JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 8, General Issue (2020)

Conceptualizing Data Curation Activities Within Two Academic Libraries

Sophia Lafferty-Hess, Julie Rudder, Moira Downey, Susan Ivey, Jennifer Darragh, Rebekah Kati

Lafferty-Hess, S., Rudder, J., Downey, M., Ivey, S., Darragh, J., & Kati, R. (2020). Conceptualizing Data Curation Activities Within Two Academic Libraries . *Journal of Librarianship and Scholarly Communication,* 8(General Issue), eP2347. https://doi.org/10.7710/2162-3309.2347

This article underwent semi-anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.7710/2162-3309.2347

PRACTICE

Conceptualizing Data Curation Activities Within Two Academic Libraries

Sophia Lafferty-Hess
Senior Research Data Management Consultant, Duke University

Julie Rudder
Head, Repository Services Department, University of North Carolina at Chapel Hill

Moira Downey
Repository Services Analyst, Duke University

Susan Ivey
Research Data & Infrastructure Librarian, North Carolina State University

Jennifer Darragh
Senior Research Data Management Consultant, Duke University

Rebekah Kati
Institutional Repository Librarian, University of North Carolina at Chapel Hill

INTRODUCTION As funders and journals increasingly create policies that require effective data management and data sharing, many institutions have developed research data management (RDM) programs to help researchers meet these mandates. While there is not a standard set of services for these RDM programs, some institutions, particularly those with repositories that accept data deposits, provide data curation services as a way to add value to research data and help make data more accessible and reusable. Stakeholder communities within the field, such as the Data Curation Network (DCN), are also developing guidelines, procedures, and best practices to support and expand data curation practices. **DESCRIPTION OF PROJECT** This paper examines the data curation activities defined by the DCN, and describes an activity undertaken by library staff at Duke University and the University of North Carolina at Chapel Hill to create a structured model of these tasks to more easily conceptualize and communicate data curation within these two institutional settings. The purpose of this paper is to describe how this model provided a basis for the implementation and expansion of data curation services at each institution and concludes with overall lessons learned. **NEXT STEPS** As we develop our services, libraries have an opportunity to make the often-invisible work of curation more transparent. This paper aims to provide a point of reference for other libraries as they consider how to scale up their data curation programs as well as contribute to discussions around prioritization of services, program assessment, and communication with stakeholders.

Received: 11/06/2019 Accepted: 05/08/2020

Correspondence: Sophia Lafferty-Hess, Box 104732, Durham, NC 27708 sophia.lafferty.hess@duke.edu

INTRODUCTION AND LITERATURE REVIEW

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 enact policies that require researchers to effectively manage their data and share that data openly (NSF, 2011; PLOS, 2014), academic libraries have been developing research data management (RDM) programs to support researchers' needs (Fearon, Gunia, Pralle, Lake & Sallans, 2013; Raboin, Reznik-Zellen, & Salo, 2012). One aspect of RDM programs may include building and supporting institutional repository systems for the stewardship and dissemination of research data. Some libraries are also including data curation services as an element of their repository programs in order to add value to research data and help make data more accessible and reusable.

In 2017, the 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 specifically 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 (DCN), 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 been examined through case studies and interviews with institutional repository staff, and efforts to establish standards around performing data curation, assessing data, and certifying "trusted" data repositories provide a solid foundation to engage in meaningful dialogues in this space (Johnston, 2017; Johnston et al., 2018; Lee & Stvilia, 2017; Lin et. al, 2019; Peer, 2014; Wilkinson et al., 2016).

Over the past two years, the Libraries at Duke University and the University Libraries at the University of North Carolina at Chapel Hill (UNC-CH) have engaged in conversations around the formulation and expansion of our repository and data curation programs. These conversations began at a Triangle Research Libraries Network (TRLN) Institute in the summer of 2017 which focused on "Supporting New Directions and Projects in Scholarly Communications." During the Institute, library repository and RDM staff used the data curation activities outlined by the DCN as a starting point to discuss data curation services within our own institutions. The goals of this "thought exercise" were 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. This paper discusses the outputs of this "thought exercise," implementation and expansion of services within each institutional context, and lessons learned. This paper updates and expands upon a white paper that initially discussed the outcomes of the TRLN "thought exercise" and was published in the LIS Scholarship Archive Works in 2018 (Lafferty-Hess et al., 2018)).

