Report of the Blue Ribbon Panel on Research Data Strategy
University of Michigan, Feb. 2011

The research enterprise is producing exponentially more data with the passage of time, in almost every discipline. Success in research will be more and more tied to the ability to harness this data, including data from multiple sources. See, for example, a recent issue of *Science*, which features a special section on this topic (331(6018):692-729, Feb 11, 2011). To complement this, there are increasing mandates to make one's own data available for additional analysis by others. If the University of Michigan is to remain a leader in research, across a wide range of disciplines, an effective and available data infrastructure becomes a critical necessity.

Scientific computation has garnered attention for a much longer period than scientific data. It is also often easier to manage with shared resources that can be "rented" as needed on demand, because computation, in the absence of data, can easily be moved from one machine to another – it is the associated data that causes "stickiness" and makes resource management a challenge. Data requires not just storage but also transfer to where it is required so that it can be used with the necessary computational resources, analytical tools, and access restrictions.

To understand the complex issues involved and to help the University of Michigan (UM) develop a plan to manage them, Dan Atkins, AVP for Research Cyberinfrastructure and Chair of the UM IT Council, and Laura Patterson, AVP-ITS and UM CIO, appointed a Blue Ribbon Panel on Research Data Strategy in March 2010. See Appendix A for the panel membership and the formal charge given to it.

UM is no exception when it comes to the value to research of good data management. We present a vision for the benefit to be obtained thereby in Appendix B and several examples in Appendix C. Not only is there a large unmet need to support research data management, there is also tremendous opportunity to accelerate the pace and impact of our research through better data management.

There is a wide range of research activities at UM, with a corresponding spectrum of data needs. Nevertheless, it is possible to develop a rough categorization of the major types of user needs as shown in Appendix D. These include the need for data and metadata recording, distribution, storage, archiving, publication, protection, and analysis; the need to deal with heterogeneity; and the need for user education and consulting help.

The need for research data management is not unique to UM. Several peer institutions have recognized this need and already developed a plan. Several other organizations have also written thoughtful reports and white papers on the issues involved. We studied a broad range of written reports, identified in Appendix F, and also interviewed some leading figures in research data management, including Sayeed Choudhury, Reagan Moore, and Lincoln Stein. A record of these interviews is in Appendix G.

It is widely accepted that the infrastructure for integration of data with national and international grid projects will be important. Web services will be used for moving large amounts of data, performing service discovery in real time, and increasingly be deriving data on the basis of semantic model systems. Consequently, the underlying cyberinfrastructure must be designed to support these systems and their demands. Important "middleware" or service-based components of the research infrastructure, such as a common, scalable identifier services for samples across the research enterprise, do not currently exist. One can argue that these systems are needed even now for sample management and for linking to clinical and research data in a systematic fashion. This shortcoming is often currently overcome by labor intensive manual processes used to link specimens to research or clinical datasets, such as chart or registry review. Development of such linking systems would make integration with scientific research efforts more tractable, and would pay both short and long term dividends.

One of us, Chris Miller, also conducted an extensive survey of all LSA research active faculty members. Detailed results from this survey are in Appendix E. The key findings were:

1. 81% of the research groups use less than 2TBs/year for grant funded research and only 4% use above 10TBs annually.
2. 65% create or use files that are 1GB or less.
3. 2/3 of PIs *do not* request equipment purchases in their grants.
4. Only 10% of the researchers claim that there data *could never* be re-created.
5. 40% of the researchers need to share their data outside of U-M.
6. While most researchers back up their research data, only 27% do so daily or more often.
7. Nearly half of the PIs prefer affordable and centralized back-up services over increased centralized or localized storage. There is no interest in 3rd party solutions.
8. For nearly half of the PIs, it is either required or recommended by their funding agency to provide continued access to their data after the grant has ended.
9. Nearly half of the PIs suggest that 50% or more of their data could be deleted after the project ends.

The conclusions of the blue ribbon panel are that research is becoming more data intensive across disciplines, and believe that this trend will both continue and accelerate. A well-considered approach to data management is vital. At the current time, UM does provide some support for research data management, but this support is not nearly enough, and we are at risk of falling behind in the race for research leadership. At the same time, the panel believes there is a window of opportunity right now for UM to leapfrog the competition and establish a leadership position. Our detailed recommendations in this regard are given below.

Recommendations

The Opportunity

As areas of inquiry become more data driven in many disciplines, a typical response is to "roll your own" solution to address these needs. Such solutions are often messy and non-standard. While very effective in the short run, they usually do not scale well and become brittle barriers to the next generation of functionality. Many times, these short-term solutions become bothersome "traditions" that keep disciplines from adopting best practices for data management. In short, even if an institution, such as UM, were to offer an excellent set of data management services for free, only some fraction of faculty would avail themselves of these services.

Federal research funding agencies have begun to require data sharing plans for data resulting from research conducted with grant funds. NIH has had a data sharing plan requirement for several years now, at least for larger grants (more than $500K per year), but has started addressing this issue vigorously in the past year or so. [See http://grants.nih.gov/grants/policy/data_sharing/data_sharing_chart.doc and http://grants.nih.gov/grants/policy/data_sharing/]. NSF, which has historically not had any similar requirement, has begun requiring this as part of proposals starting in January 2011. [http://www.nsf.gov/bfa/dias/policy/dmp.jsp]. As researchers learn to comply with this new requirement, UM can convert it from a burden to an opportunity both for the researchers and for UM. As the major funding organizations compel researchers to think about their data management and sharing, suitable channels at UM for convenient and affordable data management are likely to be adopted with gratitude, not only by researchers explicitly seeking solutions for their data management needs, but also by researchers who do not yet do so. In other words, we can create a milieu in which researchers don't just get by with creating the minimum data management plan acceptable to funding agencies, but actually develop a robust plan leveraging UM facilities. By enabling this switch, UM can not only increase the success rates for funding proposals, but also the research outcome and impact long term.

Even in the absence of any federal requirements, better research data management is a desirable end in itself, for the many reasons given above, including the rapid advancement of knowledge and success in the research enterprise of the future. As an anecdotal measure of cultural validation, there is even a useful Wikipedia page on the subject at http://en.wikipedia.org/wiki/Data_sharing. Federal mandates for data management plans provide a bully pulpit that the university can exploit to its benefit and that of the UM research community. Which brings us to the question of what precisely UM should do to maximize the benefit from this opportunity.

The heterogeneity of the UM research community suggests that a "one size" solution is not likely to fit all. Rather, **we recommend the creation of multiple service interfaces, each capable of operating at multiple scales**, and each building in a natural way from the previous one, for this purpose. Some specific service interfaces are described below. **We additionally recommend the provisioning of expert consulting on research data management** also described below.

We recommend that UM provide a specified level of data management **services to each funded research project without charge to the project, or researchers involved. These services must include not just data storage, but also assistance in multiple aspects of data management, as discussed below.** Many projects will require additional data manipulation and data management services. Costs for these additional services and additional data should be charged back to the project in most cases. And, as the research itself is typically a collaborative effort across institutions, these needs should be considered with a multi-institutional solution in mind. In most cases, it should be possible to predict these costs at the time projects are proposed, so that they can be written up in the data management plan and included in the project budget. Some suggested service interfaces are described below, along with the anticipated manner of use and suggested cost responsibility.

Compliance service

DRDA already offers a free service to UM researchers in meeting requirements with regard to data management and sharing plans. This service has already proven valuable, and **our recommendation is that this service become a focal point to educate users about other service offerings and support them in developing a strong data management plan rather than one that is minimally compliant**. We expect that in many cases explicit budget allocation for research data management and archiving activities, complemented with university facilities described below, will give proposals an extra edge.

Data management plans are not the only compliance issue researchers must contend with. Researchers working with patient data need to comply with HIPAA requirements. Many researchers may need help with digital rights management, for example to publish data while retaining rights. Additional flavors are required of the compliance service described in the preceding paragraph, to meet (and suitably leverage) such needs.

Expert Consultation

Many researchers need expert advice, on a range of topics related to the management of research data. These topics include:

- security and compliance
- use of data management service interfaces described below
- metadata recording and preparation of data for archiving
- choice of data representation
- other tasks, such as data recording and data analysis.

For many users, getting a small amount of help and direction in the beginning can make a huge difference to how well they manage their data. In fact, many potential users may not even understand the different services offered by UM, and which ones may be appropriate for them. Furthermore, a small upfront investment in doing things "the right way" can lead to significant benefits later on. **We recommend that UM should provide an expert consulting service to address these needs. Some level of basic consultation should be free, for the reasons enunciated above.** There is sometimes going to be a

need also for more involved situation-specific consultation, e.g. for help in schema design for a particular data set to be stored in a database. In such cases, there should be a mechanism to escalate smoothly from the free (to the user) standard consultation to custom consultation for a fee, performed in the context of a standard established data management environment.

Data storage service

This service provides basic no-frills storage, more in the spirit of back-up storage than of real archiving. A user provides a chunk of data, and the system stores it safely, with encryption and durability. The system may choose to store it in the cloud or in a remote data center or in any other place – this location should be transparent to the user. There may be some restrictions on this storage on account of compute requirements – see below. The basic service should provide a reasonable degree of durability against common failures, e.g. equivalent to RAID-6. For situations where a stronger guarantee of durability is required, storage solutions may be more costly, and the associated additional cost may have to be recharged to the users even if the cost of the basic service is not.

There are many commercial providers of back-up storage for PCs and servers. The University may even choose to negotiate a group discount and outsource this task. Note that research data back-up is slightly different from PC back-up in that one is not looking for differences between thousands of files in hundreds of directories, most of which have not changed. Rather, there is likely to be a more actively managed data collection and versioning process.

The minimum service interface required is to be able to deposit a packaged unit of data, such as a directory in a file system, to retrieve metadata about previously stored data and to retrieve a desired unit of data from among those previously saved. This must be possible to do in a manner that is either independent of operating system or at least supports several common operating systems, such as Linux, MacOS and MS Windows. The service must provide positive confirmation of storage so that the user can be assured that data has actually been recorded and not merely dropped off, possibly to be lost.

We believe that a data storage service of 1 terabyte per year of project duration will meet the needs of most research projects, with moderate cost to UM. We expect approximately half that amount – 500GB per year of project duration -- will need to be archived permanently, again at UM cost. Some larger projects will require managing more data than this. Costs of storage beyond the base level can be charged back to users, as long as they are competitive with commercial and "roll-your-own" options. In 2011, we would expect this cost to be approximately $200 per terabyte per year.

Commercial vendors of similar services, such as Amazon S3, Mozy, and Carbonite, all place restrictions on the amount of data transferred to and from storage. While some restrictions are probably required to be able to manage the storage, commercial vendors may be setting very low limits because of internet bandwidth restrictions. We recommend that these be negotiated depending on the needs of individual research projects – occasional transfers may suffice for many projects, but others may need to

perform data intensive computing with concomitant large access capacity. Internet2 may prove helpful in this regard.

Standard UM computer accounts are per individual. If there is substantial allocation and use of the data storage service, and other data management facilities, on a per project basis, then there must be a clean notion of a project account, through which the PI, or designee, can provide access to project participants. For some projects, it may be necessary to have finer access control management, without giving access to all project-related data to all project personnel. Also, the design of this facility should permit convenient managed sharing with individuals, such as external collaborators, who do not have a UM uniqname. Authentication capabilities, like those provided by Shibboleth [http://shibboleth.internet2.edu/], could be a means of enabling controlled access in a standardized way, leveraging existing institutional infrastructure. Note that commercially available data storage solutions are not likely to provide this type of capability. If UM contracts to a third party vendor, at least a thin customization layer on top is indicated.

Variants of the above scheme may be required for a few units, where the mode of work is different. For example, in the creative arts, digital creations, or high fidelity recordings of traditional creations, can easily run into several terabytes, and the work may not be funded by research grants. A larger UM provision may be required, without tie to research projects. On the other hand, there may be future revenue, in terms of licensing fees and sales, that could be taxed to defray some of the cost in a deferred manner.

