Against the Grain

Volume 29 | Issue 5 Article 10

November 2017

Data Management and Preservation of Digital Research Data

Sayeed Choudhury
John Hopkins University and Institute for Data Intensive Engineering and Science, sayeed@jhu.edu

Data Management and Preservation of Digital Research Data

by **Sayeed Choudhury** (Associate Dean for Research Data Management, Sheridan Libraries, Johns Hopkins University and Institute for Data Intensive Engineering and Science) <sayeed@jhu.edu>

This article outlines a set of "principles for navigation" for how libraries could evolve data management services to support the changing needs of researchers. While the article provides a brief overview of the historical and current landscape, the recommendations are forward looking. The key principles of navigation:

- *Libraries need to move data management, particularly preservation, closer to the active phase of research.*
- *Linked data graphs offer a pathway to this active phase of research.*

By adopting these principles, libraries can shift the current "deposit and download data" approach to a more dynamic, iterative approach that fosters data use and preservation directly throughout the research and teaching process.

One of the central tenets of this article relates to the current gap between library-based data management services and the evolving nature of data-intensive research and teaching. At the crux of this gap is research libraries' current inability to connect effectively our data management services to the research workflows associated with increasingly large, complex data. Furthermore, libraries' data management services have yet to cohere into broader infrastructure. The principles of navigation concept refers to a recommendation from a report from an **NSF** funded workshop about infrastructure (Edwards et al. 2007). The authors of this report note the following two major points, successful infrastructure develops when smaller-scale community-based systems cohere and the socio-technical dimensions of infrastructure development, are as important as the technological dimensions.

From a socio-technical perspective, in terms of demand, many researchers make the following type of request: "I have 50 terabytes of data…could you help me preserve and provide access to them?" From the supply side, most libraries have focused on the "long tail" of research data which are typically characterized by a small number of researchers working together to use spreadsheets or standard database software. This misalignment of demand and supply has profound implications for the evolution of library-based data management services and the corresponding support for research and teaching.

These observations are based on the **Johns Hopkins University** Data Management Services (JHUDMS)[1] and the Data Conservancy[2] program. This article also reflects insight gained from the author's role as a member of the Executive Committee for the Institute for Data Intensive Engineering and Science (IDIES)[3] based at **JHU**.

JHUDMS was directly launched in response to **NSF's** announcement requiring data management plans as part of proposal submissions. Since the announcement of the White House Office of Science and Technology Policy (OSTP) memoranda on public access to publications and data (Holdren 2013), JHUDMS expanded to provide support for proposal submissions to other funding agencies.

JHUDMS data management consultants provide three types of services: consulting, training and archiving. Consultants offer pre-proposal submission support for creating data management plans. JHUDMS experience has demonstrated that this specific engagement is like a reference interview in that the consultation creates a deeper understanding of data management needs.

Barbrow et al. (2017) mentioned the JHUDMS training resources in their review article of research data management services.[4] One of the most encouraging aspects of these training efforts is that even seemingly simple contributions, such as file naming conventions, are appreciated by researchers. The fundamental premise behind these training efforts is that the data management plan is ideally the *beginning* of the process.

JHUDMS now provides additional services such as assigning DOIs. They support the **JHU** Data Archive but also suggest appropriate alternatives (e.g., ICPSR). The **JHU** Data Archive currently consists of a custom-built storage system and Dataverse but it is being migrated to a Fedora and Open Science Framework (OSF) platform.

While there have been successes for JHUDMS, there remains potential for growth. Much of JHUDMS' experiences with data management services have been transactional, rather than inspirational. Most research libraries could point to a collection (particularly special collections) and connect it to a faculty success story related to research or teaching. Most research libraries would find it more challenging to do the same with data under their stewardship. While this disconnect is a function of the relatively modest amounts of data in question, it also relates to a lack of integration between data management services and the research or teaching environments of our researchers.

