUL Data Working Group Report (2008).
⟨http://hdl.handle.net/1813/10903⟩

"Stewardship of Digital Research Data." Research Information Network (2008).
⟨http://www.rin.ac.uk/data-principles⟩

"To Share or not to Share: Publication and Quality Assurance of Research Data Outputs." Research Information Network (2008).
⟨http://www.rin.ac.uk/data-publication⟩

D Tables

These tables are available on the following pages:

Table 1, Types of Data, Metadata, and Other Assets

Table 2, Storage Systems, Data Collection Rate and Size

Table 3, Needs

Table 4, Examples of Lost of Threatened Data/Metadata and Issues Around Sharing Data

Table 1: Types of Data, Metadata, and Other Assets

Disc area	Data type	Applications, data processing workflows, and metadata types
Medical Sciences [TO]	• MRI, PET EEG data in vendor-proprietary formats • Statistical analyses in R, S, SPSS formats • Associated medical information (skin conductance, heart rate) • Survey data (print and electronic), patient questionnaires/interviews in text, Excel, and FileMaker/Mac formats) • Logs and lab notebooks mostly analog; a few on Google Docs/Spreadsheets	• ePon software for recording responses to stimuli • Programs and scripts that associate and transform data, these are dependent on external software libraries • Data are in incompatible formats with incompatible metadata; they would like to do meta-analysis, but it isn't feasible
Biomedical Sciences [BO, JW, KE]	• Flow cytometry profiles ("FLL" format.) • Electrophysiology data • Microscopy images, gel images (TIFF files.) • In future, will gather Microarray reader data (Affymetrix) and load into Excel or other numerical-analysis software • Excel files • Analog notebooks from students	• Use MATLAB data analysis but it does not assist well with metadata creation
Life Sciences [DY]	• Digital images of specimens, photographed by students • Specimen metadata handwritten and physically attached to specimen with labels. Some museums use barcoding; not used here because infrastructure doesn't exist. • Specimen collectors currently take GPS units with them; data goes into their computers or onto the handwritten labels. • No plans to mass-digitize labels; need labor to do data entry. • Sometimes grad students working with part of the collection digitize a few of the labels. • Raw sequence data kept, then software extracts alleles for specific loci, which go into the database. • Primary data (analog notebooks, field notes, etc) ends up in the trash after collectors retire	• Specimen metadata: DNA, descriptions, event metadata (when/where collected etc). GPS used for locations.
Life Sciences	• Nucleic-acid sequence data • DNA fingerprints	• Specimen metadata: location, label information • Write own software to track data in a MySQL database

[MB,KC]	• Chromatograms (proprietary file format) • Digital images of specimens (600 DPI) • 250K specimens in simple spreadsheet from UW-Madison; 350K from all of System • Working with Aluka Project to digitize 4500 valuable specimens	designed by faculty, cross-linking and standardization are only just beginning to happen • Raw sequence data kept, then software extracts alleles for specific loci, which go into the database. • Biomapper software for visualizing data, in geographic contexts
Earth/Space Science [SB]	• Raster images, aerial photos, satellite multispectral images • Also generates spatial data in GIS format and vector data • Not usually raw streamed data from satellites but often processed geotifs, derivative products • Formats are mostly open	• Difficult to share visualizations for raster imagery (can be 54 MB in size), has used tools to share in web mapping service format (makes a jpg or png with the ability to zoom) and server figure but staff time is a barrier to making it possible to send out all images this way • Uses a tiff to png Linux program, runs current viewer for daily satellite image they put online, but would like to migrate to a better understood mechanism • Used ESRI spatial data engine (Arc SDE) to manage and store vector type data but have moved away from storing imagery in DB, instead referencing data in with postgreSQL or PostGIS so manage location of imagery but not imagery itself
Earth/Space Science [JS]	• Local GIS data, in raster and vector formats with associated tabular data • Has data for 25 counties in WI	• Some counties don't provide metadata so librarian creates brief core metadata for users so they can interpret images, such as currency and projection of image • Partners from Historical Society and State Cartographer's Office develop metadata strategy so can compare historic images with current ones
Earth/Space Science [JB]	• Tabular government census and GIS data • Community planning data, addressing data:, geocode and plot on map	• Trying to provide visualization packages to researchers and students in courses, but they don't work well, would prefer to build one that is customized
Earth/Space Science [RG]*	• Raw data from satellite sensors • Processed data	• Use MatLab, IDL, Fortran for processing • Validation workflows involve large number of data files on the SAN, scheduled runs on cluster systems • Re-test and validate data, bringing in more datasets from various instruments • Do automated testing, then sorting and filing
Social Sciences	• Very large survey results, produced locally: text files, audio recordings (WAV downsampled to mp3)	• SAS, SPSS used to manipulate data, sometimes in the form of automated scripts