There may be regulatory, or other, restrictions on some research data. For example, HIPAA regulations have impact on patient-related data. Where significant numbers of users have special requirements associated with their research data, the provided data management services must meet these requirements if they are to be used. UM should provide a few additional "flavors" of data management services to meet these needs, in addition to the "vanilla" service described above. If these needs are discussed at the time data management plans are written for funding proposals, it should not be difficult to identify cases where needs are not met. We believe that a small number of flavors of service (possible as few as two flavors, in addition to the basic) will suffice to cover the needs of almost all research projects.

Metadata recording service

Recording of metadata must be a key part of a strategy to manage data. It is crucial that appropriate metadata be recorded with data, for any data that will be reused or shared. Determining the right metadata to record is a critical aspect of making the associated data much more effective, but has rarely been a priority for researchers. In addition, entering metadata can be laborious, and have little (perceived) short-term benefit. In consequence, many researchers may be inclined to minimize both the effort to correctly identify the most useful metadata and the amount of metadata recorded. Any tools or advice that can aid researchers, in either better defining the needed metadata or automating the creation of metadata, can increase compliance, accuracy and, ultimately, the effectiveness of the stored data. **We therefore recommend that any storage service for research data**

should include some sort of free tools/facility/encouragement for researchers to define and create appropriate metadata concerning their data. This could be viewed as part of the "price" researchers pay to get access to institutional storage systems.

One simple benefit of metadata recording for the researcher is that the metadata can be used to label the data set and hence make it easier to find in the future (if the intent is to share). In addition, appropriate selection and use of metadata can make the underlying data useful for many (unforeseen) future uses, potentially far beyond the original research.

Several discipline-specific data resource indexing systems are available today. Obviously, an institution like UM is not interested in any one specific discipline. But if we go across disciplines, it is not clear that there is much we can do better than commercial search engines, such as those used by the Michigan Libraries or Google. For these reasons, **we recommend that Michigan not set up any special indexing system for its archive of research data. Rather, we recommend that these follow from inter-institutional collaborations and emerging best practices, some of which maybe discipline-specific**. Towards this end, we urge developing ties across institutions, something the UM library system does already.

UM already has a public archiving system already in place, Deep Blue. However, this is primarily geared towards research publications rather than research data. Ideally, there would be little distinction between the management and accessibility of primary research data and derivative intellectual products. As such, extending the scope of Deep Blue to embrace research data would be a natural next step. However, there are many practical difficulties in getting to this ideal. As such, we make no specific recommendations regarding Deep Blue's role in archiving research data.

Data publication service

Some research projects will benefit from (or maybe required to) make (some) datasets available over the web, for public access. For such publication to be meaningful, it will usually be necessary that suitable metadata be recorded with the data.

Services like Xerox Docushare, Microsoft's Sharepoint, BSCW and others provide some of this type of capability, though these commercial tools may be appropriate only for smaller, simpler projects. See http://sites.google.com/site/toolsgsd/tools-1/knowledge-management-tools for a list.

UM already provides a basic web-hosting service as a free public good to members of it research community. **We recommend a data publication service that combines this web hosting service with the data storage service to provide a unified service with natural functionality**. We expect this data publication service to be free, being a combination of two free services.

Data search service

Beyond having data sets discoverable at the granularity of full data sets, some researchers would like to have their data sets be queriable at fine granularity. In other words,

electronic publication may be desired not just of the data set as a whole, but with users permitted to access individual data items from the data set through a query interface.

For at least some scenarios, defining such an interface is not too hard, with very basic input from the researcher. This is closely coupled with the topic of metadata above. **We recommend that a minimum data search service be provided by UM for free as an extension to the data publication service**. We expect that many research project that avail of this service will require that it be augmented with expert consulting (which would be charged to the research project), especially where more complex searches are required.

Data licensing service

Whereas we may desire maximal unimpeded distribution for many products of our intellectual effort, there could be situations in which there is a desire to license or sell rights to some data, or digital creation. Where such sale is at large granularity, we currently have a mechanism through the office of technology transfer, under OVPR, and there are no new significant data management requirements. However, there could also be an opportunity for revenue at small granularity, for example, through a small fee charged per access to some digital artifact. **We recommend the provisioning of a data licensing service for this purpose, with costs charged back to users against revenues realized through licensing and sale**.

Data recording service

Before a researcher has data to share, there must be data recorded, whether through manual data input from paper or through electronic recording from scientific instruments, or other means. The number of different possible sources is large, and most researchers already have a standard way to record the data, so there does not appear to be a possibility to have a standardized institutional service offered in this regard. However, expert consulting is likely to be valuable, particularly in metadata capture. Even though much of this consulting may be customized to the specific project, **since metadata capture benefits the research enterprise more than the individual researcher, we recommend that the recharge for this expert consulting exclude the fraction related to metadata capture and preparation for archiving**.

Data analysis service

Typically, a researcher is interested in the data set not for itself, but because it can offer up information of value through analysis. The analysis is typically problem-specific, and requires considerable domain knowledge, and hence is not something that can be standardized as a service. Nevertheless, there are many tasks common to multiple analysis pipelines. A set of "lower-level" tools supporting common tasks can be offered as a data analysis creation service. In particular, these tools would include methods for data selection, data access and data visualization. Such building blocks would be designed to interface with the various research storage services provided by the University once specific implementations are in place. There already is a free UM service for statistical analysis: The Center for Statistical Consultation and Research

(CSCAR). **The proposed data analysis service should appropriately extend CSCAR facilities and work with the data storage service described above**.

Data with computation service

Often, researchers have to perform extensive computations with their data. This implies that "bandwidth" between the computational resource and the storage resource is a critical parameter. For data intensive computing, a fast growing need in many disciplines, computational services without corresponding data capabilities are of little use.

UM has already developed plans for providing computation as a service, and for hosting processors purchased by individual researchers. Extending these plans to include some sort of **"data with computation" plan** makes sense. While the actual shape of this appears at present to be user-specific, we believe that some exemplar storage system service able to complement the existing and planned computational services is required for UM. Once a couple of reasonable choices become available, we expect that many researchers will find ways to conform to the best fit from among the available options.

A key point to note here is that the provision of any such service imposes significant requirements on the data storage service. For example, having data tied up in "cloud" services may preclude its effective use. Similarly, investment in network infrastructure will be required, even if it does not appear to directly address data storage needs.

Conclusion

The importance and value of research data is becoming broadly recognized. Associated with this is the need to manage research data better. At the same time, funding agencies are also asking for data management plans. The University of Michigan has an opportunity to leapfrog peer institutions through investment in infrastructure to support research data. This report lays out a high level description of the set of services required for this purpose. **We recommend that full technical specifications of these services be developed next, along with system requirements, and cost estimates for system acquisition, development, maintenance, upgrades, and user support. To accomplish this on an ongoing basis, data management issues will need continued attention from the IT governance council, which in turn will need to seek input from a multiplicity of stakeholders and experts to define the best possible system for the university.**

Appendix A - Panel membership and formal charge

Membership

- H. V. Jagadish, Chair, EECS
- Ken Pienta, Medicine
- Drago Radev, EECS and SI
- Margaret Hedstrom, SI and ISR
- Richard Rood, AOSS
- Elaine Brock, DRDA
- Shawn McKee, Physics
- Paul Millis, University Audits
- Paul Killey, CAEN
- John Wilkin, University Library and HathiTrust
- Frank Manion, Cancer Center
- Margaret (Maggie) Levenstein, ISR and Ross School of Business
- Mary Simoni, Music
- Chris Miller, Astronomy
- Ron Barbret, Financial Operations
- Joanna Mirecki-Millunchick, Materials Science
- Heather Carlson, Medicinal Chemistry
- Cindy Wells, ITS

Charge

The Blue Ribbon Panel of Experts is charged to provide guidance to university leadership on the why, what, and how of research data strategy. The Panel is also encouraged to seek input from others far and wide through surveys, townhall meetings, structured testimony, use of social media, and/or any other mechanisms you wish to use.

Research communities at various rates are recognizing that e-science is now both computationally (modeling & simulation) and data intensive (data mining) and growing numbers are advocates of *data exploration* some are calling a fourth science paradigm, with the other three being *empirical, theoretical,* and *computational*[1]. Data sets are becoming the new instruments of science. Scientific data is currently doubling every year, and over the next 10 years this doubling time will shorten to 3-4 months.

It is difficult to project data sizes that will be commonly used in sciences in 10 years from now, but clearly they will be in the hundreds of Petabytes, beyond anyone's reach today. Dealing with such amounts of data will be very difficult even on the national scale, impossible locally. Having a balanced hierarchical system will be necessary with great research university such as ours both contributing to and taking from the data universe. We need to better attend to our local needs as well as play a broader role as partners with others engaged in global scale research.

[1] The Fourth Paradigm: Data-Intensive Scientific Discovery, http://research.microsoft.com/en-us/collaboration/fourthparadigm/

As a member of the Blue Ribbon Panel of Experts you are asked to contribute to:

- A vision of the role of digital data and services (digital objects, more generally) to the University's mission of research leadership over the coming decade. Include representative samples of the research drivers that are enabled by this vision of services.
- A more specific inventory of the critical needs of the research data services needs of the UM research community for the next five years but framed within a 10 year vision. Include representative samples of the research drivers that are enabled by this vision of services. The scope of needs should include research in all areas including the arts and humanities.
- An analysis of the gap between what is needed and what is currently available on the campus. A description of the problems existing that need to be mitigated and discussion of the lost opportunity costs for not doing so.
- The approach the University should take to satisfy needs including technical architecture, organizational form, training opportunities, polices, investment levels, and cost recovery.

Appendix B – Visions of research with good data management (vignettes from selected disciplines)

Social Science Research

Social science research is being transformed by changes in the availability of data as well as the increased facility of computation. In some respects, social science data is much more difficult to create in the current environment. Survey response rates have fallen and privacy concerns have increased. In other respects, however, there is a dramatic increase in the availability of new social science "data" created by the increasing feasibility of turning administrative data and other information collected in the course of non-research activities (e.g. scanner data from stores, on-line purchases and contributions, on-line searches) into research databases. This possibility creates its own challenges in terms of privacy protection, data custodianship, maintaining the provenance and context of the data, and new demands for computational power with both increases in dataset size and detail (e.g. detailed geography). The combination of increased computational capacity and the availability of such databases has led to developments in new subfields of each of the social sciences, including agent-based modeling, local area analysis, and network analysis.

Natural Language Processing Research

The state of the art in robust, practical natural language processing depends tremendously on the availability of language models. These models are built automatically from gigabytes of data and reflect some statistical summaries of the frequency in which words appear together in naturally occurring text, both in one language or across languages. Google has been able to provide to the research community a collection of language models called the Google Web 1T 5-gram corpus which includes 1 terabyte of data. To be able to access such data in practical situations, e.g., in the development of novel NLP methods, a research group would need a parallel cluster and high-throughput data storage. Without such a framework, research beyond the state of the art in natural language processing (including speech processing) and information retrieval will be limited to industry sites only.

The ACLS cyberinfrastructure report provides a number of significant use cases for the humanities and even the social sciences. More recently, however, mathematics doctoral students at Harvard did some extraordinary work on language change with the corpus of content digitized by Google. Erez Lieberman, JB Michel and others published "Quantifying the evolutionary dynamics of language" in *Nature* in 2007 (http://dx.doi.org/10.1038/nature06137), and have continued doing work in this area. :

Biomedical Research

The impact of data-driven research is quite significant for computational biology. In fact, the growing body of data is rather unwieldy, with inconsistent data formats and poor cross-talk between different subdisciplines. All forms of biomedical research are driven by the collection, management and analysis of heterogeneous primary and secondary data in digital form. Biomedical research that leads to discovery which translates into routine standard of care requires institutionally supported systems for building and sharing data sets that comply with federal regulation, local policy, and the many systems from which these data are derived. A successful strategy for integrating these data must address governance, leadership, business model, regulatory requirements, data sharing policy, IT infrastructure and necessary tools and technologies.