Current Landscape

The most recent, relevant analysis to the role of libraries with data management is the **Association of Research Libraries (ARL)** Data Curation Spec Kit (Hudson-Vitale et al. 2017). The associated survey was designed to focus on data curation though the authors note that there remains confusion regarding the difference between data curation and data management. Some high-level findings include that most **ARL** libraries now provide some type of data curation support but there is great variability in the service offerings. Most **ARL** libraries rely on part-time effort from individuals with other responsibilities. The number of datasets under the stewardship of libraries is modest with most libraries counting less than fifty data sets. Much of the service offerings reflect the capabilities of the underlying technology. The respondent libraries indicated that the top three domains in terms of demand are social sciences, life sciences and arts & humanities. This observation resonates with the idea that most libraries have focused on the long-tail of data or spreadsheet-based research.

The author is a member of an EDUCAUSE Center for Analysis and Research (ECAR) working group[5] that is also examining data curation but with a broader viewpoint. The **ARL** SPEC Kit mentions that many libraries are considering which other units within their organizations should be involved in providing data curation services. The ECAR working group is comprised of individuals from different units within the university (e.g., library, central IT) from a range of institutions (e.g., R1 university, community college). Consequently, the ECAR report will describe findings from a broader constituency and offer recommendations for building an institution-wide strategy for data curation services.

A special issue of the *International Federation of Library Associations and Institutions (IFLA) Journal* (Volume 43, No. 1 – March 2017) featured multiple articles on global research data management services. Broadly speaking, institutions in Europe, Australia and New Zealand leverage funder mandates and national strategies and institutions in other regions of the world are conducting needs assessments as initial steps in developing research data services.

Within the U.S., an important driver for the creation of data management services was the OSTP memoranda. While these memoranda were created to foster greater sharing of data, there is a healthy degree of pragmatism within the guidelines. The memoranda acknowledge that there are certain conditions under which data should not be shared (e.g., privacy issues, national security issues). The memoranda further acknowledge that costs should be considered when managing data. While there is some movement by funders to encourage deposit into repositories or attach identifiers to data, such actions are not actually required.

The OSTP memoranda inspired libraries to launch data management services to assist researchers with their data management plans and actions. It has now been four years since

continued on page 32

Data Management and Preservation ...
from page 30

these memoranda were published. While the responses from federal funding agencies is undoubtedly evolving, there are a few current observations worth sharing.

Federal funding agencies continue to respond with a high degree of variability. A recent analysis by **Kriesberg** et al. (2017) affirms that "while some agencies, particularly those with a long history of supporting and conducting science, scored well, other responses indicate that some agencies have only taken a few steps towards implementing policies that comply with the memo." Given this type of environment, many researchers may have reached out for help initially but over time felt more confident in their own ability to manage data. Whether they are correct or not does not change their *perception* that they can manage data without help from the library.

How might libraries advance their data management services to engage our researchers more effectively, particularly as it relates to their research and teaching needs? The data management program at **JHU** might be instructive in this regard given the **JHU Sheridan Libraries'** long-term engagement with the Sloan Digital Sky Survey (SDSS).

Terminology

As evidenced within the **ARL** Spec Kit, there remains confusion regarding the term data management, which is often used interchangeably with data curation or data preservation. The Digital Curation Centre's (DCC) definition of data curation as "maintaining, preserving and adding value to digital research data throughout its lifecycle" is often cited, particularly since it reflects the research findings from information science researchers. **Hudson-Vitale** et al. (2017) built upon this definition of data curation by adding an emphasis on the "usefulness to scholarly and educational activities." The ECAR data curation working group has defined data curation as "the process by which data is put into a state and managed such that it can be understood and used by interested parties across disciplines and organizations."

Phillips et al. (2013) outlines the **National Digital Stewardship Alliance's (NDSA)** levels of digital preservation along the facets of storage and geographic location, file fixity and data integrity, information security, metadata, and file formats. The **NDSA** approach usefully affirms the importance of levels of service. Data management is not binary and it is an ongoing process.

For the launch of JHUDMS, **Choudhury** et al. (2013) developed a data management stack model comprising storage, archiving, preservation and curation. Fundamentally, this model delineates each of these layers of data management. Storage is defined as bits on tape, disk, cloud, etc. with backup and restore services. Archiving focuses on data integrity through fixity, identifiers, etc. Preservation is defined as providing enough metadata, context, representation information, etc. such that someone other than the original data producer can use and interpret the original data. Curation is defined similarly to the DCC. With this stack model as a reference, it reinforces the notion that data management services can span a wide range of capabilities.

