Findings from Data Assessments with Faculty in the Sciences and Engineering
March 15, 2011

1. Overview

In May 2010, Steve Girvin, Deputy Provost for Science and Technology, invited nine faculty members in the sciences and engineering to participate in an assessment of their data storage and management needs. The assessments were intended to investigate data-related practices in specific disciplines that with the Research Data Task Force Report[1] and Mass Storage Working Group Survey[2] could be used to help generate cross-disciplinary requirements, inform use cases, and streamline services. The ultimate goal is to develop recommendations for the Scientific Computing Strategic Planning Advisory Committee on an institutionally supported, centrally managed research infrastructure that supports the creation, use, management, and stewardship of digital content in the sciences, engineering, and social sciences at Yale.

Staff from the Office of Digital Assets and Infrastructure (ODAI) conducted the assessments with faculty in the sciences and engineering over the summer of 2010. An interview was conducted with each faculty member, and at the faculty member's discretion, with graduate students, postdoctoral researchers, and support staff. The interviews were recorded and summarized.[3] The faculty member reviewed, edited, and approved the final version of the assessment.[4]

This document (current as of September 2010) has been prepared to report findings from the data assessments with faculty in the sciences and engineering to the Mass Storage Working Group. Additional assessments are being conducted with faculty in the social sciences and humanities over the course of the 2010-2011 academic year. The complete findings will be reported to the Office of the Provost by June 30, 2011.

This document is organized as follows. A summary of findings (Section 2) and participant profile (Section 3) are followed by the detailed findings (Section 4), which are organized under four major headings: Storage and Backup (Section 4.1); Computing (Section 4.2); Collaboration (Section 4.3); and Data Management and Sharing (Section 4.4). Sample interview questions are provided in Appendix A. Researcher profiles are provided in Appendix B.

2. Summary Findings

- **Storage and Backup**: The storage needs of the research groups interviewed range from a terabyte to multiple terabytes per group. Some groups are using nearly all of their available storage for active research. Most groups are retaining primary (and ancillary) data from past projects indefinitely and discarding intermediate stages of data once processing is complete. The groups generally require high performance storage for data in active research. They consider low performance, offline storage acceptable for archived data. One investigator with particularly large data sets expressed concern about the frequency of backups and whether or not backups could be kept up-to-date. Investigators expressed a willingness to wait more than a day for backed up data to be retrieved and up to a week to restore an entire system.

- **Computing**: The extent of computing varies across research groups. Three groups use centrally supported high performance computing (HPC) facilities; five groups use group servers; and one investigator is directing a center with its own infrastructure. While researchers commented favorably on access to the Bulldog clusters, they expressed frustration with the Lustre file system and with staffing levels.

- **Collaboration**: Research groups, ranging from a few to dozens of members, expressed a need to work collaboratively on large data sets. Typically, collaboration is project-based, with multiple group members involved in each project. Some of the collaboration (e.g., code development) happens in real time. Some researchers are using tools for source

[1] Office of Digital Assets and Infrastructure. *Research Data Task Force Report*. March 15, 2010. Revised April 20, 2010. http://www.odai.yale.edu/sites/default/files/file/research data report final v2.pdf

[2] Office of Digital Assets and Infrastructure. Mass Storage Working Group. http://www.odai.yale.edu/mass-storage-working-group

[3] To protect confidentiality, the audio recordings of the interviews will be deleted by June 30, 2011. ODAI will retain in confidence the interview transcripts and assessments.

[4] The assessments provide in-depth information about particular labs. However, the findings from the assessments are not necessarily representative of all faculty and should be evaluated with information gathered using complementary methodologies (e.g., surveys).

control, version tracking, and device synchronization. Some collaboration involves non-Yale researchers. In some cases, the scale of data has limited the extent of collaboration.

- **Data Management and Sharing:** Data management is not well integrated into active research. Graduate students, who typically have responsibility for managing data, want to move on with their research. In a few cases, researchers who submit data to repositories or with publications gradually adopted data management practices in active research to anticipate data submission requirements. One investigator pointed out that practices for sharing primary data are not well understood or established in his field.

3. Profile of Science and Engineering Participants

Participants were drawn from science and engineering departments in the Faculty of Arts and Sciences and the Yale School of Medicine, particularly those known to be generating large amounts of data in their research. In order to help identify common needs, participants were asked to describe their research, focusing on data from a recent project. Researcher profiles are provided in Appendix B.

