# Archival Options Appraisal: An Examination of Data Preservation Methods and Repositories

Carl Lew
Spring 2018
LIS 2674
Dr. Lyon

**Introduction**

Scholarship is built through sharing knowledge and information. Increasingly, scholarly findings are built upon digitally generated data. While this situation allows for rapid exchange of information, it comes to nothing if data is illegible, improperly described or stored in unsecure systems. For data to be used, widely accepted standards need to be followed so scientists and other researchers can access, analyze, cite and reuse digitally published data findings. Archivists, and the standards that guide their profession, are well suited to aid such researchers in their quest to store, describe and make accessible data. Since literacy developed, archivists have organized, described and made information, i.e. data, accessible. The twenty-first century forced archivists to establish systems and protocols that enable digitally published data to persist in a way that is coherent (is the data described properly?), findable (can other researchers access the data?), reusable (can other researchers build on or refute data based knowledge?) and verifiable (is the data what it purports to be?). Discussing archival theory and the practices connected to the management of related special collections provides a framework in which to situate current dilemmas in increasingly digital, data-driven scholarship. The archival theory of post-custodialism allows archivists a way to preserve and provide access to information they do not directly control – a critical, perhaps existential, challenge archivists currently face. The Archives of Scientific Philosophy at the University Pittsburgh displays the personal and institutional commitments necessary to effectively respond to the challenges of archiving specialized data. Examining digital repositories that serve both public and scholarly communities that are maintained or participated in by the University of Pittsburgh, provide archival options that can be developed with a commitment to data archiving.

Data can serve a variety of ends. While archivists may not possess the specialized knowledge of a doctoral level biologist, mathematician or philosopher, it is crucial for archivists to understand the needs of research communities. Without communication between archivists and scholars, data-driven research will be unable to successfully and routinely occur.

**Challenges**

A crucial challenge for archivists and scholarly researchers who want to preserve and make data accessible face is the development and maintenance effective communications over a substantial length of time. Continuity of best-practices and standards develops through collaboration with experts, and demand that archival professionals be at least familiar with the subjects held in their repository – why are researchers gathering data, how are they doing so and to what end? A library system that aims to be an intellectual access point for students and faculty will necessarily need to spend resources that train and develop their employees in a variety of contexts. A large school like the University of Pittsburgh can facilitate lifelong learning for their archival professionals. Doctoral research into specialized fields genomics and astrophysics may demand specialized knowledge that most archivists, who often have a background primarily in the humanities and an MLS, do not possess. The challenge for specialized repositories then, is finding candidates with not only rudimentary knowledge of such fields, but also with the desire and ability to continue learning about experimental methods and data collection practices for scholars who may be disinterested in archiving data, or who simply "out-class" the archivist whose amateur understanding could lead to conversational dead ends. Effective communication between traditionally disparate groups of experts depends on the adaptability of those groups – for them to modify their language,

concerns and understandings to facilitate the preservation and access of data. Adaptability needs to be a shared goal that facilitates the construction of new knowledge upon previous research throughout an institution. To do so requires groups of scholars to understand the fundamentals of the relevant scholarly ecosystem.

*Useful Framework: Archives of Philosophical Science*

To provide an example of how such communicative methods can benefit a scholarly community, it is beneficial to look outside the hard-sciences. The organization and preservation of the Archives of Scientific Philosophy at the University of Pittsburgh provides crucial lessons regarding data infrastructure and support services, and a glimpse into ways specialized knowledge can grow and be captured for continued analysis and use.

A collection from the mid-twentieth century – the Frank Plumpton Ramsey Papers[1] – provide access to specialized scholarly work that have been preserved digitally. For our purposes, the specific nature of the collection is irrelevant. But the infrastructure challenges to the preservation and access of these materials are similar to those of scientific research. As is the process of depositing the research into a digital repository. The Ramsey Papers provide philosophers of science with a view of how scholarship in a specific field, which displayed its own conventions and vocabulary, and developed in the mind of a leading thinker. Containing descriptions of content and context in the finding aid of the collection, the Ramsey Papers show how archivists can serve special collections. Whether one is digitizing paper that is falling apart, or entering data collected in the field or lab to a secure database, the end goal is

[1] "Guide to the Frank Plumpton Ramsey Papers, 1920-1930 ASP.1983.01," University Library System, University of Pittsburgh, http://digital2.library.pitt.edu/islandora/object/pitt%3AUS-PPiU-asp198301/viewer#ref12.

the same – to provide others with access to specific information for the sake of analysis, scrutiny and developing scholarship.

