

Research Data Task Force Report Yale University Research Data Interviews

Executive Summary

Office of Digital Assets and Infrastructure
Yale University
3/15/2010

Submitted by the Research Data Task Force:

Jonathan Lizee, co-chair, Information Technology Services (ITS)
Ann Green, co-chair, Office of Digital Assets and Infrastructure (ODAI)
Paul Gluhosky, ITS/Academic Media & Technology
Stefan Kramer, Social Science Library
Youn Noh, Library Cataloging and Metadata/ODAI
Andrew Shimp, Science Library
Judy Spak, Medical Library

Meg Bellinger, Sponsor (ODAI)

Research Data Task Force Report

Executive Summary

The core mission of Yale University is to create, disseminate, and preserve knowledge. Over the past 40 years, technology has played an increasing role in this process. The ways in which technology is used across disciplines are as diverse as the subject matter to which it is applied. And while technology has offered many benefits to scholars it has also come at a cost in terms of time spent and level of commitment required from faculty and support structures of the University to leverage technology in the academic enterprise. In order to better understand the challenges posed by the massive accumulation of research outputs and the demands of managing and sharing digital content, this report provides an overview of current practices at Yale University, drawn from interviews with 34 faculty members, with a focus on the associated benefits and risks.

The mission of the Office of Digital Assets and Infrastructure (ODAI) is to draw upon the extensive faculty and staff resources of the campus to accelerate the development of Yale's digital content infrastructure into a world-class resource that ensures Yale's digital assets will be discoverable, accessible, and usable for teaching and research both now and in the future. ODAI's mandate includes goals to define and develop foundational tools, systems and platforms based on open, interoperable architecture that will meet the digital content management needs of faculty, staff, and future learners, and to create the means to identify and protect Yale digital assets. The digital outputs of research activity at Yale are key assets that fall under ODAI's mandate. The 2009 Working Group of ODAI reinforced this mandate when it identified as a priority the need to support efforts in managing the life cycle of digital assets produced during the research and teaching activities at Yale, particularly those that are not included in the formal collections of the museums, galleries, and libraries. ODAI sponsored the Research Data Task Force (RDTF) and appointed selected staff from ITS, Library, and ODAI to develop a project to jointly **research and define faculty requirements and essential components of a coherent technical infrastructure, service definitions, and a comprehensive policy layer to support the life-cycle management of research data**. This effort is strategically aligned with other digital infrastructure initiatives in ODAI including the development of an enterprise digital asset management system and planning for digital preservation services supported by a secure storage environment.

In order to understand the current state, breadth, and diversity of technologies used in research, the RDTF conducted interviews with 34 faculty¹ drawn from a range of departments and professional schools from July to September 2009. (The interviewees have appointments in Yale Faculty of Arts and Sciences, Yale School of Nursing, Yale School of Medicine, Yale Forestry and Environmental Studies, and the Yale School of Management). The primary goal of the interviews was to determine the role and impact of technology during different stages of the research life cycle.² Of the thirty-four interviews, ten were in the social sciences, twenty-one in the sciences, and three in the humanities. The interview was designed to collect information directly from faculty about their data related processes and their perceived gaps in technology (hardware, software, and network), services, and policies that support the research enterprise throughout the data life cycle. In order to collect this information, a broad range of questions were asked to elicit descriptions of any common issues and possible common approaches

¹ One faculty member declined an in-person interview, but provided feedback via email. One person interviewed was research staff, not a faculty member.

² A campus wide approach to data stewardship throughout the data life cycle requires using consistent terminologies. Digital stewardship is the commitment made by the institution to appropriate long term retention and dissemination of digital assets, with the technical infrastructure and services required to meet that stewardship promise. Data curation, or the proper management of the data throughout its life cycle -- from creation to dissemination and archiving -- is necessary in order for the data to be accessible and useful over time. Guidance and tools are needed to help researchers manage, publish and share their data outputs, but to date those services are not organized in a coherent manner and are not widely available. Data preservation, the proper management of data for long term persistent access, should be offered to those data that are designated as requiring an institutional commitment to stewardship.

or best practices that might be shared across disciplines, and alternatively, to determine if any research data management practices were domain specific.

