Conceptualizing Data Curation Activities Within Two Academic Libraries

Sophia Lafferty-Hess[1], Julie Rudder[2], Moira Downey[3], Susan Ivey[4], and Jennifer Darragh[5]

May 29, 2018

Abstract

A growing focus on sharing research data that meet certain standards, such as the FAIR guiding principles, has resulted in libraries increasingly developing and scaling up support for research data. As libraries consider what new data curation services they would like to provide as part of their repository programs, there are various questions that arise surrounding scalability, resource allocation, requisite expertise, and how to communicate these services to the research community. Data curation can involve a variety of tasks and activities. Some of these activities can be managed by systems, some require human intervention, and some require highly specialized domain or data type expertise.

At the 2017 Triangle Research Libraries Network Institute, staff from the University of North Carolina at Chapel Hill and Duke University used the 47 data curation activities identified by the Data Curation Network project to create conceptual groupings of data curation activities. The results of this "thought-exercise" are discussed in this white paper. The purpose of this exercise was to provide more specificity around data curation within our individual contexts as a method to consistently discuss our current service models, identify gaps we would like to fill, and determine what is currently out of scope. We hope to foster an open and productive discussion throughout the larger academic library community about how we prioritize data curation activities as we face growing demand and limited resources.

Keywords: data curation, digital repositories, institutional repositories

---

[1]*Duke University Libraries, Research Data Management Consultant;* sophia.lafferty.hess@duke.edu
[2]*University of North Carolina at Chapel Hill Libraries, Repository Program Librarian;* jrudder@email.unc.edu
[3]*Duke University Libraries, Repository Content Analyst*; moira.downey@duke.edu
[4]*Duke University Libraries, Repository Content Analyst*; susan.ivey1@duke.edu
[5]*Duke University Libraries, Research Data Management Consultant;* jennifer.darragh@duke.edu

Introduction

Transparency and openness in science is increasingly viewed as one of the cornerstones of reliable and reproducible research (Munafò et al., 2017; Nosek et al., 2015). As funders and journals increasingly enact policies that require researchers to effectively manage their data and share that data openly (NSF, 2010; PLOS, 2014), academic libraries have been developing research data management (RDM) programs to support researchers' needs (Fearon et al., 2013; Raboin, Reznik-Zellen, & Salo, 2012). One aspect of RDM programs is building and supporting institutional repository systems for the stewardship and dissemination of research data. Some libraries are going one step further by providing *data curation services* that add value to research data and help make data more accessible and reusable.

The recent Association of Research Libraries (ARL) SPEC Kit 354 closely examined the data curation services of 80 ARL member institutions (Hudson-Vitale et al., 2017). While the survey found that 51 libraries indicated that they are providing some form of data curation services, the authors also noted that "respondents conflated data curation activities with research data management services…this indicates that a common understanding of data curation is not widespread or ubiquitous" (p. 12). Research data management is a broad term that refers to the work performed by researchers and others throughout the research lifecycle that supports the preservation, access, and use of data including writing data management plans, organizing and documenting data, formatting data, and archiving and sharing data. While data curation can be considered a part of research data management writ large it can also be more narrowly defined as "the encompassing work and actions taken by curators of a data repository in order to provide meaningful and enduring access to data." (Johnston et al., 2016).

The SPEC Kit used 47 data curation activities, initially developed by the Data Curation Network project, to assess the types of activities currently performed, as well as the activities that libraries are interested in providing in the future. These 47 activities offer a useful model for considering the scope and breadth of data curation services. The SPEC Kit concludes that they found a "wide variability in data curation services" and suggests that "as libraries grow and strengthen their positions as centers of data curation, recursive efforts to convey their activities meaningfully and consistently, both internally and externally, will be of benefit" (p. 13). How libraries engage in data curation activities has also been examined through case studies and interviews with staff working within institutional repositories (Johnston, 2017; Lee & Stvilia, 2017). Further research will provide a growing foundation for libraries to engage in dialogues around data curation service models.

In the summer of 2017 at the Triangle Research Libraries Network Institute, which took as its focus "Supporting New Directions and Projects in Scholarly Communication," library staff members from the University of North Carolina at Chapel Hill and Duke University used these 47 activities outlined by the Data Curation Network as a starting point to discuss data curation services within our own institutions (TRLN Institute, 2017). The goals of this "thought-exercise" was to demystify data curation for our local contexts and empower staff to have fruitful discussions surrounding data curation services, systems, and staffing, both within the library and with external stakeholders. These types of discussions can also provide a basis for improving, communicating, and scoping services as policies are enacted, new systems are developed, and staff decisions are made.

