Data Curation Network

A Cross-Institutional Staffing Model for Curating Research Data

July 2017

The Data Curation Network project is supported by a grant from the ALFRED P. SLOAN FOUNDATION.

Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data

Release date: 7-27-2017
This report represents the primary outcome of the project "Planning the Data Curation Network" funded 2016-2017 by the Alfred P. Sloan Foundation grant G-2016-7044.

Authors: Lisa R Johnston, University of Minnesota
Jake Carlson, University of Michigan
Cynthia Hudson-Vitale, Washington University in St. Louis
Heidi Imker, University of Illinois at Urbana--Champaign
Wendy Kozlowski, Cornell University
Robert Olendorf, Pennsylvania State University
Claire Stewart, University of Minnesota

Recommended Citation: Lisa R Johnston, Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. (2017). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. http://hdl.handle.net/11299/188654.

Table of Contents

Executive Summary

Funders increasingly require that data sets arising from sponsored research must be preserved and shared, and many publishers either require or encourage that data sets accompanying articles are made available through a publicly accessible repository. Additionally, many researchers wish to make their data available regardless of funder requirements both to enhance their impact and also to propel the concept of open science. However, the data curation activities that support these preservation and sharing activities are costly, requiring advanced curation practices training, specific technical competencies, and relevant subject expertise. Few colleges or universities will be able to hire and sustain all of the data curation expertise locally that its researchers will require, and even those with the means to do more will benefit from a collective approach that will allow them to supplement at peak times, access specialized capacity when infrequently-curated types arise, and to stabilize service levels to account for local staff transition, such as during turn-over periods. The **Data Curation Network** (DCN) provides a solution for partners of all sizes to develop or to supplement local curation expertise with the expertise of a resilient, distributed network, and it creates a funding stream to both sustain central services and support expansion of distributed expertise over time. Our model will accelerate local capacity, strengthen collaboration between libraries and disciplinary projects, and significantly enhance libraries' collective voice in conversations about the future of research data.

The Data Curation Network will serve as the "human layer" in a local data repository stack that provides expert services, incentives for collaboration, normalized curation practices, and professional development training for an emerging data curator community. Data curation enables data discovery and retrieval, maintains data quality,

adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation. Data curation requires a specialized skill set that spans a wide variety of data types (e.g., spatial/GIS, tabular, database, etc.) and discipline-specific data formats (e.g., chemical spectra, 3D images, genomic sequence, etc.). The Data Curation Network addresses this need by creating a cross-institutional staffing model that seamlessly connects expert data curators to local datasets.
The Data Curation Network model, and the research findings supporting it, are presented in this report as the primary outcome of the Alfred P. Sloan funded grant titled "Planning the Data Curation Network" that ran from May 2016–June 2017. To implement the Data Curation Network we propose:

1. A staffing and governance model that includes tiers of participation allowing some institutions to join the Network by contributing in-kind data curation staff and others to utilize the Network's curation services as end-users.
2. A submission workflow that fosters strong local connections between researchers and local curators and gives the home institution complete control to decide how to engage Network resources.
3. An implementation plan that incrementally grows the Data Curation Network with new partners and expands curation offerings over time.
4. A financial plan and sustainability model that prevails beyond the grant-supported implementation phase and enables the Data Curation Network partners to train and recruit new data curators.
5. An assessment plan that demonstrates how a networked approach to curating research data is more efficient and scalable, and that data curated by the Data Curation Network are more valuable.

Next, the Data Curation Network will be implemented to accomplish our mission to better support researchers that are faced with a growing number of requirements to ethically share their research data. Our vision for the Data Curation Network is to:

- Develop standards-driven data curation techniques for all types of repository infrastructure.

- Expand into a sustainable entity that grows beyond our initial six partner institutions.
- Demonstrate that datasets curated by the Data Curation Network are used to advance research and education in ways that are measurably of greater reuse value than non-curated data.
- Build an innovative community that enriches capacities for data curation writ large.

1.0 Introduction

Research data have value beyond their original purpose. They can be used to demonstrate findings, enable new discoveries, reproduce and validate results, and are repurposed in surprising new ways that their creator may never have imagined. Yet data, captured in a multitude of digital file formats through an ever-increasing number of techniques, are constantly at risk of falling short of their long-term reuse potential. Data can be messy and incomprehensible. They often lack important documentation, metadata, and other characteristics that might otherwise secure their long-term usefulness. In addition, their fragile, digital shells are not resilient enough to preserve them from format obsolescence and other ill-effects of digital deterioration, such as bit-rot. Finally, the reality for most data created today is that they never leave the local environment in which they were first stored and, therefore, as time goes by, become the victim of benign neglect.

On the other hand, well-curated data are valued by the scholarly communities that produce them. Professionally curated data are easier for fellow scholars and future collaborators to understand, are more likely to be trusted, and the research they represent more likely to be reproducible (Roche, Kruuk, Lanfear, & Binning, 2015; McNutt et al., 2016; Smith & Roberts, 2016; Beagrie & Houghton, 2014). As a consequence, and to counteract their ephemeral and swiftly eroding nature, requirements for digital research data to be managed, shared, and preserved have emerged. Researchers worldwide face emerging mandates and altruistic pressures to share their research data in ways that make them findable, accessible, interoperable, and reusable, or FAIR (Wilkinson et. al, 2016). For example, in the United States, many recipients of federal research funding must address how their research will be "publicly accessible to search, retrieve, and analyze" in a written Data Management Plan appended to their grant

applications (Holdren, 2013). Policies from funders such as the National Science Foundation (https://www.nsf.gov/bfa/dias/policy/dmp.jsp) and the Bill and Melinda Gates Foundation (http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy) serve as examples of these requirements.

But funders are not the only drivers for researchers to share their data. Researchers also face a growing number of publisher expectations to include digital data in the peer review process and share them alongside resulting publications. Journal data sharing policies, such as those held by PLoS ONE (http://journals.plos.org/plosone/s/data-availability) and Nature Publishing Group (http://www.nature.com/authors/policies/availability.html#data), require all data underlying their research results to be made openly available for sharing and reuse (Stodden, 2012). Often, reproducibility is a driving factor for these policies. Some disciplines have embraced the open data movement as a positive development that will foster expanded practices in validation and replication (Munafò et. al, 2017), and may even safeguard against scientific fraud or the dissemination of erroneous results (Fecher, Friesike, Hebing & Linek, 2017).

Academic and research libraries have followed research and scholarly communications trends related to research data with great interest. Libraries are experts in identifying, selecting, organizing, describing, preserving, and providing access to information materials in print and digital formats. And as a critical agency of their parent institutions, academic libraries are persistent, with demonstrated and sustainable models for providing services such as collection management, preservation, and access to a broad variety of information. Librarians and archivists understand the value, and challenges, of creating and preserving information for future generations, and recognize that specialized curatorial actions must be taken to preserve data and other materials for reuse. This curation enables discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation. Thus data curation, the active and ongoing management of data through its lifecycle of interest and usefulness, is central to our mission and has become an important role for academic research libraries as we transform our

workforce to assume greater digital stewardship responsibilities in the academy (National Research Council, 2015). For example, over the last decade institutional repositories (IRs) that were originally launched to support open access to traditional scholarship, such as articles and theses, have risen to the challenge of providing access to the many types of digital data, in a variety of formats, that the overwhelmingly multi-disciplinary institutions generate. Based on well-established archival models, such as standardized OAIS-compliant software architecture (Consultative Committee for Space Data Systems, 2011), IRs provide the technical infrastructure to make digital research data accessible, retrievable, and reliably persistent in all the ways that a trusted digital repository might aspire.