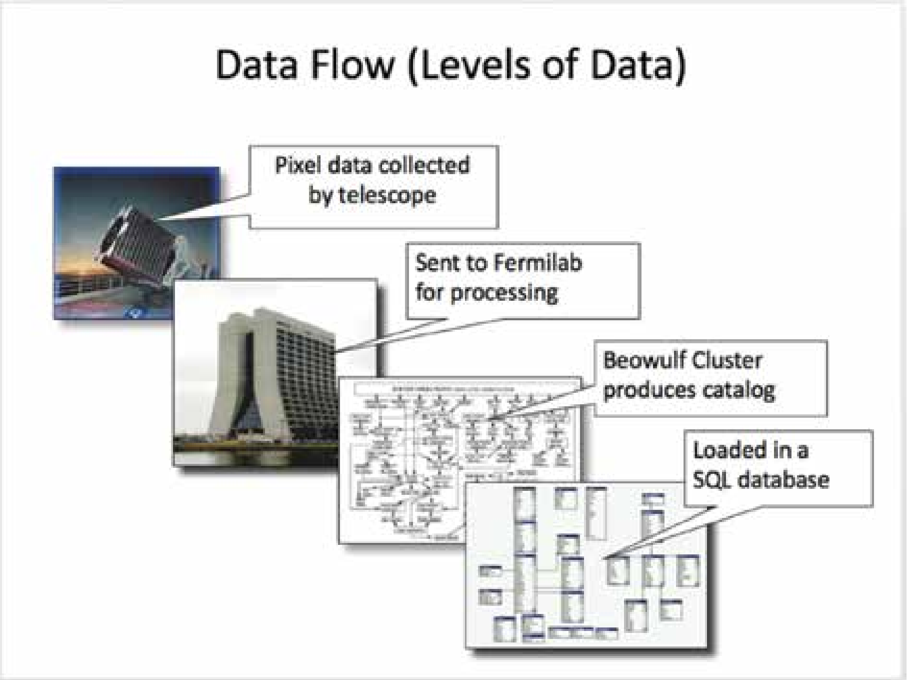
A Path Forward for Data Management Services

The **Sheridan Libraries'** data management program began in the early 2000s after a series of conversations between the author of this article and **Alexander Szalay**, a Professor of Physics and Astronomy at **JHU** and one of the principals for the SDSS project. **Szalay** was also the Principal Investigator for the **NSF** funded National Virtual Observatory (NVO).[6] While the NVO resulted in a framework and set of services for interoperability of astronomical data, it did not include mechanisms for preservation. One of the most important observations from these initial conversations relates to levels of data. **Szalay** conveyed that SDSS data are produced and processed in levels beginning with level 0 as bits from the telescope itself to level 3 as data releases in the form of SQL databases. Moving from level 0 to level 3 involves processing and calibration from raw, unprocessed data to more refined, accessible data. The figure below depicts these levels of data:

secondary analysis and re-use, they must become more involved in data management of all levels of data. **Choudhury** (2016) outlines a case that the private sector and government sector are currently making better use of data analytics at scale than libraries (and often by using other people's data).

It should be noted that working with all levels of data presents significant challenges. From a size perspective alone, the **Sheridan Libraries** has stored, archived (through fixity checking) and preserved (through a media migration) over 160 TB of SDSS data. In some of the scientific domains, the scale and complexity of data will challenge even universities' abilities to deal with them. **Professor Szalay** believes within ten years, all storage and computing will reside in third party providers such as **Amazon** Web Services.[7]

In this context, it will become critical for libraries to identify pathways to these large pools of data within third-party environments. The Center for Open Science's OSF (*https://osf.io*) provides one opportunity to do so. OSF is not a workflow tool or a repository, but rather a framework that interfaces with various tools, services and workflows. It has the merit of being used currently in the social sciences and life sciences (the domains of greatest adoption for libraries) with tens of thousands of users.