[BH]	• Raw census data, from IPUMS at the University of Minnesota and elsewhere • Statistical data from various government arms (Bureau of the Census especially) and NGOs • Statistics (often longitudinal) compiled locally • Crunched census data, produced locally • DNA samples and sequences from survey respondents and relatives	• Most data-crunching done on servers, not desktops; private version (with more information) kept on local servers indefinitely, along with data dictionaries
Arts [SH]	• Text files and spreadsheets • Images, audio, high-definition video • Images up to 4-5 GB each • Video from video models up to 90K frames, 5GB/minute of video • Image files in Photoshop format instead of TIFF for size	• Rename, copy, join large files • Video rendering often parallelized over several CPUs, can take weeks
Humanities [JM]	• Word-processing documents • Image collections on CD-ROM • Digital video, media essays • Annotations and lecture notes, associated with digital video, for teaching purposes	

* Note: Information from this set of researchers was obtained in a separate interview in August, 2007.

Table 2: Storage Systems, Data Collection Rate and Size

Disciplinary area	Storage system	Data collection rate	Current size	Future size
Medical Science	• SAN system holds data for analysis; backups are difficult	• 230 GB primary plus analysis data per study	22 TB	
Life and Biomedical Sciences	• Store data on external hard drives • Occasionally migraties data from obsolete media to hard drives • Has to take laptop to collect data, HD of laptop always full and has to transfer to storage device [JW] • microscopy facility has two 4-8 TB RAID 5 server	• Flow cytometry profiles: 50 MB/week • Electrophysiology images: 20 MB per second, 24/7 • Microscopy images: 200-400 MB after analysis, 1 GB before		• Increases expected when add microarray experiments • With new 32 channel detectors soon to be used, microscopy images will be 20-30 GB each
Life Sciences [DY]	• Physical specimens, physical tags containing descriptive and provenance information, some electronic records	• 3 million specimens, growing at 20K/year	20,000 records	
Life Sciences [KC, MB]	• Physical specimens, physical tags containing descriptive and provenance information, some electronic records • Digital records stored on external hard drives		MBs	
Earth/Space Sciences [SB]	• Uses network-attached storage, RAID-5 configured at 50-70% of capacity • Relies on RAID for backup	• Size of a typical aerial photo is approx 1 GB	10s of TBs	
Earth/Space Sciences [JS]	• Stored on an external HD, backed up, mirrored		160 GB	
Earth/Space Sciences [JB]	• Dumps onto server that backs up on tape, using RAID configuration		150-200 GB	
Earth/Space Sciences [RG]*	• Storage area network, 3 servers, 85 TB • Hooked into NCSA and clusters • Archive data to tape but would like archived data available online. • Handle data in ways they can afford in order to	• 3 TB data per experiment and approx. 100 GB/day	85 TB	• Data from the geosynchronous imaging Fourier transform sensor, likely to be deployed

	compete for grants. • TerraGrid is not quite a footprint for their needs			in 2014, will come in at 160-200 Megabits/sec; 2 TB/day of raw and processed data will need to be stored
Social Sciences	• 2 TB server for local data and data dictionaries in Social Sciences • Working copies of sensitive/restricted data kept on secured removable disk drives		2 TB	
Arts [SH]	• Back up on external USB hard drive; one copy at work, one at home, two on DVD and spends "a hundred hours a year" syncing data between computers for backup and archival • Writes own code to do incremental backups and mirrors and uses checksums to verify copies			
Humanities [JM]	• Keeps an extra hard drive for backups • Does not do any checking for data integrity			

* Note: Information from this set of researchers was obtained in a separate interview in August, 2007.