The research projects and research areas were categorized according to the methodology used in research: experimental, computational (modeling or simulations), observational, qualitative, theoretical, or applied (where application or tool development is a focus). A similar typology used for the Mass Storage Working Group Survey had suggested that Yale researchers using experimental and computational methods generate more data than researchers using other methods. The assessments with faculty in science and engineering suggest that **advances in scientific instrumentation** – such as the use of synchrotron radiation in x-ray crystallography[5] (Dr. Pyle), the development of next-generation sequencing techniques[6] (Drs. Handelsman, Krauthammer, and Mane), and improvements in CCD technology used in telescopes[7] (Dr. van Dokkum), to name a few examples – are closely tied to the data deluge.[8] More evidence is needed to determine whether or not advances in scientific instrumentation have particularly affected researchers using experimental and computational methods.

4. Detailed Findings

4.1 Storage and Backup

The faculty members who participated in the assessments are cumulatively retaining and storing data at accelerated rates but in many cases lack the resources to ensure the viability of their data over time. Appendix B provides information on **current storage requirements** for the projects that served as the focus of the assessment and for the total storage required per group for all active projects. In cases where project storage exceeds total available storage, investigators are dividing up their data for analysis and discarding intermediate stages of data. The need for high performance storage can be estimated to range from the storage available on a desktop computer (Dr. Handelsman) to 15 TB (Dr. van Dokkum, who must divide up his data for analysis).

One common practice across research areas and methodologies is to retain primary data sets. Typically, primary data sets account for the bulk of storage and are significantly reduced to produce the final data sets. This reduction may be as steep as four orders of magnitude. One of the most commonly cited reasons for retaining primary data is to be able to reproduce results or to apply new analyses. According to Dr. Pyle:

---

[5] Hendrickson WA. (2000). Synchrotron crystallography. *Trends in Biochemical Sciences* 25(12): 637-643. http://www.ncbi.nlm.nih.gov/pubmed/11116192

[6] Baker M. (2010). Next-generation sequencing: adjusting to data overload. *Nature Methods* 7: 495-499. http://www.nature.com/nmeth/journal/v7/n7/full/nmeth0710-495.html

[7] Goodman AA, Wong CG. (2009). Bringing the Night Sky Closer: Discoveries in the Data Deluge. In T Hey, S Tansley, and K Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. 39-44). Redmond, WA: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part1_goodman_wong.pdf

[8] Searching *Nature* (http://www.nature.com/) for "data deluge" returns hits from as early as 2000 – in this case, an article on the use of Hidden Markov Models in genomics: Wickware P. (2000). Borrowing methods and models. *Nature* 404(687). http://www.nature.com/nature/journal/v404/n6778/full/404687a0.html
The term is commonly used to describe the vast quantities of data being generated in e-science.

Many assumptions are made during data processing that can influence the quality and the results in the final structure. So, protecting that raw data is just vital.

Other faculty members reported that they retain primary data in order to be able to share it, as a service to their field, or as a matter of convenience, given the time and effort required to generate or obtain it.

Intermediate data generated in calibrating or analyzing primary data is typically discarded. Temporary storage for intermediate data may be significant, depending on the type of file. A storage-intensive scenario is Dr. van Dokkum's deep survey project, which produced 25 TB of images as intermediate data, which were discarded after they had been used to create the final data products. A stage of the intermediate data – calibrated images that had been corrected for artifacts but not co-added – was later recreated at the request of a group at UC Berkeley, which suggests a trade-off involving the availability of storage and student labor. In all of the other assessments, the storage needed for intermediate stages of data was less than the storage needed for primary data.

Given that primary data is commonly retained and typically accounts for the bulk of storage requirements, **storage for archived data is generally cumulative storage for primary data**. With one exception involving research using computational methods, participants felt their primary data should be retained for at least five years. Some felt it should be retained indefinitely. In fields where advances in scientific instrumentation are rapidly increasing the amount of data generated in research, retention practices for primary data may change over time. For example, the Yale Center for Genome Analysis (YCGA) has been retaining all data excluding images, typically 1.3 TB for the most common type of run, but has recently proposed to retain only the final sequences, typically 200 GB per run, for a period of 8 months to 1 year, after which responsibility for data storage and management would be distributed to individual researchers, while raw and intermediate data would be discarded after the samples had been processed. The Broad Institute, Washington University, and Cold Spring Harbor Laboratory, institutions at the forefront of genomics research, are about to implement similar policies. Given limits on additional storage purchases, the new data retention policy is necessary to extend the capacity of YCGA's 1 PB Panasas storage system over its 3-year lifespan.