Two Repository Settings

University of North Carolina at Chapel Hill (UNC-CH)

Since 2010, the Carolina Digital Repository (CDR)[1] at UNC-CH has offered a place for individual researchers to host and preserve up to 2 TBs of data. The CDR is an institutional repository that supports preservation and access for multiple content types, which include data, scholarly articles, and student papers. The repository program is currently staffed with three repository developers, two repository librarians, two content and metadata focused staff, and one metadata librarian. These staff are responsible for the systems, content, and service of the CDR and the digital collections repository, 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.

Duke University

In the fall of 2015, Duke convened a Digital Research Faculty Working Group that included a number of campus faculty and 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 data output by faculty and researchers and

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

recommended four new library staff positions to conduct this work—two Senior Research Data Management Consultants and two Digital Repository Content Analysts—as well as resources for expanded infrastructure. In 2017, staff began work to create a suite of data curation services, including policies and procedures, while simultaneously rethinking the software infrastructure required. In concert with the Content Analysts, the Data Management Consultants established a pre-publication workflow for ensuring the quality of submitted datasets in the spring of 2017. The team then collaborated with local software developers to envision and implement a new system[2] that, like UNC-CH's CDR, uses Samvera's Hyrax framework.

Engaging with the DCN curation activities and the ARL SPEC Kit through the lens of our own institutional contexts led us to envision a three-tiered model for characterizing research data curation efforts. A shared understanding of what specific curation activities entailed, as well as of the human and technical capacity available at each institution, was instrumental in guiding our thinking. Below, we will examine in more detail the conceptualization, implementation, and growth of data curation services at Duke and UNC-CH.

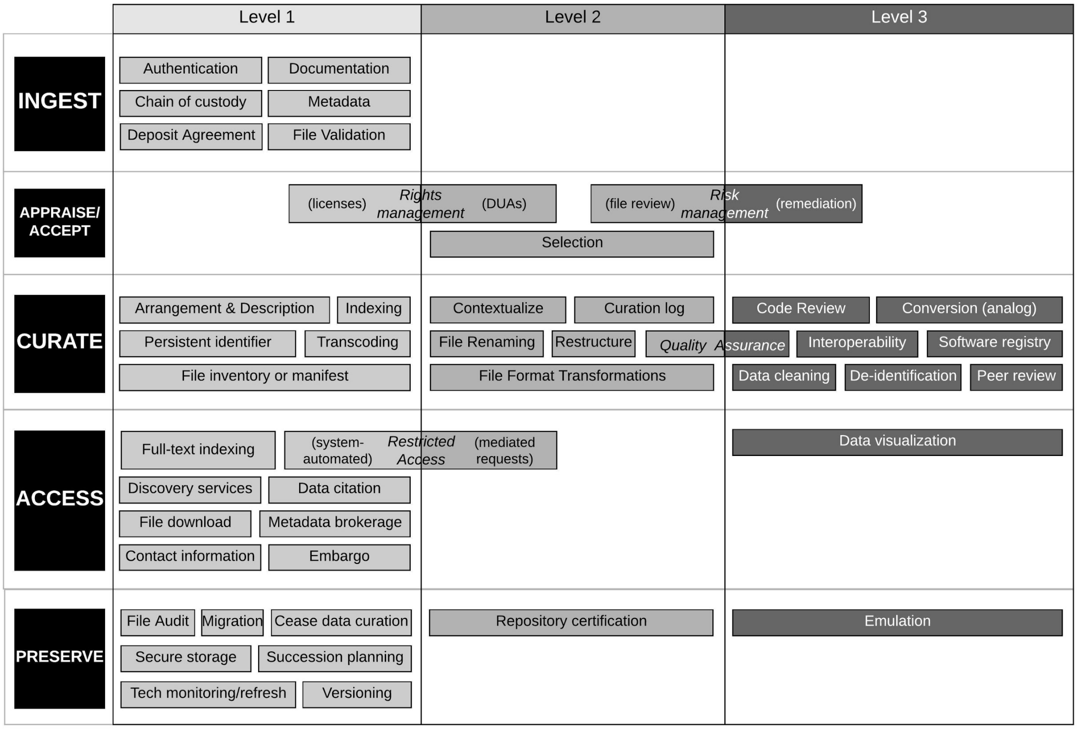
CONCEPTUALIZING DATA CURATION

The 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), which are increasingly being cited by funders and journals. Another goal for a data curation program within academic libraries is to meet the specific needs of researchers, which 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, a team of library staff from UNC-CH and Duke 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 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. In the team's conceptualization, curation involves both tasks performed by systems and those performed by human capital (to

[2] The Duke Research Data Repository, https://research.repository.duke.edu

varying degrees of human involvement). 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.