The number of key databases and resources used in biomedical research has grown exponentially from 300 to over 1,000 since 2000.[2,3] This trend can be expected to continue. The size of many of these databases now grows exponentially at a rate equal to or exceeding that of Moore's law. The time it takes GenBank, the principal database of molecular level sequence data, to double in size is now on the order of 18 months, a trend that has been in place since its inception in1982[4]. Other important databases have similar rates.[3] As of February 2010 GenBank contained 112,326,229,652 base pair of sequence data, and its distribution size was 432 gigabytes. It should be noted that many of the core biological datasets are not amenable to simple relational set theory based search approaches; rather they require searching with techniques derived from fields such as signal processing. This has implications not just for storage, but also the networking, and computing capabilities and memory architectures used to support the routine use of this data in research.

The cost per megabase of finished sequence, one of the key gold standards in the community, and widely seen as an enabler of personalized medicine, is dropping at a rate faster than that predicted by Moore's law. Current sequencing technologies use massively parallel sequencing arrays to sequence genomes, taking substantial computing and data storage resources to do so but still requiring 3 to 5 days per genome. New technologies under development both at Michigan and elsewhere use nanotechnology approaches to perform genomic sequencing in near real time, and can be combined into massively parallel arrays, resulting in an enormous cost reduction, a massive increase in data of finished sequence, and a massive data bandwidth issue involved in the handling of such data. In addition, other key biotechnologies such as methylation analysis and genome wide RNA expression analysis were predicted by Hartwell and Lander, in a report to the

[2] Galperin MY. The molecular biology database collection: 2007 Update. *Nucleic Acids Res* 2007;35:D3-4.
[3] Sinha A, Hripcsak G, Markatou M. Large Datasets in Biomedicine: A Discussion of Salient Analytic Issues. *J Am Med Inform Assoc*. 2009;16:759-767. DOI 10.1197/jamia.M2780.
[4] Genetics Sequence Data Bank Distribution Release Notes. Version 176. February 15, 2010. Available at ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt.

National Cancer Advisory Board in 2006, to decrease by at least a factor of 5 over by 2010[5], bringing such data intensive technologies into more routine use in research studies.

The national and international infrastructure for conducting biomedical research is already highly distributed, and growing more so. National scale databases and compute resources of great importance to the research community exist at a number of key locations, including the National Library of Medicine, the National Cancer Institute, and the European Molecular Biology Institute. Increasingly the services available at these sites are integrated in real time in software used at the University of Michigan and essential to research. Examples of such tools can be found at the National Center for Integrative Biomedical Informatics (NCIBI) at the University[6]. The NCI in particular, through the Cancer Biomedical Informatics Grid (caBIG)[7,8], is developing semantic services for mapping of data elements that depends on the NCI metathesaurus, as well as core intelligent data objects for representing and working with biomedical data (caCORE)[9].

Funding agencies are beginning to require use of distributed infrastructure frameworks (typically Grid-based computing) in funding for various grants and contracts, along with data sharing agreements based on these frameworks. For cancer research, use of the caBIG tools and services is now required on many grants and contracts from the NCI. Frameworks such as caBIG have a dynamic discovery capability, which means that investigators at other institutions querying datasets on the grid will automatically discover and be able to query data that exists at the University of Michigan, and visa versa.

Funding agencies, in conjunction with refereed journals, are increasingly establishing and demanding submission of well annotated data sets to national level or journal-based archives as a condition of funding and/or publication. Some of these data sets are large, in the terabyte range. As of April 1, one of the leading journals in proteomics, *Molecular and Cellular Proteomics*, will require all data sets associated with a manuscript be put in the public domain as a requirement of publication. Other journals, including *Science*, *Nature*, *Cell*, etc. are anticipated to follow next year[10].

National efforts are also occurring to normalize the nomenclature and representational aspects of certain types of research data. In cancer research this is particularly true with the data standards in the NCI metathesaurus, the NCI cancer data standards repository

[5] Hartwell LH, Lander ES. Report to National Cancer Advisory Board: NCAB Working Group on Biomedical Technology. Available at http://deainfo.nci.nih.gov/Advisory/ncab/sub-bt/NCABReport_Feb05.pdf

[6] See http://portal.ncibi.org/gateway/tryourtools.html

[7] Beck JR, and the caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid(TM): Infrastructure and applications for a worldwide research community. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems: 2007*. 2007;330 - 334.

[8] See also https://cabig.nci.nih.gov/

[9] Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. caCORE: A common infrastructure for cancer informatics. *Bioinformatics*. 2003 Dec 12;19(18):2404-2412.

[10] Private communication, Dr. Philip Andrews, Professor of Biological Chemistry, University of Michigan

(caDSR) of common data elements, and by use of domain analysis models such as BRIDG[11] for representing data objects, workflow, and vocabulary. Rules surrounding these elements are enabling caBIG, a semantic grid infrastructure used for cancer research. This infrastructure is now beginning to be extended to other NIH intra and extramural research programs. Development of some degree of semantic services and knowledge base infrastructure at University of Michigan seems an appropriate cyberinfrastructure task.

Finally, there is pressure emerging from funding agencies and journals for depositing raw datasets in national or even international repositories. As noted earlier, beginning on April 1, *Molecular and Cellular Proteomics*, a leading journal in the field will require such deposition. Dr. Andrew's Tranche, part of the *ProteomeCommons*, will be providing this infrastructure. The Tranche system is meant for providing researchers with a tool for resource sharing of proteomics information and data. The system incorporates publication support, data archiving, data provenance, and fine-grained permission control incorporating concepts such as publication escrow. Data is stored in a secure, encrypted fashion, and data replication is done between collaborating sites using a peer-to-peer networking protocol to protect against loss of data. The system can effectively operate as a laboratory information management system for project and data management, and provide the backend capabilities to support data deposition and the publication process. To truly address the needs of our research community for cyberinfrastructure requires us to embrace and provide not just data storage and compute capabilities, but flexible, scalable data management "middleware" and systems that are coupled with the workflow of what researchers do in their laboratories. If we fail to do this as part of the cyberinfrastructure of the University, we will be leaving valuable data and intellectual property "on the table". We should work to define and widely deploy cyberinfrastructure systems such as Tranche that support the needs of the community for next generation data management, and which incorporate new and emerging paradigms for scientific data sharing and publication.

Engineering Research

The recent increase in computing power has lead to a movement in engineering towards the direct integration of computational methods with experimental observations. In the last two years alone, the National Science Foundation and the National Academies published reports[12,13,14,15] outlining the current state of affairs.

---

[11] Fridsma DB, Evans J, Mead CN. The BRIDG Project: A Technical Report. *Journal of the American Medical Informatics Association*. 2008;15(2):130.

[12] Peter Cummings and Sharon Glotzer, Inventing a New America through Discovery and Innovation in Science, Engineering, and Medicine: A Vision for Research and Development in Simulation-Based Engineering and Science in the Next Decade. Report to the National Science Foundation, January 2010.

[13] Sharon Glotzer et al., International Assessment of Research and Development of Simulation-based Engineering and Science, WTEC Panel Report, (2009) (http://wtec.org/sbes/SBES-GlobalFinalReport.pdf).

[14] Diran Apelain, "Integrated Computational Materials Engineering (ICME): A "Model" for the Future?" *Journal of the Minerals, Metals and Materials Society* 60 9 (2008) http://www.springerlink.com/content/c2437206p72134q1/.

An excellent user case of this is the program developed by Ford Motor Company "to develop and design virtual aluminum engine castings in partnership with researchers at several universities. This design process would normally take years to complete on physical engine parts at considerable cost. Instead, ICME accelerates the development process by 15-25% and provides an estimated 7:1 return on investment through cost avoidance and savings." (from Cummings and Glotzer)

Another opportunity to advance engineering research is in mining of databases of existing data to find new relationships between them. The experimental method dictates that experiments are designed with a particular hypothesis in mind. Often in research, however, discoveries are made when unexpected relationships are drawn across several different data sets. Furthermore, analysis of very large or complex data sets requires tools that we have yet to fully develop. There are several examples of attempts at creating large databases for various engineering fields:

- Materials Digital Library Pathway: http://matdl.org/repository/index.php
- nanoHUB: http://nanohub.org/
- iPlant Collaborative: http://www.iplantcollaborative.org/
- Combinatorial Sciences and Materials Informatics Collaboratory (CoSMIC) http://www.mse.iastate.edu/cosmic/what2.html
- InfoChimps http://infochimps.org/
- Sloan Digital Sky Survey http://www.sdss.org/collaboration/

The fact remains that data handling, archiving, tagging, and reuse is still in its infancy. Researchers are left to their own devices in organizing and archiving their data. A coordinated and rigorous approach to data management can lead to large increases in the rate of scientific discovery.

Data Intensive Astronomy

Modern astronomical instrumentation produces images which double in size every few years. Two decades ago, the largest single astronomical images were of the order a few megabytes on a single two megapixel CCD. Today, the CCD detectors have doubled in size and astronomers have mastered the ability to create CCD mosaics to create gigapixel images. The software astronomers use to turn the raw data pixels into scientifically useful (and often pretty) images is community driven and usually designed for desktop implementation. However, the size of modern astronomical instrumentation is moving astronomers beyond the capabilities of the multi-core desktop.

In the next 5-10 years, astronomers, astrophysicists, and physicists will be routinely utilizing petascale object catalogs of stars and galaxies. These catalogs are generated from the world's astronomical observatories, which routinely collect image and spectral data on the sky. Sky surveys like PanSTARRS and the Dark Energy Survey will create

[15] Pollock et al., Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security, National Academies Press (2008) http://www.nap.edu/openbook.php?record_id=12199.

databases of billions of objects, each with hundreds of measured or inferred properties. The Large-Scale Synoptic Survey telescope will be yet a magnitude larger. A researcher's competitive edge will hinge upon his or her ability to query, access, analyze, and share these catalog data quickly and efficiently.

The cornerstone of the physical interpretations of all of these data comes from the theoretical simulations designed to match what is seen in the real Universe. Already, these simulated stars, galaxies and universes are pushing the envelope of compute resources. The simulations often require teragrid scale computing to create them and dedicated high performance computer clusters to analyze them. The simulations must then be integrated with the real data in a coherent fashion to enable fundamental scientific breakthroughs.

Most researchers at the University of Michigan work in multi-national collaborations. For astronomy, this brings new challenges, since the data are often spread throughout the world in heterogeneous archives, each with their own unique query and access mechanisms. The facilities astronomers use to collect these data are usually isolated and the bits need to be transmitted across oceans and continents before they can be analyzed. Raw pixels need to be processed and object catalogs need to be generated and then merged with rest of the world's astronomical archives. Only then can astronomers study objects individually, via representative samples through statistical distributions, or by cross-correlating between many millions of objects, each with hundreds of attributes. Scientific interpretation is made in the context of even larger simulated datasets created in the lab. This is the future of data intensive astronomical research.

High-Energy Physics Research

High energy Physics (HEP) explores the basic laws governing the interaction of matter and energy in our Universe. One of the primary ways we do this (experimentally) is by colliding particles in accelerators and carefully measuring and recording the results from large, complex detectors assembled for this purpose. As HEP has progressed, the energy threshold for new scientific results and understanding has also progressed, as well as the corresponding volume of data generated. In 2010, at the start of the Large Hadron Collider (LHC) era, we find ourselves preparing for 10's of petabytes of data per year to be processed, annotated, archived and distributed to thousands of Physicists worldwide. Within 5 years we plan to have generated and archived almost an exabyte (1000 petabytes) of data from the LHC and related simulations.

