Data Curation 101 for Theological Librarians

Clifford B. Anderson (Vanderbilt University)
Bobby Smiley (Vanderbilt University)

Abstract

A paper presented in June 2017 at the 71st Annual Meeting of the American Theological Library Association in Atlanta, Georgia; forthcoming in the 2017 Proceedings of the American Theological Association.

What is Data Curation?

Curating data is a new job skill for theological librarians. Apart from certain subfields that cross over with psychology and sociology, theological studies is not a data-driven discipline. Theological students have not needed to master statistics to study Augustine, Julian of Norwich, or Rosemary Radford Ruether. As theological researchers become interested in the digital humanities, they wind up producing data sets, which require description, preservation, and publication plans. The art of data curation is to guide researchers to sustainable and scalable practices of data sharing. Theological librarians have the opportunity to lead faculty and graduate students in these practices, steering them away from storing their data on thumb drives, network shares, and Dropbox to preserving their research in data repositories in standard formats with shared identifiers. A big task, no doubt!

Theological faculty and students can take solace in the fact that scholars in other humanities fields experience difficulty with data curation. As with the sciences and social sciences, funders' mandates to curate and preserve data impose new requirements on faculty and students in the humanities to think beyond their immediate projects to the longterm preservation of their digital products. And this has not been easy for humanities researchers. In a study of the data management plans of nineteen Office of Digital Humanities "Start-Up Grant" projects, Alex Poole notes that digital humanists may lack the ability to curate their data, despite their technical expertise. "A disjuncture persists between the types of innovative scholarly projects in which these scholars are engaging and their ability to curate the data underpinning these projects to enable future reuse."[1] Could librarians, who typically think in longer time frames,

[1] Alex H. Poole, "'A Greatly Unexplored Area': Digital Curation and Innovation in Digital Humanities," *Journal of the Association for Information Science and Technology* 68, no. 7 (July 2017): 1778, doi:10.1002/asi.23743.

assist scholars with developing preservation plans? Yes, perhaps, but scholars do not yet think of the library as a place to seek counsel for research data management. While some NEH grant recipients turned to librarians for help with data curation task, the majority muddled through on their own.[2]

Data Curation Lifecycle

The Digital Curation Centre in the United Kingdom published a "Curation Lifecycle Model," which has proved highly influential during the past ten years. As Sarah Higgins remarks, "The ... Model offers a graphical high-level overview of the lifecycle stages required for successful curation."[3] The key is that data curation should take place throughout the research process, not at the beginning and the end. Librarians may collaborate with researchers about the data management plans, metadata description, data preservation, and data publication, among other topics. Rather than attempting to address these issues upfront, librarians should check in periodically with research teams to address concerns that have popped up and to provide recommendations about best practices.

A practical way to get started with data curation at your institution is to sign up for the DMPTool or Data Management Planning Tool. The DMPTool is software-as-service provided by the UC3 (University of California Curation Center) at the California Digital Library (see https://dmptool.org/). The tool allows researchers to craft customized data management plans by selecting templates for the relevant funding agency. While any scholar can use the DMPTool without charge, institutions can also become supporting partners. Among the benefits of institutional membership is the ability to embed librarians into librarians into the data management planning workflow, allowing them, for instance, to read and comment on drafts (see https://github.com/CDLUC3/dmptool/wiki/FAQ#q-what-are-the-benefits-of-becoming-a-dmptool-partner). A goal of the DMPTool is providing a means for the "research community to gain insight into the methods and practices of research data management across the entire lifecycle at both a micro and a macro level."[4] Librarians who counsel researchers from different domains will likewise be able better to understand existing institutional patterns for data management and to guide researchers toward a common, contextual set of best practices.

[2] Ibid., 1777.

[3] Sarah Higgins, "The DCC Curation Lifecycle Model," *International Journal of Digital Curation* 3, no. 1 (December 2008): 134–40, http://ijdc.net/index.php/ijdc/article/view/69.