	Level 1	Level 2	Level 3
INGEST	Authentication Documentation Chain of custody Metadata Deposit Agreement File Validation		
APPRAISE/ ACCEPT	(licenses) *Rights management* (DUAs)	(file review) *Risk management* (remediation) Selection	
CURATE	Arrangement & Description Indexing Persistent identifier Transcoding File inventory or manifest	Contextualize Curation log File Renaming Restructure *Quality Assurance* File Format Transformations	Code Review Conversion (analog) Interoperability Software registry Data cleaning De-identification Peer review
ACCESS	Full-text indexing (system-automated) *Restricted Access* (mediated requests) Discovery services Data citation File download Metadata brokerage Contact information Embargo		Data visualization
PRESERVE	File Audit Migration Cease data curation Secure storage Succession planning Tech monitoring/refresh Versioning	Repository certification	Emulation

Table 1. Levels of Data Curation Activities

Note: Levels include all activities from the lower levels (i.e., levels two and three include activities from preceding levels). For full definitions of data curation activities see the Data Curation Network: Data Curation Terms and Activities.

Level One curation focuses on a repository program that facilitates self-deposit or lightly-mediated deposit, and is generally supported by the system or repository policies. In this level, the repository system will ideally provide many of the functionalities to support the provenance and ingest of data through standard agreements, discovery of data through metadata, indexing, and arrangement, reuse and access of data through licenses, data citations, terms of use, and a preservation environment to ensure the long-term availability of the data. However, the extent of hands-on enhancement or improvements of a dataset would be limited to minor adjustments to metadata, documentation, arrangement, or mediation of the data package in certain identified cases, such as with large file transfers.

Level Two curation builds on those tasks outlined in Level One by providing a more thorough review and potential enhancement of the data package. Level Two services are conducted by library staff with general knowledge of data management and curation best practices. At this level, staff may provide a general quality assurance of the data package, 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 protected health information (risk management), flagging and/or transforming file formats that are potentially unfriendly to preservation, renaming or restructuring files, and recording any changes within a curation log.

Level Three curation involves hands-on manipulation of datasets and specialized repository functionalities. Level Three services also often require human intervention by staff with both general knowledge of data best practices and even more specialized domain-specific and data type knowledge. In this level, staff perform more in-depth quality assurance, clean or de-identify data, run and troubleshoot code files, or provide enhanced systems for emulation or data visualization.

Some 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. 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 Two) to a more in-depth comprehensive review that, for example, leverages domain-specific expertise (Level Three). Similarly, rights management within Level One would involve facilitating data depositors to assign a license to the data package, whereas in Level Two 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.

DATA CURATION IN PRACTICE: UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Target Service Level

UNC-CH Libraries targeted Level One for our institutional repository (CDR) data curation service. Our goal was to strike the right balance and to define the specific value that we add to data deposits. We considered our current staffing levels, rate of data deposit, and other data services at the library and on campus. Although the UNC-CH Libraries offer research data management consultation and an array of other data-related services, we currently do not have any staff positions dedicated to data curation for the CDR as either generalists or with a specific domain focus. However, other UNC-CH campus entities offer local venues for research data management and various data services, and we wanted to focus on services which differentiated the CDR. Additionally, for data deposited in the CDR, we wanted to offer some data curation services, and we wanted those services to be the most useful activities that we could achieve with our staffing levels and generalist expertise. Therefore, we determined that we could add a few activities from Level Two as well, but do not feel that we can fully scale up to Level Two without additional data and subject expertise.

Implementing the Service Level

Out-of-the-box support for most Level One activities was a key reason for our selection of Hyrax as the repository platform for the CDR. Additional activities such as metadata, rights management, and contact information were handled through customization of the self-submission form. For example, we selected metadata fields appropriate for dataset submission, including methods and type of data. We also added the option for depositors to choose a license and rights statement for their data. Beyond these system-level decisions, our repository policies and practices were modified to accommodate additional Level One curation activities. As part of the CDR's preservation policy, we created succession plans in the event that the university ceases support for the repository. Our repository developers and systems staff monitor software, hardware, and infrastructure improvements.