Five specific aspects of the life cycle of research data were addressed in the interviews (see Appendix One for the text of the questionnaire):

1. Data sharing: What mechanisms are in place to allow data sharing within the research group, within the institution, and with others outside the institution?
2. Data management: How do researchers collect and manage collections of data during the research life cycle?
3. Long term persistent access and preservation: What are the life span of research data and what policies and/or strategies are in place to access and preserve data not currently in use?
4. Data ownership: What policies and practices are in place either locally or institutionally regarding propriety of data?
5. Technical infrastructure: What is the current state of existing facilities and how do these impact faculty research?

Although efforts were made to conduct interviews with faculty researchers across disciplines, far more research would be required to capture results that are representative of the entire research universe at Yale University. However, the results of the interviews provide invaluable and valid evidence of the experience of faculty and illustrate the challenges and opportunities in regard to research outputs and provide much-needed evidence to support further actions. In all cases, faculty were eager to share their experiences, which provided a very rich picture illustrating both the diverse range of technologies at play, as well as the challenges faced by researchers attempting to adapt technology into their ongoing research.

The over-riding message taken from these interviews is clear. It is critical that the University develop strategies to address the complex issues surrounding the research data life cycle and the stewardship of research outputs. As the volume of data produced, stored, shared, and repurposed over time increases exponentially and the costs of managing, documenting, and preserving those outputs increases in parallel, the University needs to understand and evaluate the incentives, costs, and mandates behind digital stewardship. This report proposes a set of strategies to move forward in efforts to meet requirements and fill gaps identified through the interviews and other information gathering efforts at Yale and beyond. Next steps will be to vet these recommendations and move toward developing business case/s and scope statement/s with priorities. As each of these recommendations are prioritized for implementation by the Office of the Provost (and the ODAI Advisory Committee), ODAI will develop project proposals with resource requirements.

Conclusions and Strategies for Moving Forward

Yale University produces vast amounts of data during the course of research activities which represent significant investments of time, funding, and intellectual activity. The installation of High Performance Computing Clusters (HPC) is a significant resource investment for the University in data creation. Increasingly, major funding is required to develop HPC data back-up solutions and improve support for general research data storage (see the ITS Strategic Plan yale.edu/itsp). However, these major investments do not include provision for data curation processes which bring practices of selection, organization, and retention policies that will optimize these investments in hardware and software.

More importantly, the value in data resources is in the discovery and knowledge that are built from them. Data “are only as useful as researchers’ ability to locate, integrate and access them.”³ Efforts need to be taken to

³ Howe, Doug Seung Yon Rhee et al (2008) “The Future of Biocuration.” *Nature*. 455/4:47.

strategically build the research data curation infrastructure to ensure that the research is discoverable, available, and usable for further research.

The fabric of science is changing, driven by a revolution in digital technologies..... These technologies generate massive data sets that fuel progress. Technologies for high-speed, high-capacity networked connectivity have changed the nature of collaboration and have also expanded opportunities to participate in science through instant access to rich information resources around the world. While these digital technologies are the engine of this revolution, digital data are the fuel. ⁴(p. 3)

As data sets grow to the terabyte scale, virtually every academic institution is facing pressures from increasing volumes of data and increasing requirements for storage and network access. In response, some research institutions are developing the administrative and technical infrastructure to support expanding multidisciplinary research projects and to enable the integration of teaching and research.⁵ Research communities are developing repositories that incorporate data management policies, standards, and tools to enable data to be more effectively shared and reused.⁶ Nationally and globally, funding agencies and associations are developing policies and guidelines to promote data sharing and stewardship.⁷

As indicated by the findings of the interviews and other needs assessments undertaken at Yale and at other academic institutions, we know that researchers are challenged by the demands of storing and managing data across the digital life cycle, of producing data management and sharing plans that adhere to best practices, rights, and policies, of describing their data in ways that make them identifiable and usable, and of determining best formats and options for storing and sharing data securely over long periods of time. Loss of data, unmet mandates from funding agencies to preserve data, and difficulties in securely sharing data with collaborators are challenges faced across multiple academic disciplines.