Yet, among the advances in technical aspects underlying a digital repository--those that provide storage, ingest, description, access, and preservation--one challenge looms large: the expertise of a data curator. The curation staff, or the "human layer" in the repository stack, bring the disciplinary knowledge and software expertise necessary for reviewing and curating data deposits to ensure that the data are reusable. Due to the heterogeneous and multidisciplinary nature of research data generated in our academic institutions, the skills and expertise required to curate data (to prepare, arrange, describe, and test data for optimal reuse) cannot be fully automated nor reasonably provided by a few experts siloed at a single institution. Multiple data curation experts are needed to effectively curate the diverse data types an IR typically receives (Bloom et. a., 2016; Johnston, 2014). Yet, given limited resources, it is unrealistic to expect that every academic library can hire a data curator for every data type (e.g., GIS, tabular spreadsheets, statistical survey, video and audio, computer code) or discipline-specific data set (e.g., genomic sequence, chemical spectra, biological image) an IR might encounter. Similarly, each type of data curation expertise might only be utilized intermittently depending on the disciplinary focus at each institution.

The **Data Curation Network (DCN)** addresses the challenge of scaling domain-specific data curation services collaboratively across a network of multiple institutions and digital repositories in order to provide expert data curation services in disciplines and domains beyond what any single institution might offer alone. The planning

phase project called "Planning the Data Curation Network" ran from May 2016–June 2017 with support from the Alfred P. Sloan foundation.[1] The project team for the DCN planning phase brought together research data librarians, data curation experts, and academic library administrators from six academic institutions that each, separately, provide repository and curation services to their campuses: the University of Minnesota, Cornell University, Penn State University, the University of Illinois, the University of Michigan, and Washington University in St. Louis.

Over the course of the year planning phase, our team sought opportunities to broadly present our work and discuss our ideas with colleagues. For example we were featured at several conferences, including the 2016 SHARE Users Meeting sponsored by the Association of Research Libraries and the Center for Open Science held in Charlottesville, NC, the Joint 8th Research Data Alliance Plenary and SciDataCon 2016 conference held in Denver, CO, the 2016 Digital Library Federation Forum held in Milwaukee, WI, the winter 2016 Coalition for Networked Information meeting in Washington, DC, the 2017 International Digital Curation Center conference held in Edinburgh, UK, the 2017 Research Data Access and Preservation summit held in Seattle, WA, the IMLS-Funded Preservation Quality Tool (PresQT) Workshop held in Notre Dame, IN, the 2017 Big 10 Academic Alliance Library conference held in West Lafayette IN, and the 2017 International Association for Social Science Information Services and Technology (IASSIST) meeting held in Lawrence, KS. As a result of these conversations it became clear that although our planning phase work was focused on the needs of US academic research institutions similar to the six represented by the project team, this model would scale to a wider range of organizational make-ups and affiliations such as federal government agencies, international academic institutions, and small- and mid-sized liberal arts colleges. We very much welcome the opportunity to explore these and other avenues for broader interpretation of the DCN model.

We preface our model for implementing the Data Curation Network with a literature review and a summary of the practical lessons

[1] The Data Curation Network Project Planning Grant narrative is available at http://hdl.handle.net/11299/188634.

learned from our interviews with peer service programs and successful collaborative networks. Next a methods section provides a summary of our research activities that informed the DCN model development, including holding focus groups with researchers, running controlled data curation pilots, and surveying the library community around their existing support and plans for future services in these areas. Finally we present our DCN model with a summary of the staffing roles, participation levels, criteria for bringing on new partner institutions, a proposed path to financial sustainability, and an implementation and assessment plan that are grounded in the measurable metrics and observed demand for data curation services across our six planning phase institutions.

2.0 Literature Review

The Data Curation Network model builds on a rich history of well-established collaborative service models in libraries. Not unlike our vast interlibrary loan networks that deliver books, articles, and other library collections across networked libraries, or the collective contributions of catalogers adding unique and specialized MARC records to national and international cataloging databases, or the more recent response to on-demand web-based user needs with the successful implementation of 24-7 library reference chat services, the DCN will build from our common need to provide scaled services in a shared way.

2.1 History of Library Collaborative Staffing

According to Weber (1976), collaborative work between libraries initiated in the later half of the 19th century. One of the first areas of librarianship to be tackled in a networked manner was indexing and cataloging. By coming together to standardize cataloging of materials, libraries of this period felt that expertise across institutions could be leveraged to create higher quality records.

The end of the nineteenth century also saw an increasing interest in libraries lending materials to one another. By lending materials libraries could meet, according to a quote in the 1898 *The Library Journal*, "the growing demands of scholars, incapable of satisfaction by any one library, and the economical management of library finances." (Stuart-Stubbs, 1975). Interlibrary loan grew out of this grassroots and informal movement into a network of material exchange that has been successful throughout the United States and abroad.

Collective collection development was also identified in the mid to late

nineteenth century as a necessity for libraries, and as a mechanism to fill patron needs and address limited budgets and limited space. This solution has come to encompass projects focused on centralized infrastructure for all types of collections and stewardship responsibilities, including digital services, print storage, preservation, and discovery, among others. For initiatives specifically around collective digital preservation and digital collection development and discovery, community focused solutions have helped solve collective issues. "LOCKSS and HathiTrust represent community-sourced solutions that have enabled academic libraries to externalize stewardship functions that were previously organized locally at a much higher cost," (Dempsey, Malpas, & Lavoie, 2014, p 30). Similar projects that have been built as community-supported solutions to digital collections include the Digital Public Library of America, the Digital Preservation Network, and Duraspace.

Recently there has been momentum around managing shared print materials, or the ability to collectively share the management and preservation of print literature while decreasing local holdings. Dempsey et. al. (2014) find that the development of, "shared print management schemes represent a cost-effective alternative to institution-scale solutions, redistributing the costs of library stewardship across a broader pool of participants," (p30). Rather than this initiative being driven by community-focused solutions like those mentioned above, Dempsey et. al. find that consortia are playing a strong role in organizing libraries for shared print services based on geography. Examples of this include the 2CUL project, CIC Cooperative Cataloging Partnership, the Association of Southeastern Research Libraries, the Committee on Institutional Cooperation, the Statewide California Electronic Library Consortium, and the Western Regional Storage Trust.

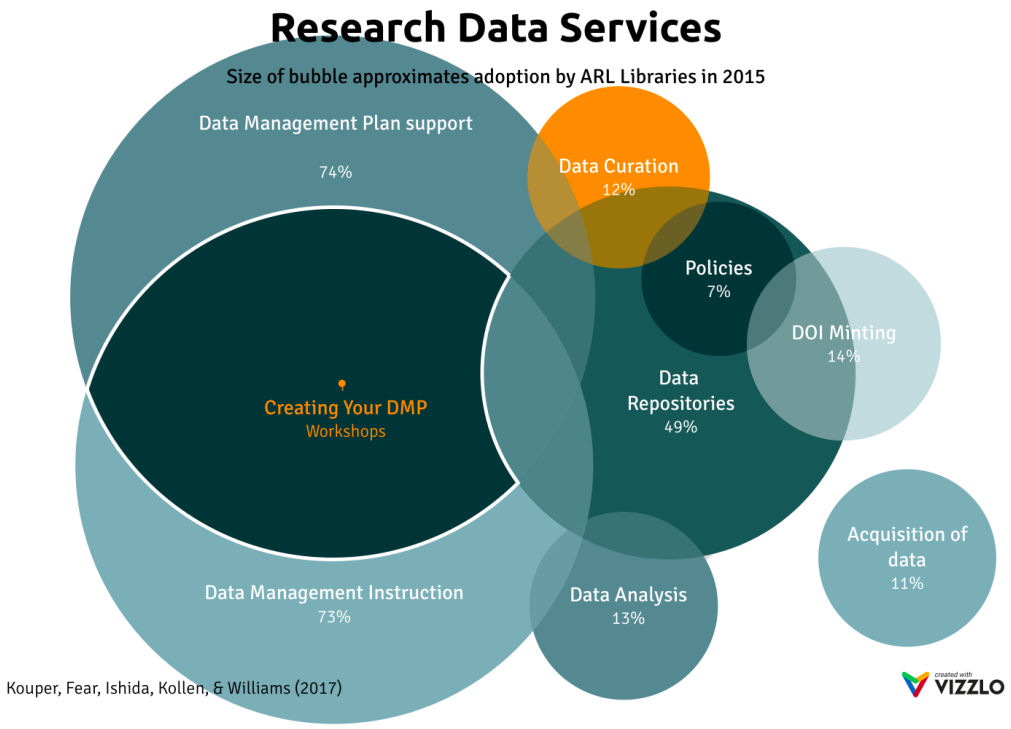
In a similar vein, the appeal for a network of expertise model for delivering unique library services has been expressed through recent research (Kirchner et al., 2015). The authors recommend "...a pilot project in which experts at multiple institutions consciously create a shared approach to address specialized information needs or to solve a common problem" (p17). Additionally, Erway (2012) calls for a collaborative expert network for handling the variety of born digital

media handled in the nation's libraries.