[4] Andrew Sallans and Martin Donnelly, "DMP Online and DMPTool: Different Strategies Towards a Shared Goal," *International Journal of Digital Curation* 7, no. 2 (October 2012): 128, doi:10.2218/ijdc.v7i2.235.

Ethics of Data Collection

The collection and curation of research data involves ethical considerations as well as technical perspectives. For instance, researchers who work with human subjects should already be working with institutional review boards (IRBs) to review and refine their data collection and dissemination practices.[5]

Researchers working in areas that do not fall under the purview of institutional review boards should also consider the ethical implications of their data collection and management.[6] In an article titled "What is Data Ethics?" that introduces a themed issue on the topic, Luciano Floridi and Mariarosaria Taddeo provide an expansive definition of data ethics.

> ... Data ethics can be defined as the branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).[7]

The point of such reflection is to steer a course between "social rejection" and "legal prohibition" of data-driven research.[8] While theological librarians may not become specialists in the complexities of data ethics, they should learn to spot potential issues about data confidentiality, for example. If research data contains potentially identifying information about human subjects or traditional knowledge about indigenous communities, librarians should generally refer researchers to specialists who can help redact the data before public release.

Data Repositories

Where should researchers preserve and make accessible their data? An existing institutional archive is an option. While institutional repositories are more typically geared toward the dissemination of pre-prints and post-prints of academic papers, most can also handle datasets.[9] Given that major divinity schools and

[5] In Laura Stark, *Behind Closed Doors: IRBs and the Making of Ethical Research* (Chicago: University Of Chicago Press, 2012), Laura Stark details how religious believers from the so-called 'peace churches' like the Mennonites played a role in the development of institutional review boards.

[6] See John Leslie King, "Humans in Computing: Growing Responsibilities for Researchers," *Communications of the ACM* 58, no. 3 (2015): 31–33, doi:10.1145/2723675.

[7] Luciano Floridi and Mariarosaria Taddeo, "What Is Data Ethics?" *Philosophical Transactions of the Royal Society A* 374, no. 2083 (December 2016): 3, doi:10.1098/rsta.2016.0360.

[8] Ibid., 1.

[9] On the use of DSpace as a data repository, see Yin Zhang and Hsin-liang Chen, "Data Management and Curation Practices: The Case of Using DSpace and Implications," *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2015): 1–4,

seminaries already maintain institutional repositories, librarians at such schools may naturally turn to them when seeking longterm storage for datasets. Adapting existing institutional repositories to handle research data may require making significant changes to workflows[10] and may not provide adequate metadata[11] or, when dealing with big data, enough file storage and network throughput.[12]

When advising researchers, librarians should also consider so-called "domain" and "disciplinary" repositories[13] as well as generic data repositories. By contrast to institutional repositories, these varieties of repository serve broader and narrower communities of practice. An example of a domain repository for religion and theological studies is the ATLA Digital Library (see http://dl.atla.com/). A sermon database might function as a disciplinary repository; sermons belong to the general domain of theology, but fall more narrowly into the category of practical theology. To foster discoverability of their datasets (and to take advantage of domain and discipline-specific functionality), researchers may wish to deposit data in these kinds of repositories. Generic repositories like Figshare (see https://figshare.com/) and Zenodo (see https://zenodo.org/) serve data-driven researchers across fields and domains. Among the high-level tools for data management they provide are versioning of Document Object Identifiers (DOIs), altmetric analysis, large file storage, integration with Github, and exporting metadata to citation management software, though products differ in capability and cost. By contrast, institutional repository software may not support this functionality, rendering them less suitable for depositing specialized forms of research data.[14]

Lastly, theological librarians should also assist researchers with finding the appropriate licenses for research data. While the Creative Commons licenses are familiar, they may not be the best choice for data. In the United States, at least, not all data is not straightforwardly copyrightable; copyright requires that the data contain a "modicum of creativity."[15] The act of assembling research data, while arduous, may not merit copyright in a legal sense. Even if it does, making research data available without copyright restrictions (by using a CC0 or public domain license; see https://creativecommons.org/share-your-work/public-domain/cc0/) will foster the greatest downstream reuse. Researchers who fear that they will losing credit for their work can still ask that scholars who draw on their datasets cite them (with the appropriate DOI) in their publications.