Data Flow (Levels of Data)

Pixel data collected by telescope

Sent to Fermilab for processing

Beowulf Cluster produces catalog

Loaded in a SQL database

The collaboration with the NVO identified yet another level 4 of data that result from analyses of level 3 data releases. These level 4 data are the ones cited in publications. With few, notable exceptions (e.g., **University of California San Diego's** Chronopolis), most libraries' research data management services target level 4 data.

While all levels of data are important, the issue with focusing on level 4 data is that they represent the end of a story related to data-intensive research. If libraries wish to support

The **Sheridan Libraries** has been working with the Center for Open Science to integrate Data Conservancy capabilities into OSF (Choudhury et al. 2017). Specifically, we are building the capability to package and ingest data in a linked data ready format into a Fedora repository that will be available as one of the default storage options within OSF. OSF does not currently include robust metadata capabilities. Rather than focus on metadata in the traditional sense, we are considering

continued on page 34

Data Management and Preservation ...
from page 32

whether it would be more desirable to support linked data within OSF project through the RMap[8] service. Researchers would be able to review linked data graphs and connections generated via RMap. This type of compound object represents a new form of publication that connects articles, data and software. Equally importantly, libraries would be able to connect the level 4 data within their repositories and OSF projects to the earlier levels of data that are used for research and teaching. This concept has already been demonstrated through a linked data representation of **ARL's** SHARE network and a pilot data rescue effort.[9]

This approach may help address the current issue that libraries' data management programs seem disconnected from the evolving nature of data-intensive research and teaching. Arguably, the physical sciences and engineering are developing the capabilities that social scientists and humanists ultimately adopt. If libraries wish to become more involved in the type of analytics, re-use, visualization, etc. that are the hallmarks of data-intensive research, there is an urgent need to develop infrastructure that is embedded within researchers' workflows and processes.

Acknowledgements

The author wishes to acknowledge **Alex Szlaay** for the figure on levels of data and to recognize the work of his colleagues. The work described in this article has been funded by **NSF**, **IMLS**, the **Sloan Foundation**, and **Johns Hopkins University**.

References

Barbrow, Sarah, **Denise Brush**, and **Julie Goldman**. "Research data management and services: Resources for novice data librarians." *College & Research Libraries News* 78.5 (2017): 274. Web. *https://doi.org/10.5860/crln.78.5.274*

Choudhury, G. Sayeed, **Rick Johnson**, **Jeffrey Spies**, and **David Wilcox**. **Choudhury, G. Sayeed**. "Data management at scale." *Information Services & Use* 36.1-2 (2016): 27-33. Print." Proc. of International Digital Curation Conference, Edinburgh. N.p., 2017. Web. 30 July 2017. *http://www.dcc.ac.uk/events/idcc17/research-practice-papers*

Choudhury, G. Sayeed. "Data management at scale." *Information Services & Use* 36.1-2 (2016): 27-33. Print.

Choudhury, G. Sayeed, **Carole L. Palmer**, **Karen S. Baker**, and **Timothy DiLauro**. "Levels of Services and Curation for High Functioning Data." Proc. of International Digital Curation Conference, Amsterdam. N.p., 2013. Web. 30 July 2017. *http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster192.pdf*

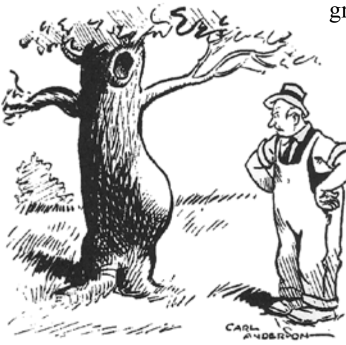
Edwards, Paul N., **Jackson J. Steven**, **Bowker C. Geoffrey**, and **Knobel P. Cory**. *Final report of the workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures*. Rep. National Science Foundation, Jan. 2007. Web. 30 July 2017. *http://hdl.handle.net/2027.42/49353*