2.2 Current Support for Data Curation in Academic Libraries

Data curation is a subset of a broader suite of research data management services (figure 1). A number of studies and surveys have explored the extent of research data services provided by academic libraries and found that support for research data management, including data curation, has increased steadily over time (Soehner, Steeves, & Ward, 2010; Tenopir, et al., 2011; Tenopir et al., 2015). More recent explorations by Lee & Stvilia (2017) found that support for data curation in libraries is mainly built upon existing and local IRs. IRs only account for a small percent of the data repositories available to researchers, while discipline-specific data repositories (e.g., ICPSR, GenBank) and general-purpose repositories for data (e.g., FigShare, Zenodo) are enjoying growing use (Kindling et al., 2017).

Figure 1: Data curation as a component of research data services



Collaborative projects related to research data management (though not specifically focused on data curation services) are also underway. The Research Data Alliance (RDA, https://www.rd-alliance.org) launched as a community-driven international organization in 2013 and its special interest groups provide a venue for developing and establishing standards for data curation, such as those by the Publishing Data Workflows group (Bloom et al, 2015) and the newly formed Assessment of Data Fitness for Use working group. The Stewardship Gap project is an 18-month Sloan funded project (http://www.colorado.edu/ibs/cupc/stewardship_gap) that reports gaps in how sponsored research data is preserved for future generations (York, Gutmann, & Berman, 2016). Educational preparation for data curation services, like the DigCCuRR Professional Institute (https://ils.unc.edu/digccurr/institute.html) and the CLIR data curation post-doctoral fellowship program (https://www.clir.org/fellowships/postdoc), as well as information sharing networks such as the Digital Liberal Arts Exchange (https://dlaexchange.wordpress.com) and the DataQ Project (http://researchdataq.org) and leading the way in training data curators on relevant best practices in the field as well as providing valuable forums for community building and networking. Finally, academic research library initiatives focused around data curation issues provide a platform for peer groups to share experiences and best practice. Groups such as the SHARE Curation Associates program (http://www.share-research.org/about/our-team/associates-program), which is highly focused on computational-thinking competencies and technical skill development of repository staff in the United States, the UK-based JISC Research Data Shared Service Project (https://www.jisc.ac.uk/rd/projects/research-data-shared-service), which seeks to build shared software and repository infrastructure for higher education institutions in the UK, and other emerging collaborative efforts such as the Curating for Reproducibility Consortium (http://cure.web.unc.edu) combine staff and best practices for furthing data curation service offerings in libraries.

3.0 Methodology

Our cross-institutional team held regular discussions via bi-weekly conference calls and two in-person meetings over the course of the one year planning grant to develop the Data Curation Network. The project kicked off with a two-day meeting held June 30-July 1, 2016 in Minneapolis, Minnesota, and facilitated by Santiago Fernandez-Gimenez, a team collaboration expert based at the University of Minnesota. Key outcomes from this meeting allowed us to focus our efforts around a shared vision for the DCN in the next 3-5 years, acknowledge the potential barriers, and prioritize strategic directions for our team in the planning phase.[2] Discussed in greater detail in this section, the following research activities were performed to inform and develop the DCN:

1. **Baseline assessment** of the local services, staffing, and repository technologies in place for data curation at each of the six planning phase institutions.
2. **Researcher engagement** to evaluate the importance of 35 different data curation activities for academic researchers, and to understand their current habits and needs.
3. **Data curation pilots** with staff to identify local practice and identify potential implementation issues and activities, including a normalization of curation process.
4. **Engagement with the research library community** to understand levels of current support for data curation services.
5. **Financial model** development to sustain the DCN post-grant based on various cost models and existing networks.
6. **Metrics tracking** of the ongoing demand (and response) for data curation services across our six institutions over the one-year planning phase.

[2] The summer 2016 DCN Meeting outputs, including our Vision, Barriers, Metrics, and our Strategic Directions, are available at http://hdl.handle.net/11299/188637.

3.1 Baseline Assessment of Local Curation Services

To understand the existing levels of support for data curation across our six planning institutions, we ran a baseline assessment in May 2016 that compared our services, repository technologies, local policies, and staffing and organizational structures.[3] Our results indicated a strong alignment: we all provide data curation services that were aimed at institutionally-affiliated users, each operate a data repository using one of the three popular open source technologies (DSpace, Hydra/Fedora, and Bepress), and all are committed to providing data curation services with similar levels of staffing and well-aligned policies in place (figure 2).

Figure 2: Baseline comparison of six data repository and curation services

Institution	Local Data Repository	Technology Platform	Staffing Full (FT) and Part Time (PT)
University of Minnesota Twin Cities	Data Repository for the University of Minnesota (DRUM)	DSpace 6.x Launched Nov 2014	0 FT / 7 PT (director, coordinator, and 5 data curators at 10% FTE each)
Cornell University	eCommons at Cornell	DSpace 5.x Launched Fall 2012	0 FT /2 PT (data curators)
University of Illinois	Illinois Data Bank	Custom Ruby-on-rails Launched May 2016	1 FT (developer) /8 PT (director, curation, subject and functional specialists at 5-30% FTE each)
University of Michigan	Deep Blue Data	Sufia 7.x (Hydra and Fedora) Launched Feb 2016	1 FT / 5 PT (RDS manager, project manager, developers, subject specialists at 5-20% FTE each)
Penn State University	ScholarShere	Sufia 7.x (Hydra and Fedora) Launched Fall 2012	0FT / 5 PT (product owner 40%, scrum master 10%, project manager and developers 75%)
Washington University in St. Louis	Digital Research Materials Repository (DRMR)	Digital Commons BePress Launched Jan 2015	0 FT / 5 PT (coordinator, repository manager, copyright specialist, subject specialists)

[3] Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. *Journal of eScience Librarianship 6*(1): e1102. https://doi.org/10.7191/jeslib.2017.1102.

The baseline assessment greatly influenced how we envisioned the DCN submission workflow. For example, in our cohort, most data curation services for local deposits occurred post-ingest, meaning that the dataset was first self-deposited to the local system by a researcher and either automatically accepted or accepted following appraisal (e.g., meet local policy, etc.). In other words, the data went "live" for public viewing and access before curation staff took further action (figure 3). Our results suggested that the DCN should utilize a similar post-ingest curatorial review workflow to alleviate any concern about gaining access to datasets that are not publicly available (e.g., behind password protection) or interacting with unfamiliar repository technologies.

Similarly, a common limitation found in our baseline assessment is an inability to host or publish large and active data sets. Acceptable deposit sizes range from 500MB to 15 GB per file (larger ingests mediated), and no institution offers repository services for active databases. Anticipating innovations in this area, we intentionally developed the DCN model independent from local repository infrastructure.

Figure 3: Comparison of workflows for data curation at six institutions

	Pre-ingest Curation?		Mediated vs Self-deposit?		Accept/Reject Stage?		Public	Post-ingest curation		
Workflow Steps by Institution	**Consult only**	**Staging Area for deposit**	**Mediated deposit**	**Self-deposit**	**Approval to accept or reject**	**Auto Accept**	**Go Live Here**	**Review meta-data**	**Review metadata and files**	**Add DOI**
Minnesota	X			X	X		X		X	X
Cornell	X		X*	X		X	X		X	X*
Illinois	X			X		X	X		X*	X
Michigan	X			X		X	X		X*	X*
Penn State	X			X		X	X			
Wash U	X		X	X		X	X		X	X

* On request

3.2 Researcher Engagement Sessions

Building on other user-needs assessments of researchers, performed via survey (Tenopir et al., 2011) and focus groups (Bardyn, Resnick, & Camina, 2012), the DCN team engaged researchers on the importance and utilization of data curation activities. Between October 21, 2016 and November 18, 2016 the team held six focus groups, one at each of the planning institutions, that were aimed at identifying the data curation areas where the DCN should place its focus. Using a mixed-methods approach (discussed in detail in our full report[4]) we identified the data curation activities most important to our researchers, identified which activities were currently happening for their data, and, finally, asked how those activities were happening and level of satisfaction with the results.