Two Repository Settings

University of North Carolina at Chapel Hill

Since 2010, the Carolina Digital Repository[1] at University of North Carolina at Chapel Hill has offered a place for individual researchers to host and preserve up to 2 TBs of data. The Carolina Digital Repository (CDR) is an institutional repository that supports preservation and access for multiple content types, which include data, scholarly articles, mediated deposit of student papers, and the born-digital collections of the Louis Round Wilson Special Collections Library. CDR is based on the Fedora Commons framework and built using the OAIS framework model for preservation (Consultative Committee for Space Data Systems, 2012). Until recently, data deposits have been mediated and supported by one repository librarian and three software developers. In addition, UNC's Odum Institute provides a data archive and data archiving services that are available to UNC researchers (Odum Institute, n.d.). UNC Libraries also offers research data management consultation and visualization services.

In 2015, the UNC Faculty Council passed an Open Access resolution and with it came support for additional resources. The CDR is now staffed with four repository developers, two librarians, one support staff, one student worker and metadata support from another unit. These staff are responsible for the systems, content, and service of the CDR, and research data support continues to be only one part of the responsibilities and focus of the team. The CDR currently offers self-deposit for research data, as well as a mediated deposit service for larger files and collections. In 2018, the repository team plans to migrate the IR content (which includes support for research data) to a new system based on the community supported repository platform, Hyrax. These changes present us with a great opportunity to define our data curation services and systems to better support researchers' needs.

[1] https://cdr.lib.unc.edu/

Duke University

Duke University has operated with an Open Access (OA) policy in place since 2010, when the University's Academic Council voted in favor of a new digital repository for scholarly publications and writings (Mock, 2010). The Duke Digital Repository (DDR)[2], which grew out of this mandate to support open access publishing on campus, has broadened in scope to support distributed software stacks managing a diverse array of content. In the fall of 2015, Duke convened a Digital Research Faculty Working Group that included a number of campus faculty and campus IT administrators, as well as the Associate University Librarian for Digital Strategies and Technology. The group discussed services and support for the increasing volume of digital research and data output by faculty and researchers, and explicitly endorsed the DDR as a tool to support the campus OA policy, compliance with federal funding agency publication and future data retention mandates, and long-term discovery, access, and preservation of faculty research and scholarly output ("Digital Research Faculty Working Group", n.d.). The working group's recommendations included the creation of four new library staff positions to conduct this work--two Senior Research Data Management Consultants and two Digital Repository Content Analysts.

Since bringing these four positions online, staff have worked toward creating a suite of data curation services while simultaneously rethinking the software infrastructure required. In concert with the Content Analysts, the Data Management Consultants have established a pre-publication workflow for ensuring the quality of submitted datasets. While data presently resides in a content-agnostic Samvera application rooted in the Fedora Commons framework, work has begun with the repository development team to build a dedicated application based on the Samvera community's Hyrax platform and migrate all existing data.

Conceptualizing Data Curation

A first step when developing and implementing a data curation program is to clearly identify the programmatic goals in order to more effectively measure success. One goal is to help researchers meet the FAIR (Findable, Accessible, Interoperable, and Reusable) guiding principles for scientific data management and stewardship (Wilkinson et al., 2016). Another goal for a data curation program within academic libraries is to meet the specific needs of researchers, which

[2] https://repository.duke.edu/

include assisting them with policy compliance, increasing the impact and visibility of their research, and facilitating access and reuse of data.

Keeping these two goals in mind, the TRLN team evaluated and discussed our own unique contexts specifically related to staffing models, curation activities, and internal long-term goals for each program. While looking over the extensive DCN activities, the team reflected on how to determine what are the most essential activities in the face of limited resources. The team then began a process of grouping the various types of curation activities into three distinct "levels" to provide a structured model to better conceptualize and communicate about the provision of data curation within our individual contexts.