In some sense the use of the LHC data is very simple and straightforward. We are looking for a small number of anomalous "events" in a huge sea of normal events. This is typically referred to as the needle-in-a-haystack problem. For HEP to make progress we need to be able to access and process large amounts of data with (potentially many) algorithms, looking for new physics and validating our existing understanding. To do this effectively will rely upon a number of capabilities:

- The ability to track the location and accessibility of datasets worldwide.

- High-performance WAN (wide-area network) access to large datasets, i.e. the ability to move or access large datasets in "human" timescales .
- Sufficiently detailed metadata to understand both data provenance and data content.
- Appropriately scaled "local" storage systems to allow effective processing and analysis of this data.
- The ability to curate resulting data products long-term (data for scientific papers, discoveries, codes and algorithms, tools, metadata)

None of these capabilities are unique to high-energy physics, rather it is just that HEP is perhaps hitting these issues slightly ahead of other data intensive science domains. In that sense, there exists an opportunity to utilize HEP to pilot data intensive infrastructure solutions. Conversely, if general solutions to any of the above 5 capabilities were developed they should be easily adapted for use in HEP.

Looking forward 5 years, we would like to see the University of Michigan able to provide data intensive scientists the capability to quickly and effectively access and manage their datasets of interest, whether or not the data itself is hosted locally or externally. We should be planning to routinely be able to handle datasets at the 100's of terabytes in size with access to federated and distributed datasets in the exabyte range. Tools to "curate" data/datasets should be widely available on campus and centrally supported including the resilient, archival storage needed to keep critical results and derived datasets.

Networking infrastructure, tools and support should allow terabit/sec external data access should be in place, certainly if not within 5 years, within 5-10 years. Corresponding "cache-like" storage capabilities related to moving data into or out of the University at high data-rates should be developed and put into production.

Appendix C - Examples – Value to research of good data management

Use Case #1: Metadata schema for media-rich content

It is very difficult to search audio and video data because the data are multi-dimensional, there are multiple metadata schemas, music information retrieval is underdeveloped, and the human skills necessary to formulate oral queries introduce error. Considering audio, it is commonly held that there are four fundamental variables that describe sound: frequency, amplitude, time, and timbre. Although each of these variables is theoretically infinite, they are limited by both human perception and the device that produces the sound. These variables are further complicated when the sound source is not monophonic. Multiple impoverished metadata schemas such as ID3 and Dublin Core are currently in use. The application of these schemas to an audio artifact is time consuming and requires significant knowledge. Research in Music Information Retrieval (MIR) indicates that the more complex the music, the less likely it is that an oral query will yield usable results.

The field needs is a metadata schema that describes all salient attributes of an artifact using a schema that is automatically generated and embedded in a file format that supports multiple modes of query such as text, singing, or optical recognition. There would be a corresponding method for human validation of the integrity of the automatically generated schema that would be used by multiple content delivery agents ranging from the library to the iTunes Store. Each derivation of an artifact would have its own unique schema and the schema would track the lineage of the derivation. Range variant searches across multiple artifacts would yield relationships among schema that would motivate human inquiry about collections of artifacts.

Use Case #2: Human perception and cognition of large-scale data sets

Sensory input in our daily lives is vital to our safety and our capacity to interact with the physical world. But perceptual modalities such as sound, touch and smell are under utilized in our comprehension of large-scale data sets. We need to explore human perception and cognition within the context of the philosophy of aesthetics to yield a cogent understanding of and interaction with large-scale time-variant data sets. This exploration should not be limited to the physical world but should explore the capacity for human perception and cognition in virtual reality. Ideally, a gaming environment encourages exploration that yields the advancement of knowledge.

Use Case #3: Documentation for historical social science data

About 15 years ago, two researchers (at the time they were on the faculties at Stanford and Harvard) received NSF funding to create a dataset from Census of Manufactures that were done by the U.S. Bureau in the 1930s. These Censuses had never been put into machine readable format; the extant paper records are in the National Archives in

Maryland. Assisted by several graduate and undergraduate students, the researchers coded a small portion of the data in the Census, covering automobile related industries. The data were put into now obsolete database and spreadsheet programs and stored on now obsolete media. They wrote two or three papers based on the data. They prepared no documentation of the dataset. The data are clearly valuable in and of their own right. They are also potentially valuable today as part of a larger project to code the rest of the manufacturing censuses of the period, link those censuses to longitudinally linked Economic Census data from the post-WWII period, and link them to other contemporary datasets that have since been created (e.g. firm credit histories from Dun and Bradstreet, import –export data, etc.). Committee member Margaret Levenstein obtained the various floppy disks and tapes on which the data were stored. ICPSR has been able to extract most, but not all, of the data and store it in a durable format. We are in the process of creating documentation for the data, but that requires relying on the memory of the PIs and former graduate students, which has not been easy. There is no question about the provenance of the data, but there is ambiguity about its place in the public domain. Census has no rules regarding making public business data (unlike household data, which legally becomes public after 72 years), but the data is *de facto* public since its transfer to the National Archives. The extant dataset is small and can easily be analyzed on any existing computer, but its potential use has increased because of the potential to link it to later, must larger, datasets as well as the entire census from the period. This also requires that the two datasets be organized, cleaned, etc. in a similar fashion, and that is an enormous challenge because one dataset is created by academics and the other by the U.S. Census Bureau.

Use Case #4: Confidentiality of Census data

A doctoral student in survey methodology is writing his dissertation on a confidentiality protection technique that uses synthetic data to create small area estimates from the American Community Survey and the National Health Interview Study. These are both large, national surveys that are used to produce estimates of demographic, economic, and health outcomes for the nation and its constituent parts. Despite their large size, producing local area estimates raises confidentiality concerns. For example, it takes three years of data to produce estimates of population, etc. from ACS for locations of under 60,000 residents and no estimates are produced for areas of under 20,000 residents. This student is creating synthetic data, based on actual ACS and NHIS data, and using that synthetic data to produce estimates for small geographic areas that would not meet confidentiality protection limits. He will then evaluate the quality of the estimates produced from the synthetic data. Because of the confidentiality of the data, all analysis is conducted within the Michigan Census Research Data Center on servers at US Census Bureau. Because of the computational intensity of the programs to create synthetic data, we had to have this project moved to a different server (purchased with NSF funding for a different synthetic data project) than we usually log into from the MCRDC. That required negotiation with Census and other researchers at other research data centers and our researcher was given access to these computing facilities for a very limited period of

time (which we are ignoring and counting on Census bureaucracy not to notice when we have passed that time limit).

Use Case #5: Privacy concerns with integrated data sets

We have been asked by the Sloan Foundation to prepare a proposal to link the Health and Retirement Survey (a panel study of older adults conducted by the University of Michigan's Survey Research Center) to the Longitudinal Employer Household Dataset (a US Census Bureau dataset of all employers linked to employees, with earnings histories for each employee). Linking these two large, complex, confidential datasets has enormous potential for research on labor markets, especially the relationship of an aging population/workforce and labor markets. This project has been under discussion for at least five years and is regularly stymied both by concerns about confidentiality protection and concerns about the appropriate use of confidential information. For example, HRS respondents have given us the name of their employer and their SSN, but the permission that they grant with that information specifies what the information can be used for. It does not specifically say that it can be used to link to LEHD, so both the funder (NIA) and HRS have been very reluctant to permit the linkage. Access to both these data sets is restricted, to protect confidentiality; researchers access to a linked dataset would have to meet the restrictions of both data providers, and the providers would have to agree to accept some of the confidentiality protection of the other provider (i.e. the dataset would not be available in HRS's restricted use enclave, but only in a US Census Bureau restricted use enclave). This increases researcher burden and reduces the usefulness of the linked dataset.

Use Case #6: Data-mining application in Biomedical Science

Genome mining and the search for evolutionarily significant patterns in genetic information: size of datasets is huge – for EACH species – and it is frequently unknown which regions of the genome encode important information.

Protein structural mining and the search for evolutionarily significant patterns in structure-function relationships: patterns are numerous and frequently unknown, many are pre-computed with different applications in mind.

Systems biology couples multiple forms of disparate data – toxicology, high-throughput screening, cellular localization of metabolites and proteins, genomic screening (up-regulated versus down-regulated expression patterns).

Protein modeling simulations include:

- Protein folding predictions (CASP competitions, folding@home initiative, etc)
- Simulations of very large systems of macromolecular complexes (ribosome, telomerase, cytoplasm…)
- Simulations based on polarizable force fields (or QM/MM)

- Docking and scoring for structure-based drug design: in its grandest application, molecular databases are millions of compounds and docking those molecules with all drug targets in the human body and in pathogenic organisms leads to petabytes of information… incorporating protein flexibility significantly increases computing and storage issues… Protein-protein docking does not have as many potential pairings, but the degrees of freedom are 3-orders of magnitude higher than docking small molecules

High throughput technology and large scale genetic screening is in a state of relative infancy but already changing the fashion in which research is being conducted. The detection of Single Nucleotide Polymorphisms (SNPs) as basic genetic markers for linking underlying genetics to observed disease phenotype is now taking place on a large scale. The ability to aggregate and analyze this data successfully rests on the ability to collaborate and integrate data from multiple disparate groups, exchange large data sets while adhering to regulatory constraints, and to access and integrate substantial international scientific resources (such as GenBank, PDB, KEGG, etc.) to support the process.

Much has been written about the impact of genomics on the ability to perform "personalized medicine" and to conduct comparative effectiveness research. Key to this task will be the merging of data from the electronic medical record with other sources such as population risk factors, personal behaviors and traits, family risk assessment data, longitudinal survival data, and genetic profiles, to name but a few. Integration of these data will demand access to knowledge bases and ontologies necessary to integrate these diverse data sets, all of which are collected with different research foci.

The ability of cancer researchers and bioinformaticians to exploit all the information sources discussed above in a meaningful way depends on a variety of key cyberinfrastructure components, and demands recognition of the manner in which the rapidly evolving data, regulatory, and funding environments impact our needs for cyberinfrastructure support.

Use Case #7: Laboratory project and data management, and publication support

Dr. Philip Andrews, Professor of Biological Chemistry, says, "In Proteomics, our data sets are quite variable in size, reflecting the nature of the experiments and the mass spectrometry technologies used. Several proprietary file formats are currently in use and while we do have a standard format available now it is not yet broadly used and leaves behind some critical information from the original formats. Our laboratory generates 4-5 TB of data per year in a variety of formats. Other proteomics researchers on campus probably generate a similar amount of data. We do expect this amount to increase as higher speed technologies come online. So this is not a huge amount but there are a range of file types and keeping track of them is a challenge without an appropriate data system. The biggest issue is data management in the context of project management. We are switching over to using proteomecommons.org (developed by Dr. Andrews) for this

now from our own LIMS. Since we are dealing with raw data from a variety of instruments, we use a number of software tools to process, interpret and visualize the data sets, so the computational component of the system needs to be at least partially integrated with the storage system. An added aspect of proteomics data is that as of April 1, the leading journal in our field, *Molecular and Cellular Proteomics* will require all data sets associated with a manuscript be put in the public domain as a requirement of publication. Other journals, including Science, Nature, Cell, etc. are anticipated to follow next year."

Use Case #8: Core Facility Support

Dr. Bob Lyons, Assistant Professor, Biological Chemistry and Director, DNA Sequencing Core, says, "Our needs are immediate, depending on technology that is currently in place and in use, or else is highly likely to be in place within a few months. "
"We currently have three ultra high-throughput DNA sequencers that are each generating data at a rate of roughly 0.6 TB/day. A fourth instrument is expected in mid-April. Upgrades to three of the four machines are expected perhaps as early as mid-summer. Those upgrades will roughly double the daily data production for those three instruments, giving us an estimated data production rate of between 3 and 4 terabytes per day. That data production rate, if all of it is transferred to our clients (which indeed is usually the case) represents roughly half of the maximal theoretical capacity of our local gigabit network. In terms of real-world network performance, it probably represents a much greater fraction of capacity, and will probably have some deleterious effects on the local network resources."