In total we engaged with 91 researchers representing a good mix of experience (e.g., faculty, graduate student, post-doc) and disciplines, that directly informed the DCN model. We found that most of the data curation activities presented to researchers were viewed as important or having value to themselves or to their communities of practice. The activities that ranked most highly across two or more groups were:

- *(Create) Documentation* (ranked 4.6/5)
- *Secure Storage* (ranked 4.4/5)
- *Persistent Identifier* (ranked 4.3/5)
- *Quality Assurance* (ranked 4.3/5)
- *Software Registry* (ranked 4.1/5)
- *Data Visualization* (ranked 4.0/5)
- *File Audit* (ranked 4.0/5)
- *(Create) Metadata* (ranked 4.0/5)
- *Code Review* (ranked 3.9/5)
- *Contextualize* (ranked 3.9/5)
- *Versioning* (ranked 3.9/5)
- *File Format Transformations* (ranked 3.8/5)

Only four activities presented to researchers were ranked below a “3” on a 5-point scale and these were: *Emulation, Restricted Access,*

[4] Data Curation Network Special Report (March 2017) "Results of the Fall 2016 Researcher Engagement Sessions" with links to supplemental files, http://hdl.handle.net/11299/188641.

Correspondence (with data author), and *Full-Text Indexing*. Our focus groups also revealed that while researchers were actively engaged in a variety of data curation activities for their data, no activity was happening in a satisfactory way for a majority of respondents. The activity that came the closest was *Secure Storage*, which was happening for 75% of our researchers and in ways that satisfied 38% of our researchers (figure 4).

Figure 4: Levels of satisfaction with the top 12 data curation activities* for researchers

"Does this activity happen for your data?"					"If Yes, Are You Satisfied with the Results?"			
Rank	**Data Curation Activity**	**"Yes"**	**"No"**	**Other** ***	**Yes, Satisfied**	**Somewhat satisfied**	**No, not satisfied**	**No Answer**
#1	Documentation	80%	9%	10%	26%	46%	10%	18%
#2	Secure Storage	75%	17%	9%	38%	18%	3%	40%
#8	Metadata	63%	24%	14%	29%	31%	8%	33%
#6	Data Visualization	58%	25%	16%	13%	33%	4%	50%
#9	Versioning	56%	30%	14%	13%	37%	12%	37%
#12	File Format Transforms.	55%	27%	17%	29%	21%	5%	45%
#3	Quality Assurance	52%	29%	20%	14%	27%	4%	54%
#5	Software Registry	41%	38%	20%	14%	21%	10%	55%
#10	Contextualize	38%	45%	16%	8%	24%	14%	54%
#11	Code Review	38%	34%	27%	22%	14%	5%	58%
#4	Persistent Identifier	37%	44%	18%	19%	33%	11%	37%
#7	File Audit	16%	57%	26%	2%	14%	14%	69%

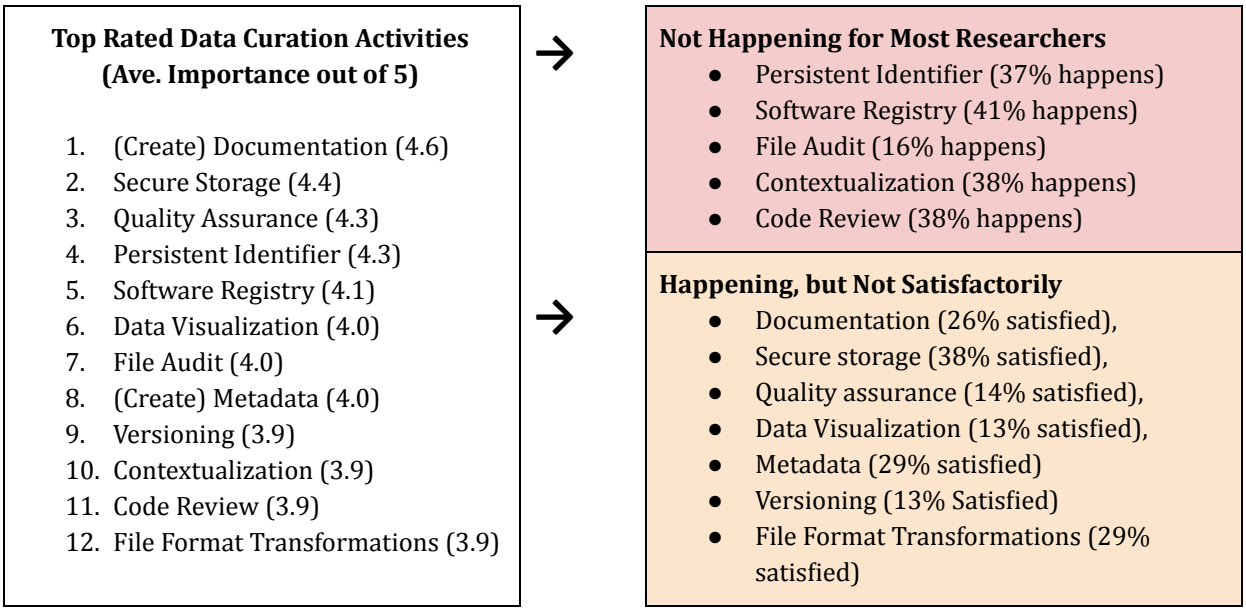
* Based on several dictionary and glossary sources we defined 35 “Data Curation Activities” used in our researcher engagement sessions. Full list and definitions at http://hdl.handle.net/11299/188638..

** Other responses included “I Don't Know,” “Sometimes,” “Not applicable,” and not answered.

Our focus groups included discussions that revealed the various ways in which researchers engage in data curation activities as well as the barriers preventing them. This gave our project key issues to address and specific areas of curation for the DCN model to focus on. For example, we identified “gaps” in highly valued data curation activities that either did not happen for a majority (>50%) of researchers or

happened in a unsatisfactory way (figure 5). The DCN model will most benefit from emphasizing, investing in, and/or heavily promoting these highly valued services that may not be available to many researchers, including: minting and managing persistent identifiers, maintaining a software registry, providing tools and support for auditing file integrity, creating and managing metadata that place data within context of related publication sources, and providing code review services.

Figure 5: Gaps and areas of opportunity for Data Curation Network services



Similarly, the DCN might support better tools and/or best practices to increase the levels of satisfaction for these commonly occurring data curation activities that are falling short of expectations, including maintaining up-to-date data documentation templates that could be used by a variety of researchers, providing best practices for secure storage, creating quality assurance checklists and review procedures for a variety of data formats and types, recommending best practices or tools for data visualization, promoting better adoption of metadata standards across disciplines, recommending tools and file naming schemas for versioning datasets, and by being more transparent about the conditions and procedures that call for file format transformations.