The Archives of Philosophical Science at the University of Pittsburgh provides archivists and scholars with a roadmap to successful data preservation. Infrastructural challenges posed by digitization are met with archival tools that allow for the dynamic use of digitized and born-digital data.[2] But hard-sciences demand an additional framework to include the "raw data" of experiments, training notes and field books. These resources provide relevant documentation to scientists who seek to build on or scrutinize previous hypothesis, experiments and conclusions.

*Useful Framework: Post-custodialism*

Archival theory is useful when examining the problem of preserving and making available data during this era. Most archival functions are embedded in the international data standards like the FAIR Principles. Archival appraisal, description, preservation and accessibility are intimately tied to current data sharing practices. But in a specialized scholarly, often scientific, environment, archivists need to modify their practices and expectations for materials. Post-custodialism is an archival framework that expresses the notion "that archivists will no longer physically acquire and maintain records, but that they will provide management oversight for records that will remain in the custody of the record creators."[3] The theory seems counter-intuitive to the idea of archives, and indeed, it does not fully capture the nuance of research data

[2] Islandora is the solution Pitt employs. See: "About," Islandora, https://islandora.ca/about.
[3] "Post Custodial Theory of Archives," Society of American Archivists, 2005, https://www2.archivists.org/glossary/terms/p/postcustodial-theory-of-archives.

archiving.[4] But post-custodial theory is a useful tool for thinking about managing data that may need to remain in the hands of the creator for continued analysis and use.

Scientific data displays characteristics that encourage post-custodial theory – they are created by a vast amount of people and institutions at a rapid pace, and the data presents infrastructure challenges to archivists, creators and users that demand creative solutions. To parse through these records and maintain robust, secure systems, participants need to collaborate to overcome archival challenges and explore the opportunities presented by digital records. During the late-twentieth century, Gerald Ham encouraged archivists to explore and "utilize…modern technology to provide easy and centralized access to increasingly complex and decentralized holdings."[5] Scientific research in an increasingly digital environment requires shared custody of data and related records. And I do think this is, in certain cases, not only necessary, but desirable. Digital preservation is a complex, fluid process that benefits from collective efforts.[6]

An R1, or doctrinal research, institution like the University of Pittsburgh must devote resources to the development of data repositories if the school wants to promote their mission and be a destination for scholarly researchers. Through proper channels of communication and systems analysis, archivists can work with research communities to actively maintain copies of fixed digital records that remain secure while being accessed remotely by a vast number of users. In this environment, archivists are not sole proprietors of a collection. They are active managers

[4] Digital archives certainly ingest, manage, maintain and control information in their care.
[5] Gerald Ham, "Archival Strategies for the Post-Custodial Era," *American Archivist* 44, no. 3 (Summer 1981): 211.
[6] I certainly lack technical expertise in the area of preserving digital data, records and related information. As a budding professional, it is crucial for me to recognize and rectify the situation. The latter requires further education and guidance – it would be dangerous to try alone.

who collaborate with and guide creators and technological specialists toward the preservation, organization and continued integrity of the records for scholarly or public use.

Archivists need to be aware of the records universe they are inhabiting and understand scholarly nuance. Though post-custodialism is not a perfect theory, it is a useful adaptation of archival thought and systems for the dynamic records universes of science. Rather than demanding sole control over unique records, archivists need to be comfortable with ensuring the fixity (i.e. integrity and security) of copies of potentially active data and related records. Scientific data captured and recorded digitally may need to be scrutinized immediately, and certainly continually. A scientific environment requires archivists to develop and adhere to systems that ensure access to secure, preserved and organized digital collections.