"We store all of the data we generate for at least 1 month, and we need to store some of it for up to 3 months (the exact fraction is unpredictable and depends on our clients' handling of the data). Our short-term storage needs are therefore on the order of 120-250 TB. However, some additional space is necessary simply to allow for the ebb and flow of data, so 150-300 TB would be more realistic. Longer-term storage is also needed: About 10% of the total data stream should be stored essentially forever. That equates to roughly 150-300 TB per year."

Use Case #9: The History of the Milky Way Through the Orbits of its Stars

Stars make up the bulk of the observed light in galaxies. We can infer intrinsic stellar properties like their masses, temperatures, metal abundances, surface gravities, among others, from their observed colors and spectra. Recently, astronomers have found what appears to be a correlation of these properties with their orbits around the Galaxy. Understanding this correlation can unlock the formation and evolutionary history of our Milky Way. Hidden within the high-dimensional phase-space of stars lies the dynamical memory of the formation of our Galaxy.

Current datasets are computationally challenging to work with, since the orbits of tens of thousands of stars must be determined simultaneously. In the near future, these stellar datasets will increase in size by an order of magnitude. Researchers need the computational infrastructure to model these orbits and the tools to analyze and visualize the complex and high-dimensional results. The goal is to unlock the dynamical formation history of the Milky Way galaxy in order to learn about past collisions and mergers between our own Galaxy and our neighbors.

Use Case #10: The Formation and Evolution of Galaxies

Galaxies are often categorized as either old, dead, and red or young, alive, and blue. The former have featureless elliptical shapes and have long ago lost most of the fuel they need to make new stars. The latter have spiral arms and are full of gas and dust: the ingredients of stellar birth. Stars are born when this gas and dust is compressed. This compression can occur via many mechanisms, which usually involve some sort of dynamical nudge. Hidden in the stellar light of galaxies is the history of its current and past star-formation. The colors, shapes, and intrinsic brightnesses of galaxies can explain when the nudges occurred and what might have caused them in the first place. This is the study of galaxy evolution.

To solve the many mysteries of galaxy evolution, astronomers use large statistical samples containing millions of galaxies. They correlate the many hundreds of observed properties, searching for the clues which lead to new insight about when galaxies were born and how they have evolved. Astronomers continue to create new sky surveys to collect larger and larger datasets. The current state-of-the art is the Sloan Digital Sky Survey with 200 million galaxies. Soon to come are data from Pan-STARRS and the Dark Energy Survey and finally the Large-Scale Synoptic Telescope with billions of galaxies each measured at a 1000 different epochs in time.

In the past, only a single flux measurement was observed at each pixel in the CCD. Or perhaps the spectrum of a single spatial position in a galaxy was observed. Today, instrumentation is evolving into new Integral Field Units (or IFUs), which collect the spectrum of light in a spatial map. The simple analogy is that astronomers have developed the equivalent of 3D medical imaging for galaxies. Now and in the future, astronomers need to manipulate data cubes that spectrally slice through galaxies at many locations. These spectral tomograms need to be analyzed individually and statistically over large samples of millions of galaxies. Astronomers need the facilities to correlate larger and larger datasets each of which is growing more complex. They need to be able to visualize the high-dimensional data and their correlations. They need to manipulate new 3-D data cubes, perform modern statistical analyses, and share the results in such a way that others can incorporate the knowledge for even more detailed analyses. Galaxies are complex beasts, but deciphering the physical processes which govern their evolution is one of the most important challenges of our time.

Use Case #11: Galaxy Clusters and Cosmology

Clusters of galaxies are excellent tracers of the dark matter and baryons in the Universe and are an important ingredient in the famous "Cosmic Triangle" which delineates dark matter, dark energy, and the space-time geometry of the Universe (Bahcall 1999). The study of the parameters which govern the birth and expansion of the Universe is called Cosmology, and galaxy clusters play an important role.

In order to use galaxy clusters for precision cosmology, astronomers are required to know the details of their observational selection at a level of precision that is greater than the final statistical analysis. Current and future galaxy catalogs require significant computational resources to perform the clustering. The precision of the selection process can be measured in a controlled environment through mock universes, each with known ingredients and each larger than the real Universe. This methodology provides a built-in learning and prediction feedback loop: mock universes are created to tune the cluster selection in real data and the real data are used to improve the realism of the mocks. The final precise list of clusters in the Universe is then analyzed to infer the values of cosmological parameters like the amount of dark energy or dark matter in the Universe.

Cosmologists want to treat these mock universes in the same statistical vein as they do mathematical models, where parameter dependencies can be systematically varied. The only way to achieve this is through Teragrid-scale resources utilized to create the many mock universes. However, the analysis of even a single universe is already at the limits of multi-core desktop computational resources. As the observed data increases by an order of magnitude every 5-7 years and the mock universes keep pace, cosmologists will need new computational infrastructure that enables them to conduct integrated analyses of the mocks and the real data. In the end, they need environments which seamlessly connect the world's largest super-computing facilities with local high performance clusters, and which continues all they way to their multi-core desktop.

Use Case #12: High-Energy Physics Data

Data in HEP comes in two types: simulated data generated by detailed Monte-Carlo methods and "real" data coming from the experimental apparatus. Both types of data are needed to undertake HEP science and both types of data for the LHC are in the multi-petabyte/year range.

The current HEP infrastructure for LHC has developed a Tier-n data hierarchy with the "Tier-0" located at CERN (Geneva, Switzerland, location of the LHC) with a number of national scale Tier-1 sites around the world. Associated with each Tier-1 is a number of regional Tier-2 centers (around 100 worldwide). The real data produced by the experiments follows a well determined path from the experimental apparatus to the Tier-0, out to the Tier-1 sites and eventually some summarized data products going to the Tier-2 sites. While real data is being collected there are also large-scale simulation runs going on at Tier-2 sites (and some Tier-1s) which also must be processed and distributed in

parallel. Outside of the formal data hierarchy are also institutional Tier-3 centers which support local physicists data analysis needs.

While the real and simulated data flows are well understood and predicted there is potentially much more data flow and access associated with 1000's of physicists world-wide analyzing datasets of interest. The impact of this potentially chaotic and large-scale effort is one of the greatest uncertainties for our distributed infrastructure.

One of the first tasks LHC physicists need to do is to calibrate our complex experiment's components. There a millions of electronic channels of differing types with an LHC experiment like ATLAS. Without a detailed understanding of the response of the detector components and their associated electronics, triggering and data processing steps, physicists will be unable to determine if interesting results from the data come from new physics or poorly understood data artifacts! The step of calibrating the experiment and its corresponding data processing chain is a data and compute intensive effort involving both real and simulated data. For LHC we foresee at least 1 year to do this. This type of activity is not unique by any means to HEP. Calibrating and understanding experimental components in detail is required in any experimental science.

After the LHC experiments are well understood things get interesting. Many different groups of physicists will be vying to mine the data searching for new physics in a large number of possible areas of interest. The ability of physicists to search for data matching specific criteria, move or access that data and then process that data with their algorithms will determine how fast new discoveries can be made. Problems in any part of this chain (data lookup/discovery, access, processing) can be the bottleneck to scientific discoveries.

Currently HEP collaborations (virtual organizations (VOs)) have created tools to help create and manage metadata concerning the large amount of real and simulated data. These tools and frameworks are not as well developed as other codes and applications within HEP and certainly suffer from fragility and slow performance in many cases. Improving the tools in this area would be a significant boon to data intensive science.

Once datasets of interest are identified by the VO-specific tools there is the separate issue of how a physicist or group of physicists access that data. Because of the potential size of data involved the general idea is to send the algorithms to the data, rather than bring the data to the algorithm (local CPU farm). However this doesn't always work well either. Sometimes the data is located in places which have no currently available CPU resources to process with. Having a high-performance data distribution infrastructure in place is a critical component of providing an effective environment to do data intensive science and decreasing time-to-discovery.

One thing that should be noted is that while discovering, moving and processing data are all vital to the data intensive scientist we also need more and better tools in the area of data curation. How will we document and persist our important scientific results and

their corresponding raw and processed data? This is critical to enable even further scientific discoveries from combining with other data resources in the future.

Appendix D – User Needs

- Data Recording Challenges
    - Data are often spread throughout the world in heterogeneous archives
    - High speed instrumentation will almost certainly require some types of high-speed data caching arrangement in relatively close physical proximity to the instrumentation.
    - Expensive simulations can also create large data sets.
    - There is pressure emerging from funding agencies and journals for depositing raw datasets in national or even international repositories.
- Documentation and Metadata Challenges
    - How will we document and persist our important scientific results and their corresponding raw and processed data? This is critical to enable even further scientific discoveries from combining with other data resources in the future.
    - Tools to "curate" data/datasets (data for scientific papers, discoveries, codes and algorithms, tools, metadata) should be widely available on campus and centrally supported including the resilient, archival storage needed to keep critical results and derived datasets. Conversely, there should be tools/middleware that will permit tracing from published work to the raw data and experimental methods from which published results were derived.
    - To-date High-Energy Physics (HEP) virtual organizations (VOs) have developed tools to create and manage metadata for the large amount of real and simulated data they produce. These tools and frameworks are not as well developed as other codes and applications within HEP and certainly suffer from fragility and slow performance in many cases. Improving the tools and best practices in this area would be a significant boon to data intensive science.
- Data Distribution Challenges
    - The ability to track the location and accessibility of datasets worldwide.
    - Frameworks such as caBIG have a dynamic discovery capability, which means that investigators at other institutions querying datasets on the grid will automatically discover and be able to query data that exists at UM, and vice versa.
    - High-performance WAN (wide-area network) access to large datasets, i.e. the ability to move or access large datasets in "human" timescales.
    - There is large, fragmented investment in data and software infrastructure.
- Heterogeneity Challenges
    - Inconsistent data formats and poor cross-talk between different sub-disciplines.

    - Each data set with its own unique query and access mechanisms.
    - It is very difficult to search audio and video data because the data are multi-dimensional, there are multiple metadata schemas, music information retrieval is underdeveloped, and the human skills necessary to formulate oral queries introduce error. The field needs is a metadata schema that describes all salient attributes of an artifact using a schema that is automatically generated and embedded in a file format that supports multiple modes of query such as text, singing, or optical recognition.
    - In support of basic scientific computation, provide a single mass storage **interface** that is easy to use, secure, and efficient. (SRM [https://sdm.lbl.gov/srm-wg/] and SRB [http://www.sdsc.edu/srb/index.php/Main_Page] are attempts)
- Data Analysis Challenges
    - Domain experts may need support.
    - Data distribution and heterogeneity exacerbate problems.
    - Increasingly difficult to find data even after it has been stored.
- Data Persistence Challenges
    - Today, critical data is often stored on workstations
    - Need safe persistence, but also need availability.
    - Need appropriate combination of backup and redundancy.
    - Not everything needs to be saved, and some things that need to be saved only need to be saved for a short period of time (rather than "forever").
    - Reproducible data can often be thrown away after some period of time. What this period is may very across fields. This discarding of old data may actually aid in progress in some fields, for example because much better instruments become available, or much finer granularity data can be stored. Of course, irreproducible data should never be thrown away, and most data may be irreproducible in some fields, such as most social sciences. There is a more subtle challenge of identifying precise what is reproducible, since reproducibility may require some assumptions. For example, astronomy movie data. But the same thing may apply whenever someone wishes to study that which had previously been considered immutable. For example, a study of change in allergic response due to industrialization may need data from an old allergy trial that would normally have been considered reproducible scientific data.
    - Where archiving is relevant, the efforts need to take into account the great deal of thought that's been done on what makes archiving effective. For example, service and storage architectural models, preservation metadata (e.g., PREMIS), data structures, all need to be in the mix.