---

doi:10.1002/pra2.2015.1450520100109.

[10] Consider, for example, the case study of retrofitting an existing repository for research data at the Hong Kong University of Science and Technology (Gabrielle K. W. Wong, "Exploring Research Data Hosting at the HKUST Institutional Repository," *Serials Review* 35, no. 3 (September 2009): 125–32, doi:10.1016/j.serrev.2009.04.003)

[11] Dong Joon Lee and Besiki Stvilia, "Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff," *PLOS ONE* 12, no. 3 (March 2017): 24f, doi:10.1371/journal.pone.0173987.

[12] Line Pouchard, "Revisiting the Data Lifecycle with Big Data Curation," *International Journal of Digital Curation* 10, no. 2 (2015): 187, doi:10.2218/ijdc.v10i2.342.

[13] Lee and Stvilia, "Practices of Research Data Curation in Institutional Repositories," 3.

[14] Ibid., 30.

[15] See Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991)

Collecting and Visualizing Library Data

While the majority of data curated by libraries are sourced from research, libraries themselves already have, as well as generate, a great deal of data in need curation. Circulation statistics, holdings lists, monographic expenditures, public services activity, *inter alia*, are all data libraries collect and curate. Leveraging well-curated library data holds manifold possibilities for analysis, and what follows are examples of how collections, and circulation statistics can be harnessed for data visualization.

Shortly after arriving at Vanderbilt, one of the co-authors wanted to know more about the library's holdings and circulation data. Working with Tao You, our systems librarian, he was able to obtain circulation data broken down monthly by branch library. These data were retrieved through the library's ILS API, and made available as an Excel spreadsheet. Using the data visualization software Tableau Public (see https://public.tableau.com) a series of small multiple line graphs showing monthly circulation divided annually were generated to track the frequency of check-outs for the four largest branch libraries by monographic holdings (excluding the law library).

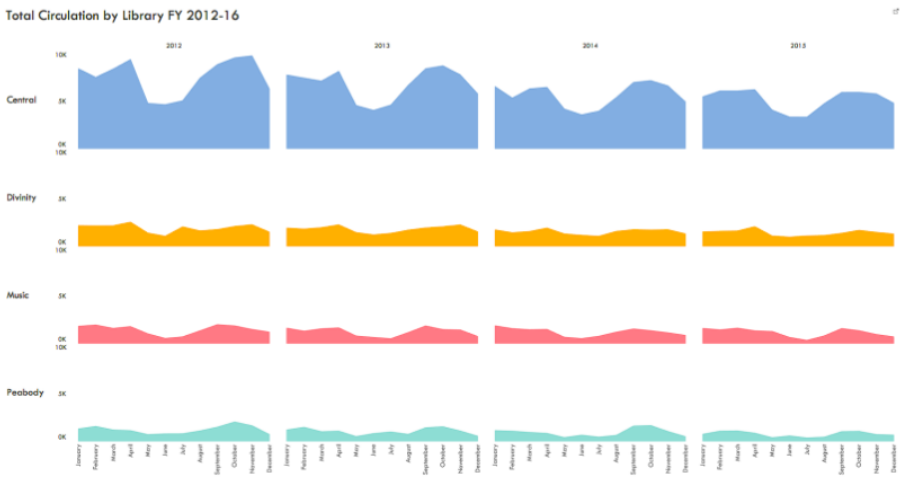

Figure 1: Small Multiples

Several features are immediately manifest: first, and not especially surprising, monthly trends match the rhythms of the academic calendar, with crests in mid-spring and late fall, and a sizeable trough during the summer months. More interesting is the diminution of total check-outs for all libraries over a four-year period, albeit with only slight variance for three of the four libraries. A cursory glance might suggest an overall decline in library use, but recognizing that the data used only includes physical monograph check-outs another trend is adumbrated. Over the period surveyed, Central Library has been embracing

an e-book preferred purchase policy, in contradistinction to, for instance, the Divinity Library, where physical monographic purchases are privileged over acquiring electronic copies (ebooks are still purchased, however). In this instance, the visualization reveals how ebook purchases can problematize traditionally used statistics (e.g., circulation data), and enjoins us to consider how ebook usage can be tracked and then mapped on physical book circulation.