3.3 Data Curation Pilots

From September 2016–November 2016 our team conducted data curation pilots with 17 curator staff in order to identify and compare the actual and individual curation practices taking place at our partner institutions. The results[5] allowed us to identify any issues, misaligned expectations, and/or conflicts prior to implementation of the Network. Namely the pilots gave us a real-world glimpse of the DCN in practice and informed how the DCN model should function, including:

- *Centralize DCN submissions with a Coordinator who performs routine checks on all submissions before assigning to DCN curator.* Not all data sets in our pilot were functional and able to be opened. Therefore our model envisions a DCN Coordinator role, separate from the DCN Curators, that will perform routine checks (risk, rights, file inventory/manifest, and file audit) and open all files to check for integrity issues before sending the assignment to the appropriate DCN Curator.
- *Assignments to DCN Curators should prioritize file format and software expertise over discipline when necessary.* The DCN Coordinator will analyze incoming datasets to the Network and make assignments based on DCN Curator expertise. However, when a curator from our pilot worked with a new or unfamiliar data format type (e.g., software they were not familiar with), they were less confident with the result. Therefore Network curators may bring a general knowledge of a discipline, but their deep expertise with software and domain file formats should be considered.
- *Allow the local curator to control decisions around collaboration with their local researchers (data authors).* Communication and back-and-forth between the researcher was unevenly carried out in our pilots. Some curators emailed descriptive questions while others requested an in-person meeting or a phone call follow-up with the data author. The DCN model should place the responsibility and choice around how to best communicate needed curatorial actions for data on the Local Curator and staff. Not only will this allow for variations in local culture and

[5] Data Curation Network Special Report, (March 2017), "Results of the Fall 2016 Data Curation Pilot," http://hdl.handle.net/11299/188640.

different levels of local support, but it may also eliminate concerns about the opportunity cost of researchers working with external DCN Curators rather than building strong relationships with the Local Curator.

- *Normalize procedures for curators to aim for rather than allowing curators to fall into a never ending quest for high standards.* Curation activities can tend to be never ending, and therefore certain minimum levels of curation must be set and activities prioritized. The DCN should develop and continuously evolve standard levels of curation that will result in well-curated data.

3.4 Surveying the Data Curation Community

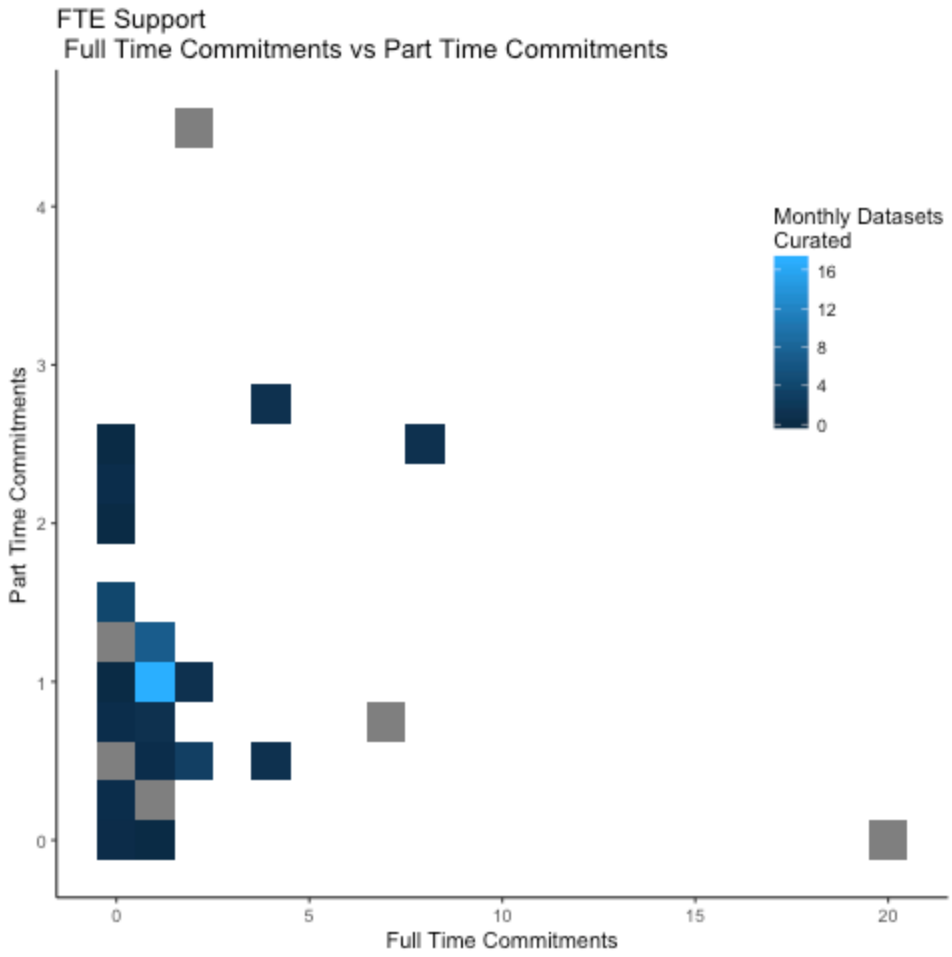
In order to design a Network of data curators, we began with a survey to better understand existing data curation services in academic libraries. Our team partnered with the Association of Research Libraries (ARL) to develop SPEC Kit #354 for data curation.[6] We surveyed the 124 ARL institutions (which include mainly academic libraries based in the US and Canada) in January 2017 to understand current data curation practice and highlight examples and best practices for other libraries to build from. Our results showed that of the 80 ARL Libraries that responded (65% response rate), 51 institutions are providing data curation services and another 13 institutions indicated that they are developing these services. Only 20% of the sample, or 16 libraries, indicated that they do not provide nor are they actively developing data curation services. Of particular note to the DCN, our survey respondents ranked having "expertise in curating certain domain data" as their greatest challenge.

Levels of staffing for providing data curation service was a key consideration of our survey. Indeed, the lack of skilled data curators was one of the challenges that the DCN is aiming to address. Our results showed that the majority of institutions place responsibility for data curation services on individuals who have other duties to carry out (partial or part-time staff). The number of partial staff ranges from one to 15 per library. The percentage of time they spend

[6] SPEC Kit #354: Data Curation, Association of Research Libraries (ARL), May 2017, http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf.

varies widely by institution, with some reporting 5–10% of time and others indicating it may be as high as 40–50% (figure 6). Twenty-eight institutions only have staff devoting a part of their time (a total of 143 individuals). Seventeen institutions have both partial focus and exclusive focus staff (88 partial and 39 exclusive). Three libraries have one person who spends all their time on data curation. Interestingly, there appeared to be little relationship between the number of data sets curated on a monthly basis and the level of staffing. An outlier reported 20 staff devoted exclusively to these activities.

Figure 6: Heat map displaying the reported staffing levels for data curation (part time or exclusive full time) vs the number of monthly data sets curated in ARL Institutions (blue scale with no response indicated in grey). Most provide data curation services with partial staff.



Our survey uncovered another key consideration: data curation activities that are most commonly supported by ARL Institutions are standard features of many IRs (such as *File Download, Deposit Agreement, Terms of Use, Embargo, Use Analytics, Discovery Services, Authentication,* and *Data Citation*). However several "Very Important" activities that rated highly with researchers in the DCN researcher focus groups were not well supported by ARL institutions, including *Quality Assurance, Software Registry, Data Visualization, File Audit, Versioning, Contextualization, Code Review,* and *File Transformations*. As a result, the DCN has an opportunity to help establish a stronger community of practice around data curation services that extend beyond IR features and utilize the network of staff working in small isolated teams across institutions.

3.5 Developing a Financial Model

There is a growing body of literature comparing the various models for supporting sustainable data curation and repository services (Kitchin, Collins, & Frost, 2015; Ember et al., 2013; Nilsen, 2017). We evaluated several financial models in order to develop a sustainable plan for supporting the DCN post grant funding phase. In particular we found the ITHAKA S+R 2016 report, "A Guide to the Best Revenue Models" useful in identifying an approach that will best support the financial needs of the DCN for the next 6 years.

Successful models in the library and information science discipline provide exemplars for collaborative sustainability. The DCN planning team engaged with several peer groups that were doing similar work with providing shared data services to learn from their experiences. For example we interviewed Anne Kenney, now former University Librarian and lead PI on the Cornell University Columbia University collaborative 2CUL project that supports shared collection development and cataloging services jointly at Columbia University and Cornell University (https://www.2cul.org), and Jonathan Markow, lead technologist at DuraSpace (http://www.duraspace.org), a distributed open source digital repository service which supports Fedora and DSpace, and whose code base and service models are supported by a global community. Their experiences taught us to emphasize the community building aspects of the DCN, versus the

economic or cost-savings benefits, and that our collaborative project must be built on trust with those that staff the project (e.g., data curators across institutions).