In addition to the curation effected through system design and policy implementation, the Institutional Repository Librarian now reviews data deposits upon submission. The goal of the review process is to provide suggestions and guidance for depositors to comply with FAIR data principles. Specifically, the review process consists of the below elements. The data curation activities to which they correspond appear in parentheses:

- Scan for personally identifiable and sensitive information according to UNC-CH's definitions (Risk Management)
- File format assessment to determine whether the data are in an open format (first step toward the File Format Transformations activity)
- License assessment (Rights Management)
- Digital Object Identifier (DOI) creation (Persistent ID)
- Locate ReadMe file or other explanatory documentation such as a codebook (Documentation).The ReadMe file mandates that depositors provide contact information in the form of names and email addresses (Contact Information).

- Contact the depositor to provide suggestions to improve the description and reuse of the dataset.

After we implemented our additional data curation services, we also began to explore whether or not library staff with data specialities and discipline-specific knowledge may be able to enhance our services even further. We compiled an expert list and matched staff to the datasets already in the CDR. Staff used the Data Curation Network's CURATE checklist (Johnston et al., 2018) to evaluate assigned datasets. We found that staff were able to make valuable edits to descriptive metadata, check for deposit completeness, and suggest valuable changes, but as suspected, time was a significant factor and adding these additional activities to daily work would be challenging to deliver in short time frames. At this time, we will use this expert list as a resource if the Institutional Repository Librarian has questions about a submitted dataset. We have also concluded that we would need to develop a UNC-CH specific dataset checklist, if we were going to grow this area of work with staff outside the CDR, perhaps having fewer steps, eliminating those steps which are automated through technology, and provide more specific directions.

DATA CURATION IN PRACTICE: DUKE UNIVERSITY

Target Service Level

At Duke, our curation workflow and the policies to support the workflow were heavily influenced by the capacity created through the staffing model recommended by the Digital Research Faculty Working Group. The four new staff members were familiar with general data curation and management best practices, but did not necessarily possess the domain knowledge required for more advanced curatorial activities for all disciplines served by the program. Given this generalist disposition, the curation team settled on a group of activities that were broadly consistent with Level Two, acknowledging that a number of curation tasks were out of scope for the present iteration of the program. While many of these out-of-scope tasks were ultimately banded together as a third level of data curation service, several Level Two tasks, such as mediated access to restricted data, currently remain outside both the human and system capacity of our program at Duke.

Implementing the Service Level

After the Institute in 2017, the data team used the DCN activities and our defined levels to identify the gaps in our curation services, focusing particularly on what functionalities we would desire from a new system. Software development staff had been given the time and mandate to develop a new, greenfield repository application dedicated to publishing and preserving research data using the Samvera community's Hyrax platform. Data curation staff were able to work with developers to implement a range of desired new features; as an example, the inclusion of versioning as a Level One curation task helped make the case that the software should be able to accommodate dataset versioning, even if it meant local software customization. Several other areas of improvement were focused around usage analytics, more robust metadata (including contact information), and embargo functionality. A few tasks that were previously carried out manually are now system-assisted, including DOI minting and requesting updates to published data.

Prior to publication, each dataset submitted to Duke's repository undergoes review by a staff member with expertise in best practices around research data. The curators perform the following steps for all datasets:

- Inspect the data for a variety of indicators that the data may be sensitive (i.e., protected health information, personally identifiable information, etc.) (Risk Management)
- Evaluate the file formats for enhancements for preservation and portability (File Format Transformations)
- Evaluate the dataset's accompanying documentation (Documentation)
- Evaluate the dataset for potential issues such as appropriate variable and value definitions, out of range values, descriptions of programs used for code files, and preferred data structures (Quality Assurance)
- Discuss potential enhancements with the depositor (if needed)
- Arrange files within the repository system to enhance access and reuse (Restructure)
- Standardize the depositor-submitted metadata using Dublin Core properties and normalizing various descriptive elements against a handful of controlled vocabularies (Metadata)
- Calculate checksums against the dataset's files to ensure bit-level integrity (File Audit)
- Register a DOI and construct a bibliographic citation for the data (Persistent Identifier and Data Citation)
- Preserve any changes to the dataset within a plain text curation log (Curation Log)
- Retain the initial submission information package for provenance purposes (Chain of Custody)

Another addition to Duke's data curation program in the Spring of 2018 was joining the DCN as a formal partner. Joining the DCN has allowed the Duke data curation team to provide more standardized Level Three service options for datasets entering our repository by harnessing more disciplinary and data type-specific knowledge possessed by our DCN partners. DCN has also provided an expanded community for discussing what different levels or "flavors" of data curation looks like across our institutions and a framework for discussing data activities within the CURATED model (Johnston et al., 2018).