Combining circulation data with holdings data evidences other trends less easily discerned from examining data in a tabular format alone. In addition to circulation data, our systems library provided holdings data by LC Class for Central Library, our largest branch library. When these are data dropped into Tableau Public, a treemap can be rendered, which is heat mapped to indicate circulation frequency over a single academic year and scaled by holdings data.

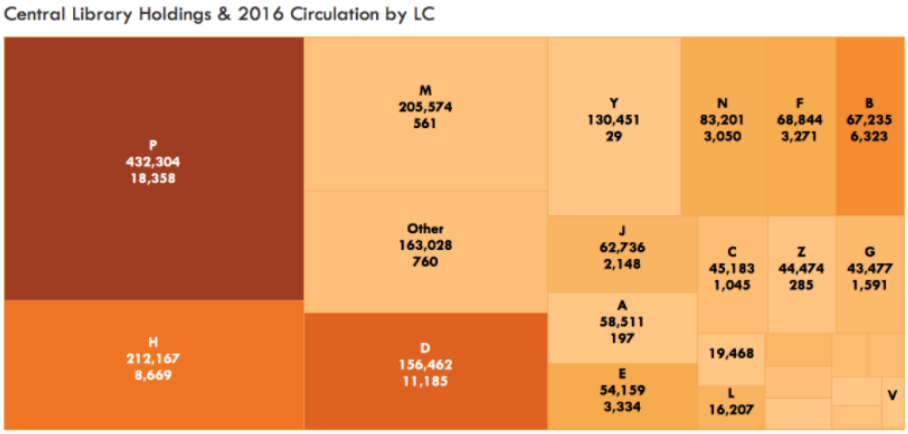

Figure 2: Circulation and Holdings Tree Map

In this treemap, several aspects of the library's collection are immediately revealed. First, the Central Library's collection strengths reside in literature (P) and social sciences (H), and the circulation frequency seems commensurate with collections of those sizes. What struck the co-author as less expected was the third largest collection by LC Class is M (Music), which surprisingly has around 50,000 more items than history (D). Indeed, Central's M collection is almost double the size of the Music Library's holdings (205,574 in Central, 104,640 listed in Music).[16] Spotting presumed anomalies like this helps elicit questions about how and why collections look the way they do — in addition to highlighting gaps and surpluses.[17]

[16]These holding statistics were also included the files provided by the Systems Librarian.

[17]A more potentially arresting example of this kind of analysis is discussed in the co-author's keynote address at the 2016 ATLA Annual Meeting; see Bobby Smiley, "Theological Librarianship in the Age of Digital Humanities," ed. Tawney Burgess, *Summary of Proceedings, 70th Annual Conference of the American Theological Library Association*, 2016, 22–32

Marshaling these data can be useful for making decisions about library space, and the off-siting or deaccessioning of material. Sourced from a highly granular Library of Congress SCAT (Statistical Category Abstract Table) from the Michigan State University Library (one of the co-author's previous employers), Figure 3 demonstrates how physical space can be visualized with these data. In the example below, another treemap displays a single floor of books broken into quadrants (that mirrors the floor's layout), broken down further by call number, and heat mapped for circulation.