With its investment in ODAI, Yale University has made an explicit commitment to building and sustaining shared digital asset infrastructure. The existing curatorial models in the libraries and museums constitute a strong foundation for building this shared infrastructure for the digital collections within these departments. However, it is the informal collections of data that require the University to build curatorial practices from the ground up. While both ITS and the University Library have some aspects of the service model needed to support the needs that are outlined in this report, additional investment will be required to develop a coherent approach to curating research data. ODAI has the resources to make a start on some of the recommendations in this report. Additional resources will be required to meet a digital stewardship commitment to appropriate long term retention and dissemination of digital assets, with the technical infrastructure and services required to meet that stewardship promise. Data curation or the proper management of the data throughout its life cycle -- from creation to dissemination and archiving -- is necessary in order for the data to be accessible and useful over time. Guidance and tools are needed to help researchers manage, publish and share their data outputs, but to date those services are not organized in a coherent manner and are not widely available. Data preservation, the proper management

⁴ Harnessing the Power of digital data for science and society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. January 2009.

<www.nitrd.gov/about/harnessing_power_web.pdf>

⁵ See, for example, the announcement of the NSF award to Johns Hopkins for the Data Conservancy

<<http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0830976>> and

<<http://releases.jhu.edu/2009/10/02/sheridan-libraries-awarded-20-million-grant/>>. The Johns Hopkins University Sheridan Libraries have been awarded \$20 million from the National Science Foundation (NSF) to build a data research infrastructure for the management of the ever-increasing amounts of digital information created for teaching and research.

⁶ For examples of domain repositories that implement policies and standards, see UniProt (<http://www.uniprot.org/>) and Linguistic Data Consortium (<http://www ldc.upenn.edu/>). For an example of a crowdsourcing approach to data management, see Galaxy Zoo (<http://www.galaxyzoo.org/>) or the Sloan Digital Sky Survey (<http://www.sdss.org/>). In the social sciences, see the Inter-university Consortium for Political and Social Research (ICPSR <http://icpsr.umich.edu/>)

⁷ See, for example, the NIH guidelines on genome-wide association studies (<http://grants.nih.gov/grants/gwas/>) and the National Academy of Sciences report, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (<http://www.nap.edu/catalog/12615.html>).

of data for long term persistent access, should be offered to those data that are designated as requiring an institutional commitment to stewardship. In this way, the University continues to fulfill its mission to create, disseminate, and preserve knowledge when that knowledge is in digital form.

Recommendations of the Task Force

- 1. Develop Digital Repository Services** that would allow researchers to easily store and manage research data outputs, to publish their datasets for online access at a stable web address and reference these datasets from publications, and be integrated into services for the preservation of the data.
- 2. Carry out Domain Specific Data Assessments** to investigate specific research centers, communities of practice, and specific academic disciplines to determine their data management, data sharing, and data curation requirements.
- 3. Develop Research Data Curation Services and Tools** to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers or knowledge-sharing databases, and/or to Yale's own digital preservation repository.
- 4. Establish a Digital Preservation Program for Research Data** to reduce the risk that valuable digital assets will be lost, and to establish a coherent, timely, and economically efficient program for persistent access to Yale University's digital research outputs.
- 5. Create a Consultation Center on Data Ownership, Intellectual Property, and Copyright** to assist researchers with the complexities of distributing, sharing, and publishing research outputs.
- 6. Address Inadequate Technical Infrastructure and Build Strategies** into business plans for addressing gaps in computing, storage, and networking.