Table 3: Needs

Need type	Specific need
Back ups and integrity checking	Medical Sciences [TO]
Network connections between computers that allow accessing data between 2 locations	• Needs storage space and connectivity between computers in his class lab, his studio, and home. Likes the idea of shared access to data (rendered image) files for students, but seems to favor a networked computer (with remote access) rather than web-based file sharing tools as the way to do this. [SH]
Data digitization, processing, and/or hardware	• Migrating past works from analog to digital formats is a need, as is a modern updated computer teaching lab outfitted for 3D work (including a 3D printer) in the Art department with full time support person to run it. [SH] • Actively pursuing creating virtual and physical media labs and communities for undergraduate work in the arts and humanities. [JM] • A challenge is making digital representations and catalog entries for their vast number of specimens—the main barrier is having funds to pay workers to do this. [DY] • Need a .5 FTE to do data entry [JB] • Need staff time to process images for online repository [SB]
Method for sharing data online with specific audiences	• Possible need for sharing subsets of highly sensitive data between collaborators via a self-serve method rather than going through a programmer [BH] • Needs infrastructure, currently digital data is being distributed manually, wants to collaborate on compus to create a repository, with access to licensed materials restricted to those with rights via their NetID [JS] • We need better ways to share and discuss data sets. Right now when other researchers wants to use their satellite data, they hire a student to request data, download and store it, then use it on their desktop. This is not sustainable because everyone is setting up their own archiving and distribution system. [RG]
Partnerships and collaborations with campus experts	• Needs staff to help move and post data to directories for people to download. In future, will publish data in mapping service for streaming into other applications so users don't have to download, has some working examples and is testing for load balance. Open Geospatial Consortium publishes standards he uses, but needs someone dedicated to this problem to make production server that can handle requests from multiple users at same time. [SB] • 1 m 2008 statewide aerial imagery in Oct-Dec of 2008 on hard drive. Is starting to move to network—attached storage, approx 400 GB total has spread out to load balance. Would be good to have help from someone with specialized IT skills for this. [SB] • Could imagine a cost benefit to collaborating with campus, DoIT, Library in these areas: Grid computing—matching scientists with grid resources, might like to be part of TerraGrid; Data curation—staff, training, consultation, with library staff; Tools sets for collaboration, data markup, visualization; Education programs--collaborative "training" on visualization, curation, standards but also raising awareness with scientists about

	the need for standards in data; Dealing with parallel and distributed computing situations [AG]
Metadata and data curation	• There may be need in the future to develop methods to code data with different identifiers--so that people see only data they have clearance for [BH] • Curation of data has several unanswered issues for them: keeping up with and selecting metadata standards, how to capture ancillary information (metadata) such as notes on subjects, experimental conditions, etc., and the need for interfaces that make it easy and natural for researchers to record information such as electronic notebooks [TO] • the absence of a permanent staff person who is a bioinformatics expert in Entomology is a problem in maintaining the continuity of the data sets they build.. Currently this work is done by students who build data sets that don't survive after they leave [DY] • Ways to annotate image data, link transformed to raw data (possibly as metadata), attach metadata when processing data [KE and JW] • Sometimes graduate students mistakenly overwrite original data when making transformations—needs to be a way to have transformations saved as metadata on the original data file [KE] • Need to be able to describe what was done to data [JW] • Ability to annotate on images, etc with disciplinary community [KE] • Need to create data sets that are interoperable with similar sets at other institutions so they can pool information and allow users to search across multiple collections. [MB]
Long term data storage/archiving/backup	• Looking at requirements for archiving data for long term so that meta-analyses across past studies, which may be important in the future could be done. The needs for meeting data storage requirements for NIH funded projects and sharing data with external collaborators are driving some discussion about archiving and they are thinking about automated processes that tag and save data in different containers. [TO] • Work close to the storage maximum of their external hard drives—often deleting data from past experiments that might have value so that they can store data from current experiments. They worry that they are forced to discard data that might have some value later on. [KE] • Generate data in several different formats, so any centralized storage solution would have to handle all of them. Backing up data is also a need but, in their view, needs to be combined with storage and sharing—not a separate step. [KE] • Needs a way to archive student digital projects so that they can be used by future students—some possibly open to the public and others available to students only. [JM] • Might have an interest in being able to point to a campus infrastructure for storing data in future grant proposals but disciplinary infrastructures might be more important. [DY] • Need for systems for storing data that are common so that people can share and re-use data. Plus, the ability to share raw data would be good, because there is only so much interpretation that can be done on published data. Would be good if the system had the ability to do processing of raw data built in, automating workflows. [MB]
Real time remote data monitoring	• Need a system that would allow monitoring the progress of experiments (flow cytometry, microscopy, gels) remotely [BO] or involving students in real-time data collection [JW]

Table 4: Examples of Lost or Threatened Data/metadata and Issues Around Sharing Data