The table below presents these three levels of curation. Curation involves both tasks performed by systems and those performed by human capital (to varying degrees of human involvement). Level 1 curation focuses on a repository program that facilitates self-deposit or mediated deposit (which means some potential human mediation of the data package), and are generally supported by the system or repository policies. Level 2 curation builds on those tasks outlined in Level 1 by providing a more thorough review and potential enhancement of the data package. Level 2 services may be conducted by library staff with general knowledge of data management and curation best practices. Level 3 curation involves the manipulation of datasets, as well as more specialized services, such as data cleaning and code review. Level 3 services also require human intervention by staff with both general knowledge of data best practices and potential domain-specific and data type knowledge.

Activities identified by the DCN can be broad and may carry across levels. These activities can be multifaceted and have been defined in various ways within information and library science. Because of this multifaceted nature, we have placed a few activities within multiple levels, and these are italicized within the table. For example, quality assurance, as defined by DCN, includes many tasks, from reviewing the documentation and metadata for completeness to validating, cleaning, and enhancing data. We see quality assurance on a continuum where a data curation program may provide a more cursory review of files (Level 2) to a more in-depth comprehensive review (Level 3). Similarly, rights management within Level 1 would involve facilitating data depositors to assign a license to the data package, whereas in Level 2 the repository would support more work-intensive procedures for access and reuse, such as facilitating the collection of Data Use Agreements. See the Appendix for a more thorough description of these activities that we have identified as spanning levels.

For full definitions of data curation activities see the Data Curation Network: Data Curation Terms and Activities.

Level 1 Curation	Level 2 Curation	Level 3 Curation
Ingest ● Authentication ● Chain of Custody ● Deposit Agreement ● Documentation ● File Validation ● Metadata Appraise/Accept ● *Rights Management (licenses)* Curate ● Arrangement & Description ● File Inventory or Manifest ● Indexing ● Persistent Identifier ● Transcoding Access ● Contact Information ● Data Citation ● Discovery Services ● Embargo ● File Download ● Full Text-Indexing ● Metadata Brokerage ● *Restricted Access (system automated)* ● Terms of Use ● Use Analytics Preserve ● File Audit ● Migration ● Secure Storage ● Succession Planning ● Tech/Monitoring Refresh ● Versioning ● Cease Data Curation	Appraise/Accept ● *Rights Management (DUAs)* ● *Risk Management (file review)* ● Selection Curate ● Contextualize ● Curation Log ● File Format Transformations ● File Renaming ● *Quality Assurance* ● Restructure Access ● *Restricted Access (mediated requests)* Preserve ● Repository Certification	Appraise/Accept ● *Risk Management (remediation)* Curate ● Code Review ● Conversion (Analog) ● Data Cleaning ● De-Identification ● Interoperability ● Peer Review ● *Quality Assurance* ● Software Registry Access ● Data Visualization Preserve ● Emulation

The purpose of this exercise was largely to identify minimal baseline curation activities expected from systems that handle research data and staff who provide data curation. It should be acknowledged that this is not intended as a prescriptive guideline and there are additional solutions for data hosting; not every institution will have the resources--staffing or otherwise--to offer these types of curatorial services. Moreover, each of the curation activities is subject to interpretation based on needs specific to the individual institution. Each institution will have to decide which level is appropriate given their available resources, and then must interpret what each curation activity will mean given their specific context. This baseline has been helpful in identifying gaps in our own systems and services as we work to move our community's data toward meeting FAIR data principles and meeting researchers' needs. To address these gaps at the narrow institutional level, this exercise has helped both institutions develop policies and procedures surrounding data curation, while also helping us prioritize resource allocation and software development. In a broader sense, this effort has helped us provide some direction to the community working with the open-source software used by both institutions by prioritizing the features and functionalities we will need in pursuit of making data FAIR.

Data Curation in Practice

University of North Carolina at Chapel Hill

At UNC, we are aiming for curation services at Level One with the hopes to add a couple of high value activities from Level Two. With our current staffing levels, rate of deposit (about 1 dataset per month), and other services at the library and on campus, we are looking for the right balance and the specific value we add. We do not currently have any staff positions dedicated to data curation as either generalists or with a specific domain expertise; however, we want to offer some curation services and we want those to be the most useful activities to help us achieve our goals. With an Institutional Repository Librarian and a Content Technician who have responsibilities that cover many services and content types, we believe we have some room to improve our services from our current offerings so we are experimenting with adding a few activities from Level Two. For example, we are in the process of rewriting our policies and procedures to include staff review of each data submission in order to ensure that contextual documentation is included and to ensure files are in open formats. As non-data specialists, we will not be reviewing the actual contents of the data at this time. In addition, we will provide suggestions for basic arrangement and description. We are also considering running datasets through a PII checker since it is relatively easy to do with existing desktop tools such as Bulk Extractor.