    - Use of legacy equipment (i.e. computers and other computer-based research devices that use obsolete applications and operating systems that are inherently insecure).
        - Inability to replace equipment
        - Inability to read data formats
        - Unusability of backups
- Size Challenge
    - We should routinely be able to handle datasets at the 100's of terabytes in size with access to federated and distributed datasets in the exabyte range.
    - GenBank, the principal database of molecular level sequence data, doubles in size is now on the order of 18 months.
- Best practices guidance
    - Increasing user base of computing-novice users, means greater need for consulting and technical support.
    - Data and metadata recording must fit into the domain workflow. The recording tends not to happen right if it is done after the fact, as a separate process.
- Challenges in Ownership, Access Rights, Identity, Etc.
    - Researchers need access to professional system administrators, data storage professionals, and information security personnel for:
        - Regular vulnerability scanning
        - Removal of unnecessary services
        - Closure of unnecessary ports on systems storing sensitive research data
        - Elimination of excessive levels of access privilege by researchers, colleagues or students
    - Lack of training on security and data handling
    - Data classification guidance
        - Data is often not classified leading to data either being over protected because everything is treated like sensitive data or everything is under-protected by treating everything as public data
    - Software licenses (researchers were unsure what was available and what they had to buy)
- Other practical issues
    - Some domains have a feast or famine cycle. Sharing may help, but easier to do with compute than with data.
    - There is substantial cost to maintain data, and particularly convenient access to it, after grant ends.
    - Much data for research may originally be recorded for other purposes. For example, medical records are central to much medical research;

administrative data is commonly used for social science research; lecture recordings may be used for education research; and so on.

- May be worthwhile to think of the overall process as regulation constraining policy choice; policy dictating process and influencing technology choice; we can consider these layers one at a time.

Appendix E - Survey of all LSA research active faculty members

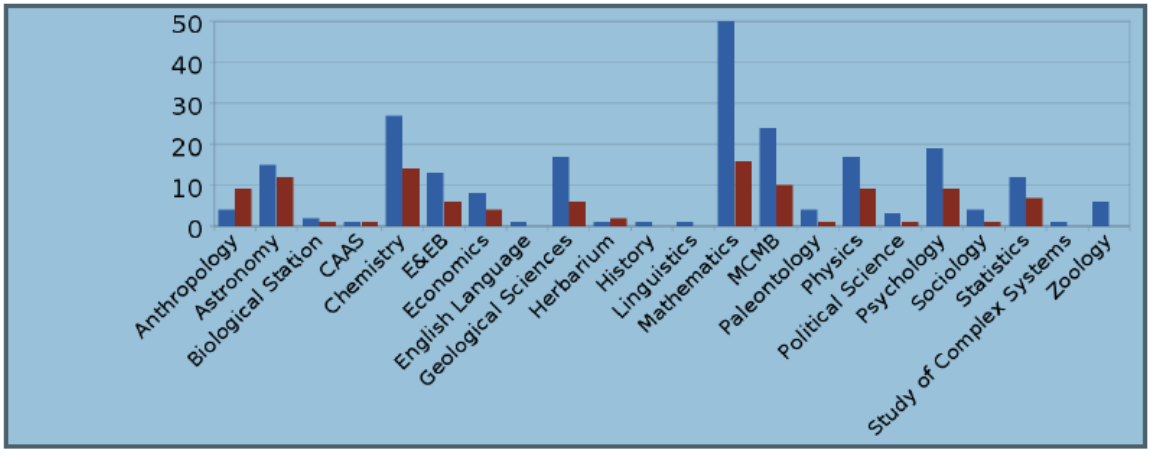
Research Data Storage, back-up, and disaster recovery faculty PI questionnaire
Christopher Miller, Asst. Professor, Astronomy

Methodology

LSA Faculty PIs who have NSF, NASA, or NIH grants were extracted from the Sponsored Awards on the Web (SAW) interface1. Duplicate entries were cleaned and faculty contact emails were correlated against lists of current faculty members via official LSA Department web-pages. A total of 234 Faculty members were then contacted by email and asked to fill out a web form located at: http://wwwpersonal.umich.edu/~christoq/data.html . Approximately one half of these faculty members responded.

Demographics

In all, 111 responses were logged to the web-form. Another half a dozen (or so) individual Faculty responded directly to me stating that they did not use any data storage facilities at U-M and had no need for any. So the official response rate is 50%, even though only 47% submitted the form itself. Please note that the first five respondents answered a form that was slightly modified compared to the rest. In Figure 1, we show the response rate per department for all LSA departments with faculty who were contacted. The blue bar indicates the number of faculty contacted and the red bar indicates the number of respondents. Note that the response rate of ~50% holds for most departments.



Summary of the Results

The full results are shown below. Here, I summarize some of the highlights:

1. 81% of the research groups use less than 2TBs/year for grant funded research and only 4% use above 10TBs annually.
2. 65% create or use files that are 1GB or less.
3. 2/3 of PIs *do not* request purchases in their grants for data storage equipment.
4. Only 10% of the researchers claim that there data *could never* be re-created.
5. 40% of the researchers need to share their data outside of U-M.

6. While most researchers back up their research data, only 27% do so daily or more often.
7. Nearly half of the PIs prefer affordable and centralized back-up services over increased centralized or localized storage. There is no interest in 3rd party solutions.
8. For nearly half of the PIs, it is either required or recommended by their funding agency to provide continued access to their data after the grant has ended.
9. Nearly half of the PIs suggest that 50% or more of their data could be deleted after the project ends.

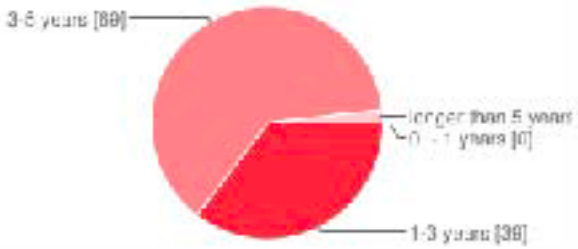
Full Report:

Does your usual funding agency require that you submit a plan for data management for grant projects?



Yes		38	35%
No		55	50%
Not sure		17	15%

What is the typical length of one of your research grant funded projects?



0 - 1 yrs		0	0%
1-3 yrs		39	35%
3-5 yrs		69	63%
> 5 yrs		2	2%

What is the maximum amount of data you collect annually on a typical research grant funded project?

less than 100 GBs		42	38%
100GBs - 1TB		33	30%
1-2TB		14	13%
2-5TB		5	5%
5-10TBs		8	7%
More than 10TBs		4	4%

48

What is the maximum size of any single research file for a typical grant funded project?

less than 1MB		5	5%
1-100MBs		41	37%
100MBs - 1GB		25	23%
1-10GB		23	21%
10-100GB		8	7%
100GB-1TB		7	6%
more than 1TB		1	1%

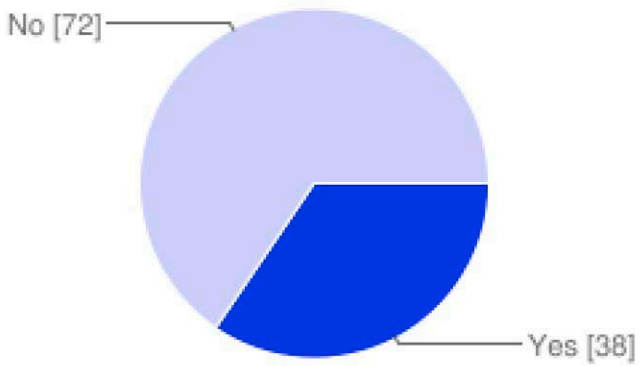
48

How many files do you generate for a typical grant funded project?

<100		33	30%
Thousands		57	52%
100,000s		12	11%
Millions		3	3%
Many millions		1	1%
Other		4	4%

66

Do your typical research grants include equipment purchases which include data storage?



Your grant funded equipment purchases for data storage are typically:

Standard desktops (22) 58%
Laptops (18) 47%
Small servers (i.e., not rack mounted) (16) 42%
Rack mounted servers (14) 37%
Cloud-based third-party services (2) 5%
Tapes or other non-spinning media. (4) 11%
Other (10) 26%
People may select more than one checkbox, so percentages may add up to more than 100%.

Please provide some insight into the value of your data.

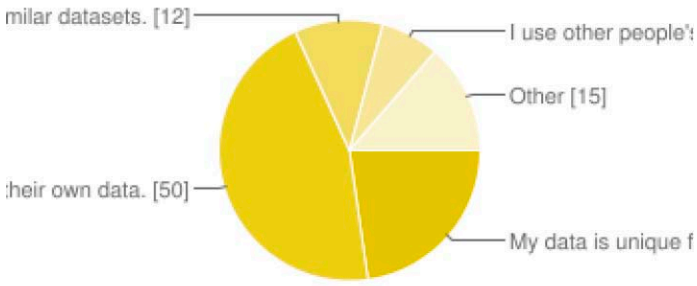
My data is unique forever. The scientific questions I am addressing could not be done so with any other dataset, now or in the future. (25) 23%

My data gives me an advantage for a little while. The scientific questions I am addressing will be done quicker and better with my data alone, but others will soon be ble to do as well with their own data. (50) 45%

My data is similar to others. The scientific questions I am addressing are simultaneously being addressed by others with similar datasets. (12) 11%

I use other people's data to address new scientific questions. Others are using the same data. (8) 7%

Other (15) 14%



Please provide some insight into how difficult it would be to re-create your data from scratch.

It is impossible to re-create my data. If my data were lost, I would not be able to meet my scientific commitments I promised in the grant proposal. (11) 10%

It would take significant time and significant resouces ($$ and/or people) to re-create my data. I would likely not be able to meet my scientific commitments for the grant in the time originally allocated if I had to start from scratch. (54) 49%

Given enough resources, I could recreate my data in a short enough period of time that I could meet my scientific commitments in the grant. (18) 16%

Re-creating my data depends on too many external parameters to estimate how I could re-create it. (5) 5%

My data is fairly easy to re-create. (8) 7%

I use other people's data. (5) 5%

Other (10) 9%

How sensitive is your data to others who want to view it during the period of the granted funded project?

The data I create should only be accessible to me. (19) 17%

The data I create should be accessible only to me and other U-M researchers and/or students. (31) 28%

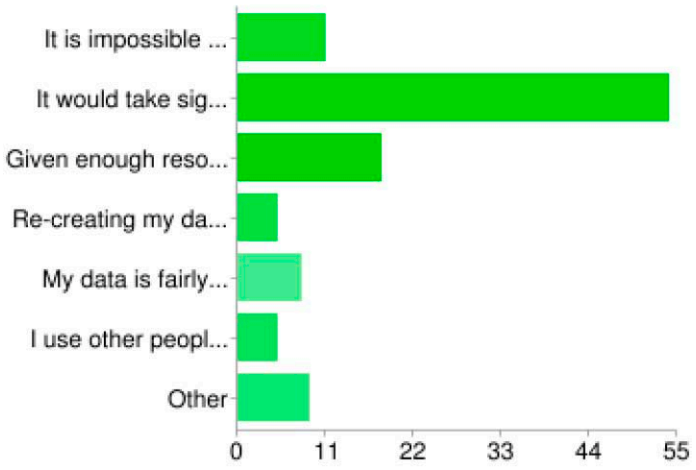
The data I create should be accessible only to me and other U-M AND non-U-M researchers and/or students. (28) 25%

The data I create can be viewed and used by other professional researchers at any location (i.e., with known identities), but not the general public. (10) 9%

The data I create can be viewed and used by anyone at any time during the research project. (7) 6%

I use other people's data and I need to meet the requirements they provide me. (5) 5%

Other (10) 9%



Do you share your data with others in order to conduct the research during the grant project period?