4.2 Computing

Computational approaches play an increasingly vital role in new and emerging lines of research, but funding and support for computational research varies from lab to lab, resulting in enormous diversity in resource allocation and provision for associated storage and data management processes. Three groups (Drs. Korenaga, O'Hern, and Pyle) use centrally supported HPC facilities. Five groups use group servers administered by ITS (Drs. Dorsey and Handelsman) or within their own groups (Drs. Brandt, Krauthammer, and van Dokkum). The ninth participant (Dr. Mane) directs the Yale Center for Genome Analysis, which has its own IT director and its own relationship to the HPC facilities and data center.

Several researchers commented favorably on the centrally supported HPC facilities, with respect to access and availability. Complaints included the frequency of problems with the Lustre file system, the relatively small size of home directories, a queuing system that underutilizes available cores, and staffing levels. One researcher commented on differences among classes of user – those who are writing their own code and can modify it to run in parallel or to insert breakpoints, and those who are using commercial software and cannot change how their jobs run – suggesting a need for different service models.

Those who do not use the centrally supported HPC facilities cited a range of reasons for not doing so: the software in their field is primarily written for Windows whereas the centrally supported clusters run on Linux; they would like to have the freedom to install software and utilities as needed; or their computing supports user interfaces.

4.3 Collaboration

Scientific research is highly collaborative, but the institutional and community infrastructure for e-science is too fragmented to adequately support collaborative work on the terabyte scale. Publicly funded community infrastructure supporting specific disciplines is not well integrated with institutional infrastructure supporting individual researchers.

The faculty members interviewed were supervising groups ranging in size from a few to dozens of members. (See Appendix B.) All of the faculty members had been or were currently involved in collaborations with researchers at other universities or in industry.

The faculty members reported using departmental servers or group servers or storage to facilitate collaboration within groups. Groups that maintain their own servers cited a number of reasons for doing so: centrally supported storage options are not cost-effective; their data supports user interfaces, so they need more administrative control; or they would like to have the freedom to install software and utilities or to develop applications as needed.

Faculty members who collaborate with outside researchers on large data sets described different strategies for collaboration. Dr. Korenaga and his collaborator mirror their data. Dr. van Dokkum and his collaborators selected a location to reduce the data (which turned out to be Yale) and shared reduced data sets. Dr. Pyle reported shipping data on terabyte drives to other specific researchers on request. Drs. Dorsey and Handelsman obtained Yale net IDs for external collaborators to give them access to group servers administered by ITS.

Participants offered a number of suggestions for facilitating collaboration. Based upon his experience at a national lab, Rob Hoy, a postdoctoral researcher in the O'Hern group, suggested having project directories systematically created for research groups:

> *There are others in the O'Hern group who work very closely together on the same data set. […] One thing I did at Sandia and I think would be a very good idea here is to have a projects directory that can be accessed by everyone in a group.*

Some collaboration (e.g., code development) happens in real time. Some of the interviewees reported that they use commercially available services for backup and device synchronization to facilitate code development, as well as other research activities. Dr. Krauthammer suggested that commercially available services might be difficult to match:

> *SugarSync is a cloud-based solution. I can sync anything from any computer, including from my iPad, my iPhone, or any other device. I think it will be hard for ITS to match these offerings.*

Faculty members involved in collaborations with research centers in Europe cited several as models. Dr. Dorsey commented favorably on her collaborations with researchers at INRIA:[9]

> *They have a server there. It was easy for all of us to get access and it seems like it's maintained very well, backed up. There are frequent messages about different kinds of things being done. Then they have code maintenance systems all built in. That might be something to look at as a model.*

Dr. van Dokkum described publicly funded data centers in Europe as an alternative to the more distributed research infrastructure in the U.S.:

> *There are several places in Europe where data coming from big telescopes are centrally analyzed. They have huge storage systems, huge computers, and a lot of people who are paid to analyze the data. For us, a student does it. There, it's actually people who are paid to do that type of thing. Sometimes those data centers will actually lead the field because they will develop the tools that others then use. They also make software available. Some software we use was developed in France at one of these data centers. Usually they're very smart people who do this 100% of the time. In the U.S., something like this does not exist.*

4.4 Data Management and Sharing

Data management is handled very much on a lab-by-lab basis and is not always well integrated into the process of research, increasing the risk of data loss and corruption and potentially limiting data sharing and reuse. Graduate students typically have responsibility for managing data in active research but, as suggested by Dr. Dorsey in her comments below, don't have much time to devote to it outside of their research:

[9] Institut national de recherche en informatique et automatique. http://www.inria.fr/

Right now [data management] is an afterthought. […] It just takes a lot of time to manage it and in some cases expertise that we don't have readily available.