We have also used this chart to help assess our next IR platform, Hyrax. It helps to determine our local feature roadmap and to advocate for the baseline capability that a data repository should

have (and the reasons why!). The chart and the activities listed have also helped us identify areas of training and expertise needed to perform these functions. We can begin to tease out the differences in research data management services and data curation activities and build a local shared vocabulary. If we decide in the future that we would like to provide more data curation services (and we have the resources to do so) this chart gives us a level of specificity to talk about those expansions, which is something that we previously lacked.

Duke University

At Duke, we are aiming for curation services in line with Level Two. Our present staffing model affords us the resources to conduct the higher-touch tasks that require extensive human intervention; both the Data Management Consultants and Repository Content Analysts are familiar with general data curation and management best practices, but perhaps do not possess the expertise required to provide the specialized services specified by Level Three for all disciplines and data types. On receipt of data from depositing researchers, the Data Management Consultants will engage in quality assurance of the data package that is consistent with Level Two, examining depositor-supplied documentation and metadata for completeness and comprehensibility, opening data files and performing a general review for potential issues, visually checking for the presence of personally identifiable information or personal health information (risk management), and flagging file formats that are potentially unfriendly to preservation.

After the initial review of the data package, the Repository Content Analysts will conduct any remediation or normalization specified by the Data Management Consultants, carrying out some of the tasks stipulated as Level Two. In particular, the Content Analysts will carry out any necessary file format transformations, file renaming or restructuring, and provide further contextualization of the data where appropriate (for example, furnishing citations for associated publications). All curation and processing steps are then recorded within a curation log.

Since the data curation program's inception in early 2017, the curation team has processed 19 datasets, with around 3 hours of work logged for each dataset and an average of approximately 6 days between submission and publication. At our current deposit velocity, Duke sees room to scale, and engaging in this exercise was particularly helpful in evaluating potential gaps in our current service while planning for future growth. Moreover, this exercise has been particularly useful as we move to a new software application in support of research data by informing both our evaluation of new systems and our specification of desired features. As an example, the inclusion of versioning as a Level One curation task has helped us make the case that any proposed software solution should be able to accommodate versioning without extensive customization being required.

Transparency and Communication

Implementing an effective data curation program relies upon communicating available services to a variety of stakeholders. Going through this exercise helped staff at both UNC and Duke to consider and implement strategies to clearly communicate and make transparent curation activities provided, why certain tasks are important, and what value they add. Communication strategies might include online documentation, targeted presentations to stakeholders, and integration of curation program details into broader RDM education initiatives. Internally, this can help us explain to library staff the services offered, how those services relate to the library's overall mission, and enable staff to cogently answer patron questions regarding curation services. Communicating the availability and value of services to the greater university community (i.e., researchers, grant managers, and others involved in supporting the faculty and graduate student research enterprise) can facilitate these services being written into formal data management plans and increase awareness of how formal data curation can help researchers meet the FAIR guiding principles and comply with growing journal data sharing policies

Measuring Success

This exercise also led our group to talk about how we should be measuring our own success. Some metrics for success could include higher numbers of deposits over time, positive researcher feedback, and download analytics as a proxy for the reusability of data. Ultimately, success will involve developing a data curation program that is both scalable as the number of deposits (hopefully) increase and flexible to respond to any growing and changing needs of the research communities that we serve. Research on what data curation services researchers value will also provide important information as we consider what services to provide and how we communicate and market our services to the broader community (Johnston et al., 2018).

Conclusion

Both institutions found this exercise to be helpful for planning our local services and communicating with internal and external stakeholders. Through this exercise we were able to provide specificity around the language and activities of data curation and to identify and document our goals around making research data FAIR and meeting researchers' needs. We hope to improve the transparency and support of our work by articulating the type of expertise needed by our staff to support these goals. Further work is needed around documenting the

baseline capabilities required by our systems to support FAIR data and these levels will help us as we develop roadmaps for future development.