LESSONS LEARNED

Assessing and Building Capacity

For any institution seeking to support research data curation, choices are necessary. Decisions around the level of support are highly localized and depend heavily on staffing models and institutional backing. In addition, it is helpful to realize that there is not a one-size-fits-all set of services and that not every institution can move beyond Level One. Even with limited staffing capacity, institutions with a self-deposit system can realize some gains and help researchers and data producers better align with FAIR principles by using the Level One activities as a guide to shaping policies and workflow, as well as a gauge for system requirements and software development. By grouping the data curation activities into levels, we were able to take a realistic look at the activities we currently support and determine the activities that we could add with our current level of staffing and development resources. We were also able to identify and begin to address existing gaps in our services at the level of both staffing and software. While further work is needed around documenting the baseline capabilities required by our systems to support FAIR data, these levels will help us as we allocate local resources and create roadmaps for future software development.

TRANSPARENCY AND COMMUNICATION

UNC-CH and Duke both found that this exercise facilitated communication and transparency within and between our institutions. Each institution considered and implemented strategies to communicate and make transparent the curation activities provided, why certain tasks are important, and what value they add to a diverse set of stakeholders. At UNC-CH, we found that the data curation levels helped us find a common language to communicate with library staff. Prior to introducing the data curation levels, it was difficult to know whether we were using similar definitions or what assumptions library staff had about our services. We needed to move beyond saying that we do "data curation" and speak in more concrete terms. The levels and curation activities provided a common language and definition in our discussions. In addition, the levels helped us talk about the types of staff we would need to support activities in Level Two and Level Three and helped us evaluate future investment in increased curation support.

At Duke, clarity around what "data curation" means within the context of library services helped us to communicate with broader campus partners and with library stakeholders regarding the value-proposition of our repository and curation services. Explaining what we would do regarding reviewing datasets for quality assurance, documentation, risk management, and file format transformations (in line with Level Two) provided concrete examples of how our model facilitated not only making data accessible but reusable. We found that while "curation" may still be an obscure term to a broader University audience, being specific about the activities performed helped to create more transparency around our work.

Inter-institutional communication was also an important outcome of this exercise. Having candid and forthright ongoing conversations among staff at UNC-CH and Duke helped to create an honest space to engage with the challenges that are common across institutions as well as to learn what worked at each institution. As libraries continue to develop new services in this space, being transparent about the services and infrastructure capabilities within our institutions will help others, including our research communities, make knowledgeable decisions regarding repository options. Transparency with our colleagues across institutions helps move forward growing conversations around how we prioritize data curation activities as we face growing demand and limited resources.

MEASURING SUCCESS AND FUTURE DIRECTIONS

Measuring the success of a data curation program remains an often difficult task that is not yet well defined. The space around data curation is a fluid and rapidly evolving one. Quantitative metrics, such as rate of deposit, tell one side of the story but are not necessarily indicative of the health or worth of such a program. As with other areas of the scholarly communications ecosystem, focus on the number of datasets deposited potentially elides the ancillary benefits of engaging in data curation. In discussing the broader impact of implementing open access policies, Mitchell (2016) highlights some of these ancillary benefits, noting that such policies "regardless of their compliance rates, create an opportunity to collect stories that give a voice to those communities of readers who don't yet have access. And these stories, in turn, help naturalize the idea of openness by increasing awareness among authors who may not realize the missed opportunities they have (and the rights they possess) to reach new communities of readers." While "naturalizing" the notion of openly available, curated data is an important outcome of implementing a data curation program, it is also one that is exceptionally difficult to quantify.