Disc. area	Description of lost/threatened data and metadata	Current data sharing practices, barriers, and security/privacy concerns
Med Sci	• Data for a single experiment is often scattered across many different notebooks and logs • Scripts that associate machine data with questionnaires, interviews are updated for new experiments but are not under version control. • Data on DAT tapes may not be retrievable in future but no plans to migrate the data to another system.	• Share with colleagues via email, FTP, mailed DVDs, MySpace, and passworded websites. Data needs to be private until all papers or theses from it are complete. • Careful about referring to animal subjects in insecure environments owing to pressure from animal-rights groups • Sharing partly governed by consent forms signed by subjects. These agreements change from experiment to experiment and are difficult to track. • NIH data-deposit requirements, but compliance is difficult because of the scattered state of the data
BioMed Science	• External hard drives are locked inside offices, no offsite backups for these data. • Constantly manipulating data and saving only what they have space for. [BO] • Space limits what data can be collected [KE, JW]	• Share data on FTP servers (insecure), wikis, and WebDAV to get around hospital's secure networking • Conflicts using the same servers and networking for research and administrative purposes. Computer-security staff are serious about data security, sometimes too serious to allow sufficient data sharing space. • Has to upload de-identified data to federal clinical trial registry [JW] • Uses wiki to share post-processed, condensed data with students and outside collaborators [JW] • Pressure to deposit in federal depositories (NIH funding) but compliance is low [KE]
Life Sci	• Some data buried in manuscript theses and dissertations (DY) • One graduate student doing bioinformatics but knowledge will be lost when he leaves [DY]	• Some DNA work has to go to GenBank • Exchange digital images with colleagues via email
Life Sci	• GIS-based online database of plant/animal specimens, somewhat obsolete [MB] • Primary data (analog notebooks, field notes, etc) ends up in the trash after collectors retire [MB]	• Have to black out location information for public consumption for species that are endangered or threatened
Earth/Space Sci [JB]	• None currently	• Recently getting medical data and are working through restrictions • Need to share visualized data but not enough staff time to prepare visualizations
Earth/Sp	• 10 year old County data in older formats, some have	• Burns CD and DVDs and students, staff, faculty have to

ace Sci [JS]	been migrated	come over and pick them up—no methods for sharing licensed data online • Agreements signed with WI counties restrict use to only UW students, staff, faculty, other states (e.g. NY and MI) have repositories of downloadable data and visualization tools open to anyone
Earth/Space Science [SB]	• None currently	• Open to anyone who wants to download, WisconsinView, currently has 5000 users who've created an account
Earth/Space Sci [RG]*	• None currently	• Currently, share data by sending emails with links to data sets in storage areas and FTP or R-Sync • Experimenting with federated storage using: Globus/Condor grid tools, MACAIDAS V visualization system (Java-based, a client/server compatibility box), and Web services to a cluster of large scale computing that would fit into the visualization system using Open DAP and ADDE (data distribution protocols to get data and metadata)
Soc Sci	• If lose funding or resources for maintaining the current secure data storage system they have now, the future of the data is in question • Local data migrated as necessary; hasn't been a major problem	• DISC presents "public face" of local data, as ASQ/SAS/Stata/SPSS • Serious privacy and confidentiality concerns with survey and DNA data
Arts	• Have to rename, copy, join large files; some external. HDs can't handle the load [SH] • Hundreds of slides and large format transparencies that may need format migration [SH]	• Exchange files on USB hard drives
Human	• CDs designed for Mac OS 9 no longer usable [JM]	• FERPA a concern for dissemination of student work

* Note: Information from this set of researchers was obtained in a separate interview in August, 2007.

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.

Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library

We suggest at this stage of our work that it is reasonable to presume that there may be no one vision for technology-enabled scholarship in a field. Ultimately, the personality of individuals combined with disciplinary tradition, the needs of the field, and affiliation with type of higher education institution will determine how widespread public sharing of non-peer-reviewed incipient ideas and data will be and what forms final archival publications will take.

Assessing the Future Landscape of Scholarly Communication: An In-depth Study of Faculty Needs and Ways of Meeting Them

Managing, preserving, and providing access to digital research data involves significant costs which will increase as the volumes of data increase. This presents significant challenges to research institutions and funders. Effective and efficient management of data requires investment in infrastructure and specialist professional support services, to ensure that data are properly selected and stored, that they can readily be accessed, and that their integrity can be assured over time. The costs and benefits of such investment must be regularly and systematically reviewed but it is clear that without investment in effective data management and access regimes, the overall efficiency of research will fall.

Research Data Principles and Guidelines