Yes		76	69%
No		34	31%

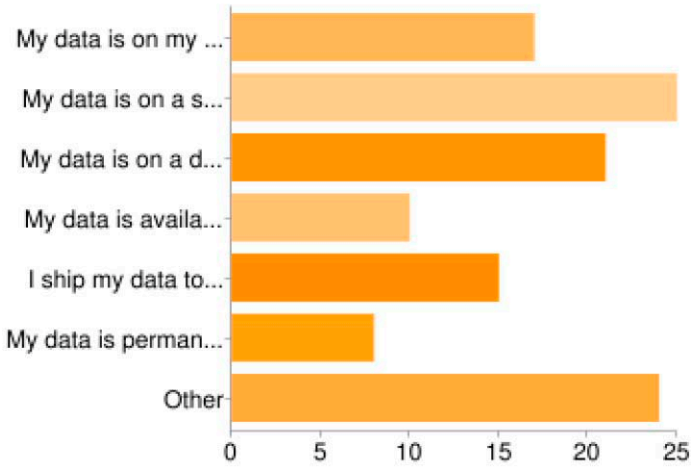
We are interested in understanding how you share your data to others.

People may select more than one checkbox, so percentages may add up to more than 100%.

My data is on my desktop/laptop and others log in to access it. (17) 22%

My data is on a server in my office or lab that has secure access mechanisms (e.g., FTP or SCP). (25) 33%

My data is on a department server that has secure access mechanisms. (21) 28%



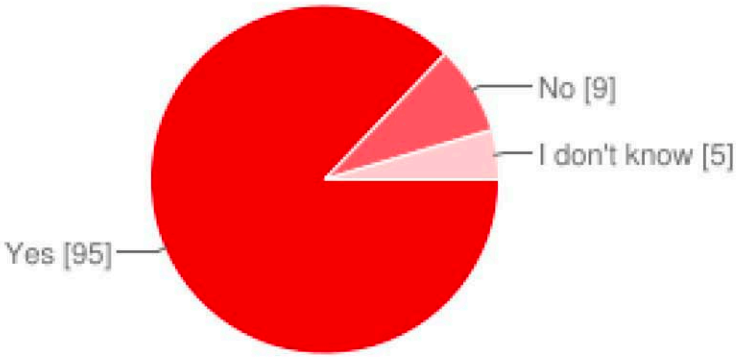
My data is available on my department or U-M website (with or without secure access). (10) 13%

I ship my data to another non-UM location where it is shared to others. (15) 20%

My data is permanently archived at the source and they share it with others. (8) 11%

Other (24) 32%

catastrophic loss?



Do you back-up your research data for your grant funded projects such that you could meet your grant commitments if your primary storage facility suffered a catastrophic loss?

How do you protect your grant funded research data during the funding period?

People may select more than one checkbox, so percentages may add up to more than 100%.

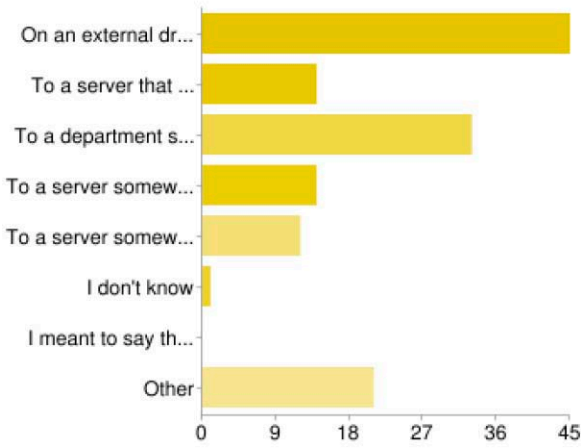
On an external drive connected to my desktop/laptop. (45) 47%

To a server that sits in my office. (14) 15%

To a department server. (33) 34%

To a server somewhere outside my building. (14) 15%

To a server somewhere off the U-M campus. (12) 13%



I don't know. (1) 1%

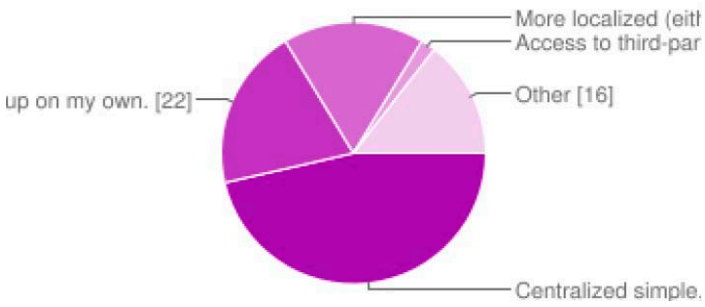
I meant to say that I don't back-up my data. (0) 0%

Other. (21) 22%

Rarely (every 6 months at best).		7	6%
Monthly		26	24%
Weekly		23	21%
Daily		23	21%
Multiple times per day		7	6%
Other		24	22%

How often do you back-up your data?

If you were offered a choice, which would you prefer?



Centralized simple, quick, and efficient automated back-up services for my research data. 51 46%

Centralized high volume and high accessibility storage space, where I can store my research data and back-up on my own. 22 20%

More localized (either in the department, office, or lab) storage space where I can store my research data. 19 17%

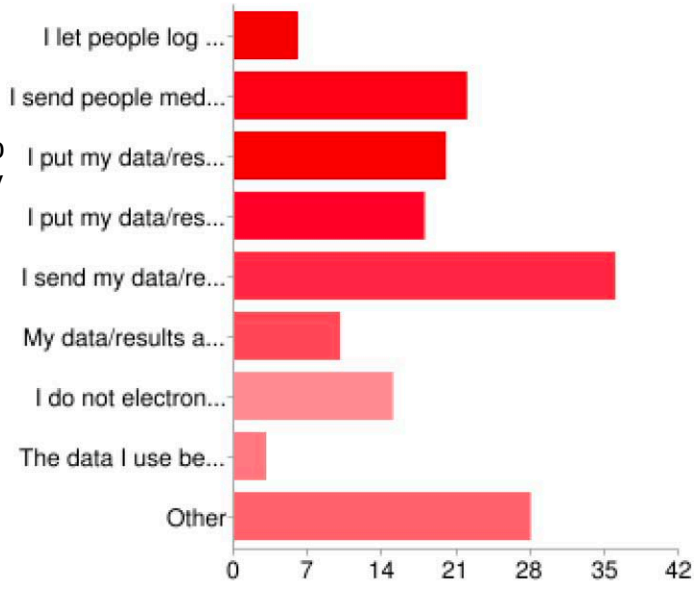
Access to third-party backup and storage solutions. 2 2%

Other 16 14%

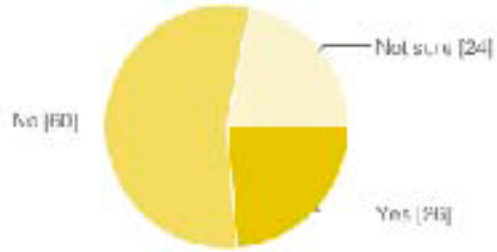
How do you electronically share your data and the results (e.g., processed data) after the project funding has ended?

I let people log into my desktop/laptop.	☐	6	5%
I send people media via the mail (i.e., DVDs, tapes, hard drives).	☐	22	20%
I put my data/results on a server in my office.	☐	20	18%
I put my data/results on a department server.	☐	18	16%
I send my data/results to an archive or journal and they share it.	☐	36	33%
My data/results are permanently archived at the source and they share it with others.	☐	10	9%
I do not electronically share my data/results after the project has ended.	☐	15	14%

People may select more than one checkbox, so percentages may add up to more than 100%.



Do the journals you publish your research in require that you make your data publicly available?



Yes		26	24%
No		60	55%
Not sure		24	22%

I let people log into my desktop/laptop

I send people media via the mail (i.e., DVDs, tapes,

Appendix F – Reference reports

The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Publishing, Jim Gray, 2009, http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Developing the UK's e-infrastructure for science and innovation, Report of the ODI e-Infrastructure Working Group, written and edited by Philip Pothem (JISC), www.nesc.ac.uk/document/OSI/index.html.

Keeping Research Data Safe Factsheet, prepared by Charles Beagrie Ltd and JISC http://www.beagrie.com/KRDS_Factsheet_0910.pdf

Data Management Practices in Social Sciences, UK Data Archive, August 2010; Veerie Van den Eynden, Libby Bishop, Laurence Horton and Louise Corti; http://www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf.

COMMUNICATION FROM THE COMMISSION TO THE COUNCIL, THE EUROPEAN PARLIAMENT, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS:
Europe's cultural heritage at the click of a mouse
Progress on the digitisation and online accessibility of cultural material and digital preservation across the EU, 2008.

Our Cultural Commonwealth: The report of the American Council of Learned Scoieties Commission on Cyberinfrastructure for the Humanities and Social Sciences, 2006, http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

Principles and Good Practice for Preserving Data International Consortium for Political and Social Research, IHSN Working Paper No 003, Dec. 2009.

STEPS TOWARD LARGE-SCALE DATA INTEGRATION IN THE SCIENCES, National Research Council, 2010.

Policy-making for Research Data in Repositories: A Guide, May 2009, Version 1.2; Ann Green, Stuart Macdonald, and Robin Rice, http://www.disc-uk.org/docs/guide.pdf

Institutional Data Management in Higher Education; Educause Center for Applied Reseach, 2009 Study of Data Management in Higher Education**;** Ronald Yanosky; http://www.educause.edu/Resources/InstitutionalDataManagementinH/191754.

Cyberinfrastructure Advisory Framework, A Net@EDU Campus Cyberinfrastructure Working Group White Paper, Educause, October 2008, Boyd Knosp; http://www.educause.edu/Resources/CyberinfrastructureAdvisoryFra/176709

University of Michigan Health System Data Warehouse Needs Assessment and Strategy, July 2010, Recombinant.

“The Continuum of Basic and Clinical Research Informatics and Health IT: Current and Future Perspectives"; Presentation to UMHS ITEC Committee, November 11, 2009, Kenneth J. Pienta, M.D and Brian D. Athey, Ph.D.

CTSA Biobank Consortium Update, The University of Texas Center for Clinical and Translational Sciences; Lorraine Frazier (PhD, Rm, FAHA, FAAN)

Research Data Task Force Report: Yale University Research Data Interviews, April 20, 2010.

Appendix G – Interview Notes

The panel interviewed some leading figures in research data management, including Sayeed Choudhury, Reagan Moore, and Lincoln Stein.

Sayeed Choudhury Conference call –

Date: 7/13/10
Attendees: Sayeed Choudhury, Margaret Hedstrom, Paul Killey, Sean McKee, Cindy Wells, John Wilkin

Background
Sayeed Choudhury, associate dean of university libraries and director of the Sheridan Libraries' digital programs, has been named to the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, funded by the National Science Foundation and the Andrew W. Mellon Foundation. Also partnering on the project are the Library of Congress, the Joint Information Systems Committee of the United Kingdom, the Council on Library and Information Resources, and the National Archives and Records. The 15-member panel includes library directors and specialists, scholars and technologists in the United States and the United Kingdom. It will convene quarterly discussions involving international experts from the academic, public and private sectors and, in late 2009, will issue a report offering recommendations for digital preservation. In addition, Choudhury has been appointed lecturer in the Whiting School of Engineering's Computer Science Department. He will lecture on digital preservation in a new undergraduate course being held this spring.

Hopkins is a big and decentralized place. There was a task force about 3 years ago that looked at eResearch. That group did not move toward a strategic roadmap, doing more grassroots. Work began with relationships between faculty and library. eScience gave Sayeed more insight into the data needs through a dozen faculty members. JHU did not want to move too quickly because they didn't want to have to turn people away.

Advice
When starting the effort, try to break the problems into smaller parts.