Dr. Krauthammer also suggested that, given the pace of projects, data management is a challenge:

We have some Google docs to keep track of the raw data, but beyond that everyone is working on his own project, and keeps data within his own folder structure. We use email to communicate the location of data files that need to be shared. People lose track of – it seems to me – of intermediate files. At one point, you have to start deleting files and you might not remember the context in which a file was generated. It's an issue of providing annotations, such as README files. We don't have a clear policy.

Dr. Dorsey observed that additional support for data management would be needed to make data shareable:

Right now it takes additional time beyond just doing the research to make it available. Generally the students want to move on. It's only out of some sense of wanting to see the stuff used by other people, wanting to share it with other people, or people asking for it, that it gets done.

Disciplines in which data sharing is an established practice tend to have more standardized data management practices. The emergence of data repositories[10] has standardized data management in many scientific disciplines. Data repositories typically impose submission requirements, develop metadata and data annotation guidelines, and provide tools to help depositors get their data into a standard format. An example of the influence of repositories was provided in one of the assessments. In order to make it easier to deposit data as required by his funders, Dr. van Dokkum modified his code for analysis to generate more header metadata. This made it easier to create the catalog for his final data products.

Some academic societies and journal publishers accept supplementary materials,[11] including data, with journal submissions, or they require authors to provide an accession number for any data cited in a paper.[12] These practices have tended to encourage labs to adopt standard practices for preparing and locally archiving data supporting publications. For example, the Pyle Lab has a standard format for archiving papers on the group's server, which they use to track all the supporting data and code for figures that appear in the paper.

Data management involves not only tracking data but also tracking samples. Investigators who are making samples available to other researchers have invested in standardizing sample management, from developing homegrown databases linked to freezer maps (Dr. Handelsman) and web interfaces for other researchers to order samples (Dr. Krauthammer) to licensing a laboratory information management system (Dr. Mane) to enable YCGA staff and users to track samples through runs and analysis pipelines in real time.

As described above, infrastructure for data sharing (e.g., the practice of submitting data to repositories and with publications) facilitates the adoption of standardized data management practices. Improved data management also makes data shareable, which enables reuse. The lifecycle approach to data management and use is exemplified by researchers in the Pyle Lab who not only are submitting structures to the Protein Data Bank (PDB) but also are downloading data sets from the PDB to develop computational approaches to solving structures. In fields where infrastructure for data sharing is not firmly established, investigators recognize the potential for reuse, as alluded to in comments by Dr. O'Hern:

I think at this moment the scientific public especially is most interested in macroscopic quantities that are derived from the data that I generate. There are 'microscopic features', e.g. locations and orientations of the particles that may prove to be important, especially since structural and mechanical properties depend sensitively on preparation history. The scientific community is more interested in bulk properties and therefore much of the microscopic data is 'thrown away'. Studies of bulk properties are described in the publications and presentations

[10] See NCBI (http://www.ncbi.nlm.nih.gov/guide/all/#Submissions_) for gene sequencing and PDB (http://deposit.rcsb.org/) for structural biology. See NOAO (http://nvo.noao.edu/) for astronomy.
[11] See PNAS (http://www.pnas.org/content/95/21/12073.full) for the physical and biological sciences and ACM (http://www.acm.org/publications/policies/dlinclusions/) for computer science.
[12] See GenBank (http://www.ncbi.nlm.nih.gov/genbank/submit.html) for gene sequencing.

> *that I give. And that is amply featured on the O'Hern web site, and stories and vignettes about the data are shown in the form of movies and pictures. Part of that may be driven by the fact that* ***there are not good ways to organize, administer, and store all of the 'microscopic' data. If we had better ways of organizing, creating databases, and understanding all of the microscopic features of the data, then I think there would be more use for it.*** *But currently, I'm judged, peer reviewed, on macroscopic descriptions of the data.*

However, even in domains where data repositories exist, these repositories typically accept final data products, so it remains the responsibility of the investigator to manage and archive primary data.[13] Few projects specifically seek funding for data management, although this practice may change with the NSF Data Management Plan requirement.[14]