Additionally, we held information exchanges with representative staff, and (when possible) reviewed MOU agreements from the NSF DataNet SEAD project (http://sead-data.net), Canada's emerging shared data service Portage Network (https://portagenetwork.ca), the Texas Digital Library consortial data repository (https://www.tdl.org/texas-data-repository), the Federal Drug Administration data policy (https://open.fda.gov), the Data Conservancy based at Johns Hopkins University Library (http://dataconservancy.org), the California Digital Library's UC3 project (http://www.cdlib.org/uc3), and the statistical data focused Curate Research Data for Reproducibility (CuRe) project.

Based on our research, anticipated costs for the DCN central services, and a benefit analysis of various stakeholders, we drafted several scenarios for support including tiered membership, fee-for-service, and, in-kind (all effort donated by institutions) models. The details of these models (ie. estimated membership fee costs per institution, tiers of participation, benefits, etc.) were vetted with library administrative staff at our institutions (e.g., Dean of the Library). Our potential scenarios to sustain the DCN included:

- *Tiered Membership Model*: Annual costs would be sustained by a DCN membership fee for all institutions. Those members that contribute curation staff could receive a cost-savings propositional to the staff time donated.
- *Hybrid Alliance + Fee-for-Services*: A core group of institutions contribute staff to the Network. Customer institutions may pay for services at a rate of time spent on curation per dataset.
- *In-kind*: An alliance of institutions contribute staff to the DCN and receive curation services. Central services (e.g. the DCN Coordinator role) will also be donated in exchange for a greater share of curation time.

One theme that emerged from our research was "membership fatigue," a result of requests for support for numerous collective projects. Therefore, our resulting 6-year financial plan to sustain the DCN

(presented in the Sustainability Plan section of this report) will enable us to transition beyond grant support to a model that sustains operations and offers users curation services on a fee-for-service basis, expanding both the funding base for the Network and opportunities for unaffiliated researchers and strategic partners (publishers, new disciplinary projects, etc.) to consume services on a pay-as-you-go model.

3.6 Local Metrics Tracking

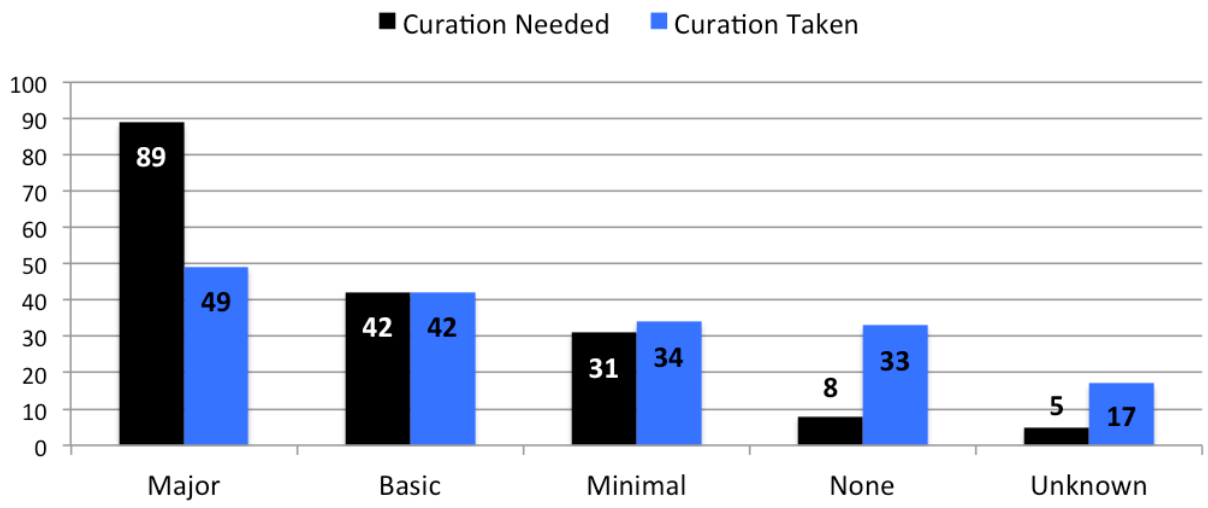
From May 1, 2016–April 30, 2017 the planning phase team tracked key metrics related to the demand for curation services across the six institutions. Using a shared Google form, we imputed the frequency, file types, disciplines, and levels of curation needed for all datasets curated by our individual institutions to better anticipate future staffing needs and demand. The results showed:

- *Frequency*: 176 datasets were curated for deposit into the six local repositories averaging nearly 3-4 datasets per week. Our metrics showed that each institution on average received around 2 data sets per month with a noticeable increase in activity during the non-academic months (May-July and December). We anticipate that not all data sets received at the local institution will be sent to the Network, due to local expertise, etc.
- *File formats*: Within the 176 data sets, we encountered 52 unique formats. The most common formats (found in 10 or more submissions) were comma-tab delimited (.csv), plain text (.txt), Microsoft Excel (.xls), 3D images (.obj), FASTQ biological sequence files (.fastq), portable documents (.pdf) and raster graphics (.tif).
- *Data types*: Spreadsheet or tabular data (found in MS Excel or CSV files) were the most common data type accounting for 27% of the submissions. The wide range of other formats included spatial GIS data, software and programming code (R, python, matlab, java), survey data (SPSS, SAS), audio/video files, databases (.acc, .dat), mass-spectrometry (.raw), Genomic sequence data (.fastq, .fa), and a range of audio/visual files.
- *Researcher demographics*: 76% of our users represented one of the scientific fields (e.g., Agricultural and Natural Resources,

22%; Engineering and Applied, 18%; Biological, 15%; Physical, 14%; Human and Health, 7%) with social sciences, library science, and humanities making up the last quarter. Anthropology, Crop and Soil Sciences, Civil, Environmental, Geo-Engineering, and Oceanography were the top disciplines represented. Additionally, nearly half of the data submitters were repeat users of their local service.

- *Level of Documentation*: 33% of the submissions lacked documentation beyond basic metadata (author, title, date), however, of those that did include documentation, 28% included a plain text readme file.
- *Levels of Curation*: Figure 7 shows the levels of curation *needed* for our sample and the level of curation level *taken* for those data. That more "Major" curation actions were needed than were taken only reinforces the need for a scaled DCN model solution.

Figure 7: Curation levels needed vs. taken for six institutions (n=175)



Legend

Major	Basic	Minimal	None
Major edits to the metadata and/or major changes to the files (new or missing)	Edits to the metadata and/or basic changes to the files	Small edits to the metadata	No edits