**Available Archival Options**

Researchers looking to publish their conclusions and data need a place to deposit their data. As an R1 school, the University of Pittsburgh must be committed to further research. To effectively endorse scientific research in a digital environment, Pitt provides or participates in a variety of projects that promote access points, integrity and preservation of data. Collaboration is crucial, and other large, R1 universities can be looked to for inspiration and as partners for continued development in this area. To build upon previous research in creative, useful ways requires inter-disciplinary and pubic facing approach – to encourage all members of an institution or place to engage with data and build their own datasets. Organizations like the Western Pennsylvania Regional Data Center (WPRDC) is a public database that can be added to and used by anyone with an interest in civic data. D-scholarship@Pitt is the University of Pittsburgh's current digital repository, open to anyone who studies at the school.

The University of Illinois project IDEALS[7] provides an additional example of ways universities like Pitt can continue to grow and develop data driven, digital scholarship. Collaborative efforts like Project Tycho and the Data Catalog Collaboration Project show how inter-institutional efforts encourage the growth of scholarly and scientific research that is not hindered by space and time.

Pittsburgh benefits from the presence of *two* R1 universities (with campuses within walking distance of one another) that encourage the highest standards in research and have encouraged the diffusion of data-driven examinations of the urban environment. The University of Pittsburgh has harnessed this opportunity when they collaborated with city leadership to develop the WPRDC – an online, civic database that opens up avenues for cooperation between the city leadership, population and resources.[8] The WPRDC allows the public, individuals and groups, who necessarily have differing skills and familiarity to data, to advocate for any number of ends, including resource allocation, transportation and public health and safety throughout the city and surrounding region. Future collaborations, throughout the United States and internationally, can use the WPRDC as a model from which to encourage civic discourse with data-driven, provable analysis – essentially creating an environment in which "smart" decisions with positive impacts upon communities.[9]

Collaboration should occur within organizations alongside outreach efforts. D-scholarship at Pitt is an institutional repository available for anyone within Pitt's scholarly

---

[7] "IDEALS Home," Illinois Digital Environment for Access to Learning and Scholarship," 2005-2013, https://www.ideals.illinois.edu/.
[8] "The Region's Data at Your Fingertips," Western Pennsylvania Regional Data Center, 2017, http://www.wprdc.org/.
[9] See Thomas H. Davenport, "Analytics 3.0," Harvard Business Review, December 2013, https://hbr.org/2013/12/analytics-30. Davenport discusses the impact of data analytics on industry, examples of which can be applied to the public sector if political will is available to do so.

community.[10] By accepting anything from undergraduate thesis on Aristotle to datasets regarding radiocarbon in Alaska, the University of Pittsburgh repository encourages the growth of scholarship by making high-level research available for curious users to peruse and potentially cite and provides alumni with space beyond personal storage for the longevity of their scholarship. Datasets, however, need to follow additional standards to be accessed, scrutinized and cited. The University of Pittsburgh provides for those needs with a guide to data sharing and digital scholarship.[11]

Developed by the University Library System, the document offers students a place to deposit their datasets with clear instructions for the format and necessary content of datasets that are supplementary to or displayed within the submission. D-scholarship at Pitts is mirrored by Illinois IDEALS out of the University of Illinois. Both repositories organize content, in part, by the community or school of scholars who developed the material, providing researchers with an opportunity to scrutinize scholarship across institutional lines.[12]

Project Tycho, of the University of Pittsburgh, and the Data Catalog Collaboration Project (DCCP) are efforts by researchers and librarians at academic health sciences libraries to share secure and useful data within and between institutions. Both are international in scope and seek to encourage the use and publication of health science data. Using the FAIR Guiding Principles as a benchmark of success, Project Tycho ensures the accessibility, integrity, preservation and use of health science data through internationally recognized standards.

---

[10] "D-Scholarship at Pitt: Institutional Repository at the University of Pittsburgh," University of Pittsburgh, 2018, http://d-scholarship.pitt.edu/.
[11] "Sharing Data and D-Scholarship@Pitt," University Library System, University of Pittsburgh, http://pitt.libguides.com/ld.php?content_id=29698922.
[12] Competently and detailed metadata regarding the provenance of the dataset, experiment and scholarship.