[13] Some exceptions are seismology and astronomy, fields in which raw data is archived and made publicly accessible through federally sponsored or globally supported initiatives.
[14] National Science Foundation. Dissemination and Sharing of Research Results. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Appendix A: Sample Interview Questions

1. Please describe your research areas.
2. Which research methodologies do you use: experimental research, models or simulations, theoretical research, observational studies, or qualitative research?
3. How many graduate students and postdoctoral researchers are you currently supervising?
4. In order to make this assessment concrete and specific, in the next few questions, we'd like to focus on your project [citation to recent paper provided by participant before interview]. What were the major findings from the project?
5. What kinds of data (content, purpose, file type) were generated at each stage of the project? How much? Where are you storing it? How long do you intend to keep it?
6. What kinds of software do you use in your research? Commercial or open source?
7. Do you write your own code? Do you use any tools for source control?
8. Do you use the high performance computing facilities at Yale? Do you have any comments?
9. How do you manage your data? Do you create any metadata for your data?
10. Is any of your data obtained from human subjects? Are there any privacy concerns?
11. What are your current total storage needs? Where is your data currently stored?
12. How do you handle backups?
13. How much will your storage needs grow over the next five years? What are the drivers for growth?
14. How do you share data with other researchers? Which stages of data do you share and when do you share it?
15. How do you share data with the public? Which stages of data do you share and when do you share it?
16. Who are your funders? Do they impose data management or data sharing requirements?
17. Do you submit data or publications to any repositories in your field? Are there submission requirements? Does the existence of repositories influence how you share data and publications?
18. Who owns and has responsibility for your data?
19. Does the University provide sufficient support for technology transfer? Are the University's policies on intellectual property clear to you? Where would you go for help?
20. Which institutions have world-class infrastructure for research in your field?
21. Do you have any questions for us?

Appendix B: Researcher Profiles

Name	Rank	Affiliation	Research Project (Additional Research Areas)	Researchers Supervised	Project Storage (TB)	Total Storage (TB)
Cynthia Brandt[15]	Associate Professor	Anesthesiology; Yale Center for Medical Informatics	Clinical outcomes among cohorts of veterans (clinical information systems)	9	0.2	1
Julie Dorsey[16]	Professor	Computer Science	Material and texture models (sketch-based modeling; photorealistic simulation of synthetic scenes)	5	0.1	10
Jo Handelsman[17]	Professor	Molecular, Cellular, and Developmental Biology	Metagenomics (microbial communities and their role in infectious disease)	8	~0.01	1
Jun Korenaga[18]	Professor	Geology and Geophysics	Active source seismology (passive source seismology; fluid mechanics)	3	~30	10
Michael Krauthammer	Associate Professor	Pathology; Yale Center for Medical Informatics	Biomedical information retrieval; statistical analysis of sequencing data	6	1-2	20
Shrikant Mane[19]	Director	Yale Center for Genome Analysis	High throughput sequencing	24	1.3	1000
Corey O'Hern	Associate Professor	Mechanical Engineering; Physics	Soft matter physics (biological physics, as applied to such problems as protein folding)	9	1	15
Anna Pyle	Professor	Molecular, Cellular, and Developmental Biology	RNA structure, RNA folding, and RNA-protein interactions (computational tools for solving RNA structures)	12	1	20
Pieter van Dokkum	Professor	Astronomy	Evolution and formation of galaxies	8	30	20

[15] Dr. Brandt is an investigator for the Veterans Health Administration, which has stringent security requirements and highly centralized IT and provides storage and backup to its investigators. Hence, her total storage needs may be less than they otherwise would be.
[16] Dr. Dorsey's total storage needs include 5 TB for an image database used in multiple projects. Dr. Dorsey estimates that 10 TB would cover her group's total storage needs over the next five years.
[17] The project storage for Dr. Handelsman's assessment is estimated from her current total storage needs, which are 140 GB. She expects her storage needs will increase at a rate of a TB per year, primarily because of 454 sequencing. Hence 1 TB is given as the total storage needed by her group.
[18] Dr. Korenaga's assessment describes an exceptional data collection project. Infrastructure to store the raw data (26 TB) exists as the national level. Dr. Korenaga estimates that 10 TB would cover his group's total storage needs over the next five years.
[19] Dr. Mane's assessment describes activities at the Yale Center for Genome Analysis (YCGA) as a whole. 1.3 TB is the output, excluding images, of the most common type of sequencing run (76 base paired end). 1 PB (738 TB usable) is the capacity of YCGA's Panasas storage system. YCGA may purchase an additional 300 TB of storage.