4.0 A Cross-Institutional Staffing Model for Curating Research Data

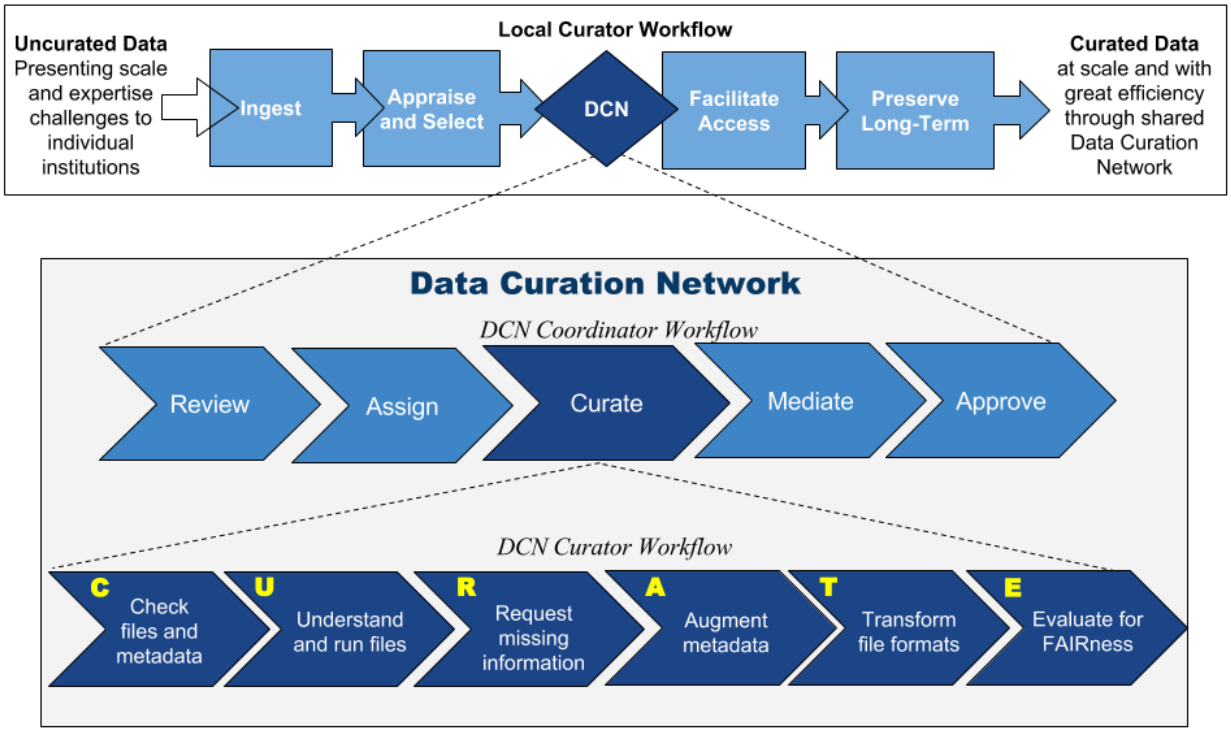
The Data Curation Network model that we propose harnesses the expertise of well-aligned institutions that collectively provide data curation services to researchers in a multitude of disciplines, ensuring that valuable scholarly datasets are findable, accessible, interoperable and reusable, or FAIR. Offered through a unique collaboration between academic libraries and disciplinary projects, DCN curators at distributed sites are matched with data sets according to their technical and disciplinary expertise, and conduct a rigorous review of the data using an established set of protocols that seamlessly fits within any local institutional workflow (figure 8).

Users of the Network will be able to more efficiently work with investigators to capture as much context and description of the data as possible, expertly review data quality and validate code, assess risks and verify file integrity, and validate and transform files. DCN curators also provide guidance around secure storage, citation and persistent identification strategies, and curated data may be deposited into the repository of the researcher's choice for ongoing stewardship.

Implementing the DCN will support and expand the data curation community. Our model will bring together staff with diverse expertise (e.g., domain-specific data curators, informaticians, digital records archivists, preservation specialists, data librarians, etc.) currently siloed in single institutions into a shared network that will collectively, and more effectively, develop standards-driven data curation techniques for all types of data housed in any repository infrastructure (e.g., Fedora/Hydra, DSpace, custom-build, etc.). By expanding local curation expertise through structured, regular

training and hosting community-wide educational opportunities, the DCN will build an innovative community that enriches capacities for data curation writ large.

Figure 8: Curation workflow for the Data Curation Network
([link to full size image])



4.1 Benefits of Using the Network

Academic libraries with existing data curation services:

- gain access to data curation expertise in more disciplines/formats than locally available.
- contribute to a larger ecosystem of data curation practice.
- participate in the development of shared standards.
- build a pipeline for training data curators and establishing professional data curation practices.
- inform and advance development of local curation services.
- smooth and stabilize services during times of staff transition and shortage.

Academic libraries with limited to no resources for data services:

- are able to provide critical new data curation services when local resources are limited (without needing to hire).
- have the opportunity for a local data curation specialist to join a larger, robust network.
- benefit from a clear roadmap, presented by DCN partners, toward data curation services maturity and scale.
- normalizing the practice of data ingest/deposits/archiving in library-hosted repositories.

Disciplinary and subject data repositories:

- receive better, more valuable data submissions from DCN partner institutions and customers.
- have potential to partner with the DCN to expand the scope of curation support for the disciplinary repository to new and/or less frequently encountered data types.
- gain access to curation staff that are housed at external institutions thereby minimizing staffing overhead costs.
- get more researchers directed to the disciplinary repository thanks to the broad network of participating institutions .
- obtain potential new revenue stream as consumption scales, should the disciplinary repository seek to join as a partner.

4.2 Roles and Responsibilities

The DCN will function through supportive organizational layers and dedicated staff that contribute to a shared governance system to be determined in the implementation phase. An important consideration of the DCN staffing model is maintaining and strengthening local relationships with researchers. Therefore, to provide opportunity for future engagement, our model incorporates the following DCN staff roles and local or institutional resources the DCN staff will interact with. Roles in the DCN include (more detailed descriptions of responsibilities for each role detailed in Appendix A):

- *DCN Users and Local Resources:*
  - **Local Researcher**: The individual responsible for the dataset. Often the creator of a dataset but may also be a representative acting on the author's' behalf (e.g., a graduate assistant).

        - **Local Curator:** The staff member who submits a dataset from their home institution to the Network. The Local Curator continues to serve as the primary contact for all communications with the Local Researcher throughout the curation process.
    - *DCN Staff:*
        - **DCN Curators:** Staff that provide expert curatorial services for the Network. They bring curation skills for specific file formats (e.g., databases, statistical survey data, video/audio files, computer code) and/or types of disciplinary data (e.g, 3D images, genomics, chemical spectra, ecological, etc.). DCN Curators take on the role of Local Curator when submitting data from their institution. DCN Curators benefit from annual training events and virtual networking with peers in the DCN.
        - **DCN Coordinator:** This individual, centrally funded through the DCN, oversees the daily operations of the Network, tracks and monitors all datasets that flow through the Network, and assigns incoming data sets to the appropriate DCN Curator.

4.3 Tiers of Participation

The DCN will operate as an alliance of partner institutions (e.g., academic libraries or disciplinary data repositories, etc.) who contribute staffing and funds to sustain and offer central services to potential users (e.g., academic libraries, publishers, or individual researchers). The proposed levels of participation will include, but are not limited to:

- **Institutional Partner**: Institutional partners, either from academic- or disciplinary-based institutions, contribute data curation staff time to the Network (at a rate of at least one 10% FTE per institution) and contribute financially to support central operations. Partner institutions may gain access to the Network for curating data sets by Local Researchers at rates established by a MOU and participate in governance functions (see details in Appendix B).

- **Institutional User**: These academic institutions and disciplinary repositories may gain access to expert data curation services on a fee-for service model or by paying membership fees. Institutional users do not provide in-kind curation staff. They do not participate in the Network governance functions.
- **Individual User**: In the future, the DCN may also offer direct services to individual researchers on a fee-for-service basis, expanding both the funding base for the Network and opportunities for unaffiliated researchers and strategic partners (publishers, new disciplinary projects, etc.).

4.4 Criteria for New Partners

Applications for new partner institutions will be considered on a rolling basis. A Memorandum of Understanding (draft presented in Appendix B) will outline the functional aspects of the model, roles and responsibilities of the staff involved, and other normative practices and expectations. Institutions interested in joining the Network will review the MOU and provide an expression of interest via an online DCN application form (to be created).

Draft DCN partnership criteria may include, but are not limited to:

- **Services and Policies**: Data curation services are currently offered to specified users (e.g., local researchers) where the data is destined to reside in a known repository (e.g, an institutional repository, a subject based or disciplinary repository, a data storage facility hosted or provided by the institution). Since data hosting and preservation are not services of the of the Data Curation Network at this time, a URL or a description of the final destination of the curated data is required.
- **Dedicated Staff**: Each DCN affiliate is required to identify dedicated staff (name, job title, expertise, etc.) to fulfill the roles and responsibilities (described in Appendix A) required at the partnership tier. Staff roles include one DCN Representative and at least one DCN Curator that bring unique expertise when curating data for the Network. These roles may be fulfilled by a single person if desired. The personnel at this

level may be asked to demonstrate their qualifications and describe their unique skills for these roles through a cover letter detailing their qualifications, certifications (e.g., Society of American Archivists Certificate of Digital Archives specialization), and previous experience.