Numbers cannot be entirely disregarded, and indicators like deposit velocity should be seen as one way to assess success. Both UNC-CH and Duke have attempted to gauge the success of our efforts through a mixture of increased numbers of deposits over time, positive researcher feedback, and download analytics as a proxy for the reusability of data. Ultimately, success will involve continuing to develop a data curation program that is both sufficiently scalable to accommodate an increasing number of deposits 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, how to tailor those services to meet specific needs, and how we communicate and market our services to the broader community (Johnston et al., 2018; Llebot and Van Tuyl, 2019).

As noted above, one essential metric for the success of a curation program is the extent to which the data it curates can be considered FAIR. Frameworks for assessing the FAIRness of individual datasets and the broader infrastructure are rapidly emerging. While the FAIR data principles are increasingly cited as a baseline for data quality, no commonly held set of core assessment criteria yet exists to facilitate cross-dataset FAIRness comparisons or allow for dataset benchmarking. To address this, the Research Data Alliance's FAIR Data Maturity Model Working Group has set out to establish a generic (domain-agnostic) and expandable self-assessment model for measuring the FAIR maturity level of a dataset ("FAIR Data Maturity Model WG Case Statement", n. d.). With regard to assessing the supporting infrastructure, Lin, et al. (2019) have advanced a group of principles (Transparency, Responsibility, User Community, Sustainability, and Technology, known as "TRUST") intended to guide development of repository applications that support FAIR data. Though these TRUST principles remain nascent, momentum is building around establishing an instrument with which to confidently assess the FAIRness of a given dataset outside the context of individual institutions.

CONCLUSION

This exercise facilitated UNC-CH and Duke staff to plan local data curation services, articulate the types of expertise needed, communicate with internal and external stakeholders, and identify gaps for future development. It also helped us prioritize activities given existing resources and communicate those priorities effectively. Finally, this exercise highlighted how cooperative models can potentially help to address knowledge gaps either through cross-institutional collaborations like the DCN or through models that engage subject liaison librarians locally. 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.

Academic libraries are currently positioned to enhance the value of the data produced by our communities and make the data more FAIR compliant. In the ever evolving space of repository options, information professionals working within institutionally based repositories have an opportunity to provide various technological and human-driven curatorial services. As we continue to make advancements in this space, as a community it is important to provide specificity around the language and activities of data curation we provide in order to make the often invisible work of curation more transparent. We 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.

ACKNOWLEDGEMENTS

The authors would like to thank Lisa Johnston for her comments on this paper, as well as thank the entire Data Curation Network team for their work on developing the list of data curation activities that provided the foundation for the model explored in this paper.

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, 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. (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; Carlson, J.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R.; Stewart, C.; Blake, M.; Herndon, J.; McGeary, T.; Hull, E. (2018). Data curation network: A cross-institutional staffing model for curating research data. *International Journal of Digital Curation*, *13*(1), 125–140. https://doi.org/10.2218/ijdc.v13i1.616

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

Lafferty-Hess, S., Rudder, J., Downey, M., Ivey, S., & Darragh, J. (2018) Conceptualizing data curation activities within two academic libraries. *LIS Scholarship Archive.* https://doi.org/10.31229/osf.io/zj5pq

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

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Hugo, W. and Mokrane, M. (2019) The TRUST principles white paper (version 0.01). Retrieved from https://docs.google.com/document/d/1UCsdnz0wk9TeMj1Dqxi8wuZ2Lu_TNVkpJ2TX48yKsec/edit

Llebot, C. & Van Tuyl, S. (2019, May 16). Peer review of research data submissions study. Presentation at the annual meeting of the Research Data Access and Preservation Association, Coral Gables, Florida. Retrieved from osf.io/uyq6b

Mitchell, C. (2016, October 26). Does the UC open access policy miss the mark? Depends on which mark [Blog post]. Retrieved from https://osc.universityofcalifornia.edu/2016/10/does-the-uc-open-access-policy-miss-the-mark-depends-on-which-mark/

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

National Science Foundation. (2011). Dissemination and sharing of research results. Retrieved from: https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestable, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys M.,…Yarkoni, T. (2015). Promoting an open research culture. Science, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Peer, L., Green, A. and Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). 263–291. https://doi.org/10.2218/ijdc.v9i1.317

PLOS. (2014). Data availability policy. Retrieved from: https://journals.plos.org/plosone/s/data-availability

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

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P.E., Bouwman, J., Brookes, A. J, Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, F.,… Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

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 Two and Level Three Curation.

Quality Assurance (Level Two): 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 Three): 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 Two): 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 Three): 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.