Pick clusters – fundamental layers – storage systems, networking layers
Above that is preservation layer – association of bits with time and
Duration layer – assuming stored, backed up, and preserved. – trans-disciplinary view and access
Connecting to application – fundamental assumption is that they will need to connect
Interfaces – discovery – thin layer for viewing into the data.

Pick areas – Astronomers, geologists (decades worth of data) –trying to pick common set of standards. Suspect that most of the commonality will be at lower levels. Trying to be very formal in their analysis. Not tractable to build unique solutions for each domain. Working from use cases – current uses – unmet needs

– extract functional needs and then identify solutions. Currently in analysis phase and building out prototype systems for Astronomy with geology looking at it.

Think about what can be leveraged

Where do you draw the line between the requirements that a campus should satisfy and where those demands cut across lots of different institutions and should be addressed in a different way. Common requirements implicitly assume that campus would want to solve. It would make sense to not stop at the campus boundary. There are lots of different places where you think that many institutions will want to do. Where are Michigan strengths.
Margaret has an IGERT award for graduate training in about what it generic across disciplines and what is unique within a research domain.

The human side of this is extremely important.

JHU is working on the data managers. Can you get access to prominent faculty member to engage them? It has been very helpful for JHU. There is nothing like faculty members talking to each other.

Think about how to get the data out

How do you make a storage service available to researchers – get data in and out? JHU is talking a lot about data systems. When thinking about preservation, can have good preservation policy, but how do you query storage system in a meaningful way. What they thought would be very basic things to do are turning out to be more difficult than they thought. Motivation

Look for storage partners

If someone else would take care of the data storage, he would be very happy. It is not something that they would like to spend their time on, but they have not been able to move it outside of their organization. Duraspace is one of their partners, they have been doing connectors into cloud providers. They are doing a lot of the work for them. Preservation policy, maybe one of the copies could be outside of their system. Concerns about cloud storage. Do the data characteristics make it not possible to move to the cloud. Conversations about network capacity, latency, bandwidth, security and transparency. Some people are not very comfortable with not knowing where the data physically resides.

Data Policy

On policy side, JHU is not addressing question of selection. First project is 140 TB for SLOAN astronomy sky project. Started conversation about whether all data are needed. Very difficult conversation to have. Initial reaction is of course it is all important. Sayeed runs out of ability to ask deeper questions because it is tied to the knowledge of the science. Looking as opportunity to better understand data. Everyone is struggling with the same issue. Margaret believes this is one of the areas where domain and context really matter. Can we assume based on use that data that has not been used is not important? Looking to be able to move between different tiers of storage based on usage.

Meta data is hard to do well

Meta data is huge area of lack of awareness and the sheer act of verifying with scientist. Very resource intensive – role traditionally done by libraries. Struggle with trying to balance faculty lack of concern about electronic and the ballooning amount of digital content. JHU library has purposely tried to discourage demand.

Additional items to consider

Consider data load time for large amounts of data, real time transactional (printing services), and need for censors.

Look for connections to applications, i.e. Demonstrating interconnections to SAKAI.

Reagan Moore Conference call

Date: 7/5/2010
Attendees: Reagan Moore, Alex Kanous, Frank Manion, Paul Millis, Margaret Hedstrom, Jeff Castle

Background

Reagan Moore, Chief Scientist, Data Intensive Cyber Environments (DICE)

Reagan W. Moore is RENCI's chief scientist for Data Intensive Cyber Environments (DICE), coordinating collaborative research efforts between RENCI and the DICE group in the UNC Chapel Hill School of Information and Library Science (SILS) in areas such as the development of data grids, digital libraries, and preservation environments. Moore is also a professor with SILS. Prior to his RENCI and UNC appointments, he was the director of data and knowledge systems at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego. Moore came to UNC Chapel Hill in August 2008 along with other members of the DICE research team. The team is known worldwide for its expertise in the development of digital data technologies, including open source software that enables sharing of data in collaborative research, publication of data in digital libraries, and preservation of data in persistent archives for use by future generations. Moore holds a Ph.D. in plasma physics from the University of California, San Diego

iRods overview

Storage resource broker, manages the properties of the data

Grid, backup tapes, other institutions

Has audit capabilities

iRods is the next generation

The genesis of iRods was to help research organizations that consist of independent research groups manage the various policy and security requirements of each group from a single point across many storage solutions. From a single management point, iRods can be used to optimize policy and security (through automated scripts) for collaboration while ensuring compliancy of agency requirements.

Many functions exist within iRods to optimize collaboration. These include, but are not limited to:

- Support of many groups with independent security and policy requirements
- Policies consist of automated procedures to:
    - Move files to backup (disk or tape)
    - Date based file expunge rules
    - Audit
    - Check-in/Check-out
- Platform independent storage solution support
- Supports cloud based storage
- Auto manipulation of data
- Can be tuned to support specific applications
- In the system, a name space is provided, challenge response, Kerberos, shibboleth
- IRB policies can be incorporated in the workflow process
- IRods can maintain an independent permission database at the project, distribution, and individual level

The system security is based on an association between ID to the file. Since user name space is managed independently of the local name space. Id authentication can be auto-managed. Australia setup a certificate system where a valid certificate creates a new user account.

Use

The iRods solution is in place for a wide variety of uses by communities wanting policies on their data or they have a need that can no longer support a single resource for storage. In production in Taiwan, Australia, France National Repository, North Carolina, used in Texas Digital Library, NSF for IRB approval, is on terra grid, running at Texas Advanced Computing Center, various research productions.

Implementation

The iRods architecture requires that each site to have a rules engine. Policies must pass through the rules engine for execution.

An approach for iRods implementation is to work with a variety of communities to gather recommendations. The communities should represent many different needs for policy and security in order to prepare for a wide variety of potential requirements for the implementation.

The iRods system is open source. It only takes a half-hour to install on a Macintosh system, which includes a postgre database and ODBC interface. An iRods implementation needs a database administrator on staff to manage the database and staff to manage the network, remote storage, security, data collection, meta-data, and ingestion policies. In Dr. Moore's experience, one person can manage 20 independent data grids. IRods is capable of digesting as many as 5 million files of data, however the speed is limited to network, server, and database latencies. Cinegrid is moving 360 mb per second over a 10 gb per second, some can push 10gb per second on a 10 gb per second. Need to overcome disk contention; usually dealt with by leveraging a solid state hard drive.
It may be possible to use a commercial interface, but this remains to be seen and depends on if the commercial interfaces have a proprietary interface. Some collaborative work with commercial providers is occurring.

Dr. Reagan's experience with iRods is the solutions he has been involved with supported a multitude of funding sources (federal, higher education, state, international).

Lincoln Stein Conference call

Date: 7/21/2010
Attendees: Lincoln Stein, H. V. Jagadish, Alex Kanous, Frank Manion, Paul Millis, Margaret Hedstrom, Cindy Wells

Background
Lincoln Stein as well known for developing informatics infrastructure for bioinformatics. He was Eric Lander's informatics lead on the human genome project at the Whitehead, and ran a program at CSHL for some years, had a lot to do with dbSNP if memory serves right, and now apparently is up in Toronto at the Ontario Institute for Cancer Research.

Notes
The technology required for cancer research has become increasingly more complex over the past several years. Accordingly, the development of innovative solutions to handle the large and complicated datasets these new technologies generate has become an essential part of the cancer research process. OICR's Informatics and Bio-computing Platform, led by Dr. Lincoln Stein, integrates and interprets this data, developing visualization tools to turn raw data into meaningful information for biologists, clinicians and trainees.
Genomic dataset at a small institution

Key need is to support the needs of the research well – biosciences research needs.

1. Special purpose cluster for express purpose of addressing needs of genome sequencing group. Optimized to get the high level of data from the sequencing to the computing. 2.5 petabyte on spinning disk. Off load to tape and send out for storage. High speed disk – scientific designs – share exact specs of the system if interested

2. Running community databases for significant databases, including plant genome. Go into detail of how capture metadata.
    a. Modicode – NIH sponsored national human genome to identify module. Involves 11 labs in North America. One set of group are using transcriptional techniques to identify variance and annotate them. The other half are using chromosome based techniques. Other groups are looking at binding of other.
    b. Give very detailed and complex and dynamic view.
    c. Challenging because of large data set and the meta data is very complex. Every experiment needs to have detailed set of metadata so that data can be compared. Including version of software being used.
        i. Used combination of structured and unstructured data to manage
        ii. Data – highly structured meta data – controlled vocabularies – systematically check files
        iii. Experimental protocols and other less structured data – using wikis with templates built on top of controlled vocabulary and autofill. Before uploading experiment, the research team has to fill out template. Each wiki page has URL that uniquely identifies the experiment that are controlled fields in file upload. From that can reconstruct the science. Each file has metadata file uploaded with it.
        iv. To distribute data to world. Genome browser. In addition have queriable database that can be used to answer more specific questions. Public available database.
    d. Challenges of project
        i. Quite difficult for members of labs to provide the level of detail in the structured and semi-structured formats. Unable to do it because don't have sufficient informatics locally. In those cases they do it for the investigator by going to the core data and having the curators do it. There are five curators.
        ii. LIMS systems are designed for experimental tracking. Some labs do not have a limb system. Some are using Excel. Each lab has custom designed systems that are designed specifically for their project and are not portable. LIMS system in general are not portable from one group to the other. If LIMS systems have metadata, they don't want to do this by hand, so curators write scripts to convert.
        iii. Always an issue in community projects curation is an issue. In every meeting, the participants ask for more metadata. It is a painful process, but necessary to make the results lasting.

iv. Project wide QC standard – implement and almost all of the data gets kicked out of system because don't meet QC.
v. Highly manual and the outcome is highly useful piece of data infrastructure. Contrast worm c-elegance database which is a single highly controlled curation compared to human sequencing which have wide variety of metadata, definitions of terms and curation techniques. This makes the data less useful for human sequencing. Political rivalries make this very difficult to resolve.

3. Pilot project with Illinois to put into private cloud. As the data gets too large, one solution is to move it to the cloud and have people pay per drink to process specific parts of the data.
    a. Collected large datasets. Initial way of distribution of files was to put them on the web and then people would download these very large files from the web.
    b. Very inefficient because of multiple copies of the data and very difficult for version control.
    c. As the project moved into the publication, this became more of a problem. Bio-nimbus private cloud which consists of eucalyptus and Hadoop on high speed file system developed at Illinois. Mirroring all primary and secondary data sets from primary location in Toronto to Illinois so there is a complete copy of the data.
    d. Each member of the consortium has an account and the ability to create virtual machine. Provide templates of virtual machines with useful software. Has been a great success.
    e. Most of the groups have embraced this. Access to compete body of data.
    f. Hoping to move entire set into Amazon and let the public do that same thing.
    g. Long term plan is move to Open Cloud Consortium. In process of Upgrading link to 10G to the consortium.
    h. Allow people to have fast access to data sets in private compute environment. Cost recovery mechanism. Pradeep is Amazon contact and will put up content for free and hoping to make money by charging for virtual machines. Problem today is the I/O speed is very low today.
    i. Amazon will need to alter their infrastructure. Only one mount of data is allowed Amazon Cloud Service today. One way is to copy. Mount on virtual machine that exports as NSF. This has been experimented with and after 6 connects performance collapses.
    j. The Illinois data storage solution is much better, Sector distributed file system.

4. Just started exploring creating a Toronto based private cloud for the medical community because it didn't make sense to duplicate the effort across the university, hospital and other research. Agreed in principal with the basic concept of charging per virtual machine. Brian Athey is looking to create similar system.

5. Next Generation sequencing going – latest generation of sequencing machines no longer make the sequencing available to you. Does the imaging on it's own little cluster. The images will be thrown away because they are so huge. Keep the base files. For the immediate future don't see any solution to the data storage problem. The current generation of algorithms is too buggy to just rely on interpretations and throw out the raw data.
    a. Problem with finishing any analysis because they can get new data if they do the analysis a couple of months later.