We expect that the curation activity placement within the three levels may shift as we learn more through delivering these services over time and as more research about the value of certain curatorial activities becomes available. We also hope this exercise will be a useful point of reference for other libraries as they consider how to scale up and communicate their data curation programs and we invite conversation and feedback on the results of this exercise.

References

Consultative Committee for Space Data Systems. (2012). Reference model for an open archival information system (OAIS) (Magenta Book No. 650.0-M-2). Washington, D.C.: National Aeronautics Space Agency. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Digital Research Faculty Working Group. (n.d.). Retrieved from: https://research.duke.edu/digital-research-faculty-working-group

Fearon, Jr., D., Gunia, B., Pralle, B., Lake, S., & Sallans, A. (2013). SPEC Kit 334: Research Data Management Services (July 2013). Association of Research Libraries. https://doi.org/10.29242/spec.334

Hudson-Vitale, C., Imker, H., Johnston, L., Carlson, J., Kozlowski, W., Olendorf, R., & Stewart, C. (2017). SPEC Kit 354: Data Curation (May 2017). Association of Research Libraries. https://doi.org/10.29242/spec.354

Johnston, Lisa R; Carlson, Jake; Hudson-Vitale, Cynthia; Imker, Heidi; Kozlowski, Wendy; Olendorf, Robert; Stewart, Claire. (2016). Definitions of Data Curation Activities used by the Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/188638.

Johnston, L. R. (Ed.). (2017). Curating research data: A handbook of current practice (Vol. 2). Association of College & Research Libraries. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988633_crd_v2_OA.pdf

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, 6(1), 2198. https://doi.org/10.7710/2162-3309.2198

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. PLOS ONE, 12(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Mock, G. (2010, March 21). Faculty move forward on Open Access policy [Blog post]. Retrieved from: https://today.duke.edu/2010/03/accessvote.html

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. Science, 348(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Odum Institute for Research in Social Science. (n.d.). "Data Archive." Retrieved from https://odum.unc.edu/archive/

Raboin, R., Reznik-Zellen, R., & Salo, D. (2012). Forging New Service Paths: Institutional Approaches to Providing Research Data Management Services. Journal of EScience Librarianship, 1(3). https://doi.org/10.7191/jeslib.2012.1021

TRLN Institute 2017. (2017). Retrieved from: https://sites.google.com/site/trlninstitue2017/home

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 3, 160018. Available at: https://doi.org/10.1038/sdata.2016.18
[Screen a: How Sharing Supports RQR]

Appendix

Quality assurance as defined by DCN includes many tasks from reviewing the documentation and metadata for completeness to validating, cleaning, and enhancing data to performing variable by variable checks to ensure all codes are properly labelled. In some cases (and in the context of the levels defined above), quality assurance may in and of itself have different "levels", which require varying levels of expertise, software, and resources. Therefore, to fully define repository service levels it is also important to further unpack quality assurance. For the levels defined above, we see two primary "levels" for quality assurance that fall within both Level 2 and Level 3 Curation.

Quality Assurance (Level 2): Ensuring that documentation and metadata are provided. Performing cursory reviews to identify errors such as lack of definitions for variables, missing codes, other potential issues that are visible during a general review of the data. This level of QA would not involve an in-depth variable by variable check, comprehensive reviews of null/blank values, or any data cleaning activities.

Quality Assurance (Level 3): Ensuring that documentation and metadata are comprehensive and complete. Perform a comprehensive review of all data files for missing labels/codes, issues with null values, out-of-range codes, etc. This level of QA would require more domain knowledge and might also include cleaning or enhancement of the data/documentation files.

We have also identified two levels for risk management:

Risk management (Level 2): Perform a cursory review for confidentiality risks inherent to human subjects data or sensitive information. This would only include a general review for direct identifiers or variables/datasets that noticeably raise questions about the legality of sharing the data. This level of risk management review would not necessarily identify potential risks of disclosure that might arise from the inclusion of indirect identifiers and would not include remediation through de-identification services.

Risk management (Level 3): A complete review for confidentiality risks inherent to human subjects data or sensitive information. This would include a variable by variable level assessment and identification of risks based on deductive disclosure. This level of review would require in-depth expertise in disclosure risks and de-identification procedures and would potentially involve remediation through de-identification services.