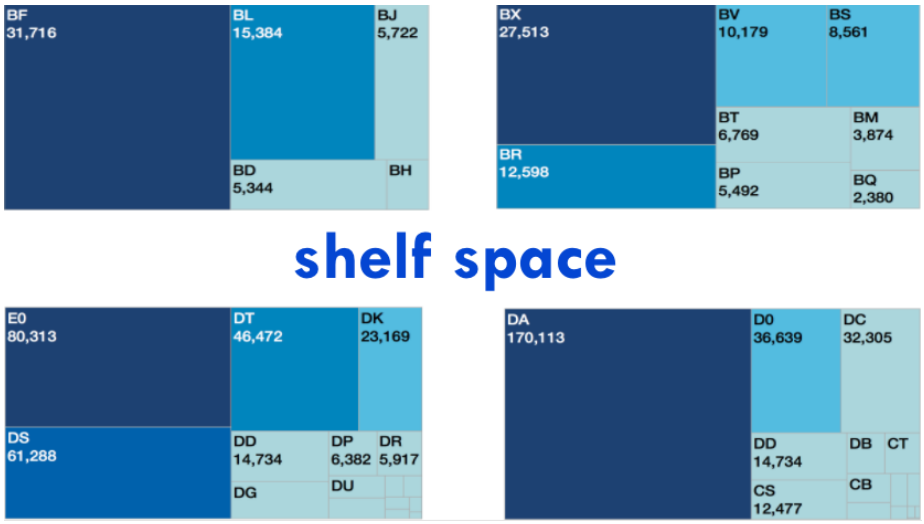

Figure 3: Floor Space Tree Map

Each quadrant represents a grouping of ranges, but it can be easily imagined that each range of shelves could also be broken down in a similar fashion. With diminishing space a critical concern, librarians with collections responsibilities are provided with additional visual tools to help inform decisions about how to manage or grow a collection.

These examples illustrate how well-curated library data be can harvested and used as potentially compelling data points for collections decisions, ranging from spotting collection trends and gaps to reimagining stacks maintenance. While these examples focus on collections, all library data are a rich source for computational visual analysis. Along with research data, the foregoing underlines how library data too needs to be curated, and in so doing, how it can be effectively deployed to aid in making data-driven decisions.

Works Cited

Floridi, Luciano, and Mariarosaria Taddeo. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society A* 374, no. 2083 (December 2016): 20160360. doi:10.1098/rsta.2016.0360.

Higgins, Sarah. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation* 3, no. 1 (December 2008): 134–40. http://ijdc.net/index.php/ijdc/article/view/69.

King, John Leslie. "Humans in Computing: Growing Responsibilities for Researchers." *Communications of the ACM* 58, no. 3 (2015): 31–33. doi:10.1145/2723675.

Lee, Dong Joon, and Besiki Stvilia. "Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff." *PLOS ONE* 12, no. 3 (March 2017): e0173987. doi:10.1371/journal.pone.0173987.

Poole, Alex H. "'A Greatly Unexplored Area': Digital Curation and Innovation in Digital Humanities." *Journal of the Association for Information Science and Technology* 68, no. 7 (July 2017): 1772–81. doi:10.1002/asi.23743.

Pouchard, Line. "Revisiting the Data Lifecycle with Big Data Curation." *International Journal of Digital Curation* 10, no. 2 (2015). doi:10.2218/ijdc.v10i2.342.

Sallans, Andrew, and Martin Donnelly. "DMP Online and DMPTool: Different Strategies Towards a Shared Goal." *International Journal of Digital Curation* 7, no. 2 (October 2012): 123–29. doi:10.2218/ijdc.v7i2.235.

Smiley, Bobby. "Theological Librarianship in the Age of Digital Humanities." Edited by Tawney Burgess. *Summary of Proceedings, 70th Annual Conference of the American Theological Library Association*, 2016, 22–32.

Stark, Laura. *Behind Closed Doors: IRBs and the Making of Ethical Research*. Chicago: University Of Chicago Press, 2012.

Wong, Gabrielle K. W. "Exploring Research Data Hosting at the HKUST Institutional Repository." *Serials Review* 35, no. 3 (September 2009): 125–32. doi:10.1016/j.serrev.2009.04.003.

Zhang, Yin, and Hsin-liang Chen. "Data Management and Curation Practices: The Case of Using DSpace and Implications." *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2015): 1–4. doi:10.1002/pra2.2015.1450520100109.