The DCCP also encourages access to clean descriptions of datasets but allows the creator or repository responsible for the data to allow or deny access to the full dataset.[13] This is a post-custodial adaptation to a wild environment where the amount scholarly data and analysis is increasing. By providing descriptions and location of datasets but leaving access decisions up to individual repositories or persons, the DCCP encourages communication between health science scholars and archivists.

**Conclusion**

Scholarship, no matter the specific research interest, is built though sharing knowledge and information. Increasingly, scientific, civic and other scholarly findings are built upon digitally generated data. While this situation allows for rapid exchange of information, it comes to nothing if data is illegible, improperly described or stored in unsecure systems. For data to be used, widely accepted standards need to be followed so researchers can access, analyze, cite and reuse digitally published data findings. Archivists, and the standards that guide their profession, are well suited to aid data such researchers in their quest to store, describe and make accessible data. Since literacy developed, archivists have organized, described and made information, i.e. data, accessible. The twenty-first century is forcing archivists to establish systems and protocols that enable digitally published data to persist in a way that is coherent (is the data described properly?), findable (can other researchers access the data?), reusable (can other researchers build on or refute knowledge through this data?) and verifiable (is the data what it purports to be?). A variety of current archival models, systems and standards examine solutions to these dilemmas and can be of use to scientists concerned with the preservation and accessibility of

[13] "About Us," Digital Catalog Collaboration Project, https://www.datacatalogcollaborationproject.org/about-us.

their data. Case studies of implementation of these archival methods at the University of Pittsburgh and surrounding region illuminate the challenges and opportunities urbane environments confront when archiving data. Successful communication and collaboration are crucial for useful data to be preserved, fixed and accessible. While archivists who come from the humanities may lack specific expertise in certain scholarly fields, without continued communication between those groups, data will die, and scientific research will be unable to responsibility and routinely build upon previous research.

**Bibliography**

Akmon, Dharma, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom. "The Application of Archival Concepts to a Data-Intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs." Archival Science 11, nos. 3-4 (2011): 329-348.

Conway, Paul. "Institutional Repositories: Is there Anything Else to Say?" *OCLC*. https://www.oclc.org/content/dam/research/events/dss/pdf/conway_presentation.pdf.

Davenport, Thomas H. "Analytics 3.0." *Harvard Business Review* (December 2013). https://hbr.org/2013/12/analytics-30.

Data Catalog Collaboration Project. "About Us." https://www.datacatalogcollaborationproject.org/about-us/.

Ham, Gerald. "Archival Strategies for the Post-Custodial Era," *American Archivist* 44, no. 3 (Summer 1981): 207-216.

Illinois Digital Environment for Access to Learning and Scholarship. "IDEALS Home" 2005-2013. https://www.ideals.illinois.edu/.

Ngoepe, Mpho. "Archival orthodoxy of post-custodial realities for digital records in South Africa." *South Africa, Archives and Manuscripts* 45, no. 1 (2017): 31-44, DOI: 10.1080/01576895.2016.1277361.

Preservica. "White Paper: Digital Preservation Maturity Model." 2014. https://preservica.com/uploads/resources/Preservica-White-Paper-Maturity-Model-2014_NEW.pdf

Society of American Archivists. "Post Custodial Theory of Archives." 2005. https://www2.archivists.org/glossary/terms/p/postcustodial-theory-of-archives.

University Library System, University of Pittsburgh. "Guide to the Frank Plumpton Ramsey Papers, 1920-1930 ASP.1983.01" http://digital2.library.pitt.edu/islandora/object/pitt%3AUS-PPiU-asp198301/viewer#ref12.

University of Pittsburgh. "D-Scholarship at Pitt: Institutional Repository at the University of Pittsburgh." 2018. http://d-scholarship.pitt.edu/.

Western Pennsylvania Regional Data Center. "The Region's Data at Your Fingertips." 2017. http://www.wprdc.org/.

Wilkerson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (2016). DOI:10.1038/sdata.2016.18.