Hudson-Vitale, Cynthia, **Heidi Imker**, **Lisa R. Johnston**, **Carlson Jake**, **Wendy Kozlowski**, and **Claire Stewart**. *SPEC Kit 354: Data Curation*. Rep. N.p., May 2017. Web. 30 July 2017. *http://publications.arl.org/Data-Curation-SPEC-Kit-354/*

Kriesberg, Adam, **Kerry Huller**, **Ricardo Punzalan**, and **Cynthia Parr**. "An Analysis of Federal Policy on Public Access to Scientific Research Data." *Data Science Journal* 16.0 (2017): 27. Web. *http://doi.org/10.5334/dsj-2017-027*

United States. White House. Office of Science and Technology Policy. *Increasing Access to the Results of Federally Funded Scientific Research*. By **John P. Holdren**. N.p., 22 Feb. 2013. Web. 30 July 2017. *https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf*

United States. Library of Congress. National Digital Stewardship Alliance. *The NDSA Levels of Digital Preservation: An Explanation and Uses*. By **Megan Phillips**, **Jefferson Bailey**, **Andrea Goethals**, and **Trevor Owens**. N.p., Feb. 2013. Web. 30 July 2017. *http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf*



Endnotes

[1] *http://dmp.data.jhu.edu/*

[2] *http://dataconservancy.org/*

[3] *http://idies.jhu.edu/*

[4] *http://dms.data.jhu.edu/training/*

[5] This working group is scheduled to publish its report in October 2017.

[6] The NVO is now known as the Virtual Astronomical Observatory.

[7] While this author was incorrect about Google Wave specifically, his CLIR thought piece "The Perfect Storm" highlighted this trend in 2009. Even if the specific predictions for timing are incorrect, the key point is the migration of storage and computing to third party providers seems realistic.

[8] *http://rmap-project.info/*

[9] The Data Conservancy YouTube channel (*https://www.youtube.com/user/dataconservancy*) and a recent blog post (*http://dataconservancy.org/data-conservancy-and-data-rescue-boulder-pilot/*) provide further information.

Rumors
from page 28

Media Briefings team is: **Matthew Ismail**, Managing Editor and the brains behind the Briefings, **Leah Hinds**, Executive Director, Charleston Conference, and **Tom Gilson**, liaison to **ATG Media**.

While we are on the subject of **peer review**, **Publons** has announced the winners of the **2017 Publons Peer Review Awards**, honoring the top contributors to peer review across all the world's journals. **Publons Peer Review Awards** were established in 2016 to celebrate the essential role peer reviewers play in bringing trust and efficiency to scholarly communication. It's thanks to their critical eye and devotion to sound science that high quality, impactful research is communicated to the world faster and more often. **Publons Awards** are designed to recognize both the quantity and quality of reviewers' efforts, and timed to coincide with **Peer Review Week** (September 11-17), a global event celebrating the critical role of peer review in science and research. Winners were selected from more than 190,000 researchers on **Publons' global reviewer database**. Following this announcement, **Publons** will reveal recipients of their **Inaugural Sentinel Award** — for outstanding advocacy, innovation or contribution to scholarly peer review. The shortlist was handpicked by a panel of judges from across the publishing industry and includes individual reviewers, career peer review advocates and experts. As we all know, **Publons** is part of **Clarivate Analytics**.

http://publons.com/awards/.

Erin Gallagher and the **Charleston Conference Directors** have been hard at work on the **Up and Comers** awards. These are librarians, library staff, vendors, publishers, MLIS students, instructors, consultants, and researchers who are new to their field or are in **the early years of the profession**. **Up and Comers** are passionate about the future of libraries. They innovate, inspire, collaborate, and take risks. They are future library leaders and change makers. And they all have one thing in common: they deserve to be celebrated. The **2017 Up and Comers** will be recognized in the **December17-January18 issue** of ***Against the Grain***, and 20 of these brilliant rising stars will be profiled in the same issue. In addition, they will be featured in a series of scheduled podcast interviews that will be posted on the **ATGthePodcast.com** website.

Gosh! Just heard from **October Ivins**! She and **Will Wakeling** are moving to Italy in December! They are buying a villa in the Abruzzo with five bedrooms for all the UK

continued on page 55