- **Training Support**: The institution is willing to support annual travel costs to send DCN Representatives and DCN Curators to training events (expected to be held on an annual basis). The DCN may charge a nominal registration fee for training events solely to cover costs. Professional development stipends may also be available for additional travel to specialized training that will benefit personnel in the Network.

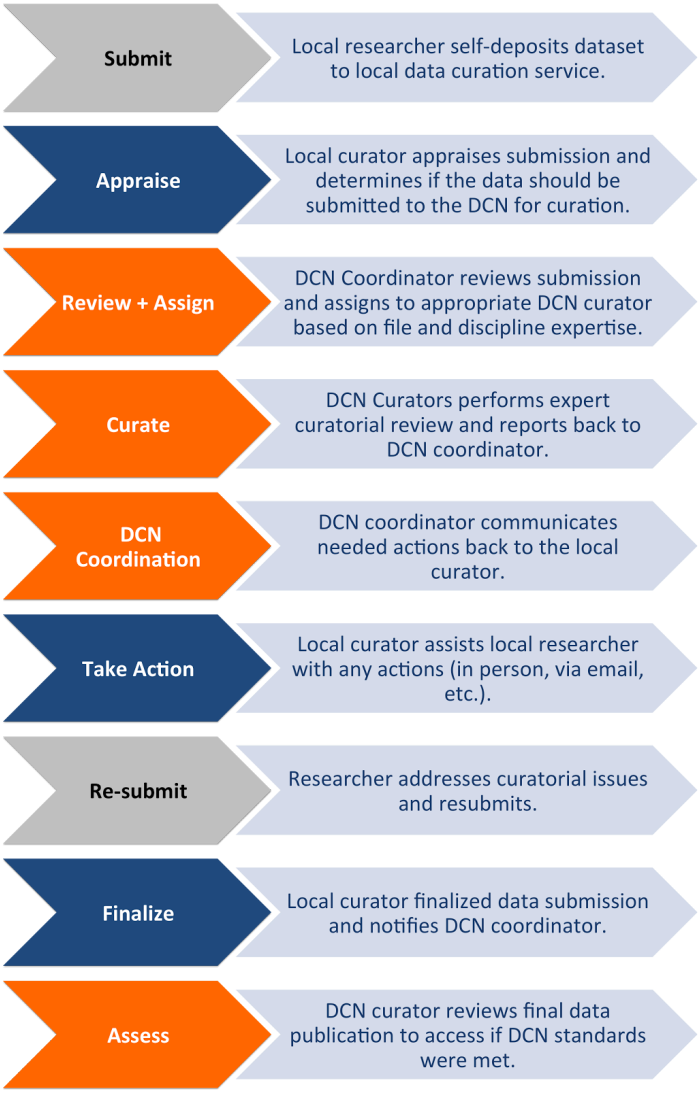
4.5 DCN Submission Workflow

The DCN model is intended to accommodate a wide variety of local curation workflows while remaining repository-technology agnostic. The submission workflow assumes that all technical functionality (ingest, storage, access, dissemination, and preservation) are the responsibility of the local institution. Therefore, local researchers may submit data to their local curation service like normal. Then the Local Curator must determine if the dataset should be submitted to the DCN for expert curation and review. Figure 9 briefly describes this process while more detailed workflows and curator checklists are presented in Appendix C.

Datasets received by the Network will be handled via a submission tracking tool (functional requirements listed in Appendix D) to track where a dataset is in the DCN workflow and the duration in each step. DCN submissions receive a preliminary check from the DCN Coordinator before being assigned to an appropriate DCN Curator (based on expertise match and availability). Once assigned a dataset, the DCN Curator is responsible for reporting any questions, changes, augmentations, and corrections for the data back to the the Local Curator. Researchers may choose not to take recommend actions and therefore the last step in the DCN workflow is for the DCN Curator to assess the final result in order to determine if it meets standards for FAIRness (Dunning, de Smaele & Böhmer, 2017).

Any issues (e.g., problems with a particular dataset) can be discussed at the regular curator virtual meetings where all DCN curators may participate. Here peers may recommend additional actions be taken or collaborate on resolutions for copyright issues, documentation, etc.

Figure 9: DCN Workflow Steps



DCN Curators take standardized and file type specific actions when reviewing the data for fitness for reuse using their expert skills and domain specific knowledge. Specifically, curators will take **CURATE** steps (detailed in Appendix C) for each data sets that include:

C – Check data files and read documentation

U – Understand the data (try to), if not…

R – Request missing information or changes

A – Augment the submission with metadata for findability

T – Transform file formats for reuse and long-term preservation.

E – Evaluate and rate the overall submission for FAIRness.

5.0 Implementing the Data Curation Network

Next, our team will launch a valuable new service that will benefit researchers, their disciplines, and the end users of research data world-wide. The implementation phase of the Data Curation Network will put the model presented here into action by incrementally adding new partners from academic institutions and disciplinary organizations (figure 10). Our proposed curation-as-service model will allow the DCN to grow and sustain with controlled member-driven expansion into new service areas in the years to come. Finally a two-pronged assessment approach will track and assess DCN success and also aim to demonstrate that data curated by the Network are more valuable to users than non-curated data. Along the way the project team will develop and share standards-driven data curation techniques, measure the impact of data curation services, and provide essential training to a cohort of data curators.

Figure 10: Six year plan for implementing the Data Curation Network

	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Support	Sloan Grant	Grant Funded (Y1-Y2) transition to partnership model (Y3)			Curation-as-service (Y4-6)		
Timing	2016-17	2017-19		2020-22		2022-2023	
Phase	Planning	Implementation		Transition		Sustaining	
Partners	6 academic institutions	8 academic institutions and 2 disciplinary partners			Recruit new partners as use and demand dictate		

5.1 Phased Implementation Plan

The first two years of a three-year implementation phase, the DCN will aim to be supported by startup grant funding and the contributed efforts of the six planning phase institutions (Minnesota, Cornell, Illinois, Michigan, Penn State, and Wash U.) plus two additional academic partners and two disciplinary partners.

Each of the 8-10 partners will contribute a minimum of 5% of a DCN representatives' time and also contribute between 5-10% FTE of 1-2 additional data curation specialists. A lead institution (currently the University of Minnesota) will also contribute 15% of the DCN Lead's time (Lisa Johnston) to provide overall direction and supervise a full-time DCN Coordinator, to be funded by the grant. Disciplinary partners may commit either 5-10% of a specialist's time or some other in-kind service that will add value to the Network. Depending on the disciplinary partner, this could be submission access to their repository, reduced or eliminated fees to partners, or some other benefit.

During the implementation phase, several activities will take place. DCN staff will establish communication channels (e.g, a shared listserv, Slack, regular video conferencing etc.) and set up the submission tracking form. An in-person DCN meeting will bring DCN Curators together for training and networking. Another key activity will be to establish and maintain an up-to-date skills inventory of DCN Curators to document available curation expertise and identify gaps for future recruitment.

The implementation phase of the DCN will continue to track trends in the types of domains or file types that come to the Network and work to recruit new institutions that might fill any gaps in expertise support. Capacity for curating data in the Network will grow as new partners join. For example, we found from our one-year of metric tracking that curators spend an average of 2 hours to curate a dataset (ranging from less than 1 hour to 8+ hours). In year 3, if each institution contributes 10% of a DCN curator time (assuming 10% FTE = 16 hours/month) then with 10 institutions the DCN will have roughly 160 curation hours or the capacity to curate an average of 80

data sets each month. Finally, the DCN will establish a public facing directory of datasets that were successfully curated by the Network. This web resource will be directional and link to the distributed and locally housed datasets. The technical mechanism for bringing together the DCN-approved data sets will aim to utilize the OAI-standardized metadata from each institution's open API feeds to function autonomously.

5.2 Sustainability Plan

Our proposed model will allow the DCN to grow and sustain with controlled expansion into new service areas in the years to come. In the third year of the implementation phase, the DCN will transition to a self-sustaining service where institutional and disciplinary partners contribute data curation staff and share the central operations costs.

The core partner institutions will share any central costs so that the Data Curation Network will continue beyond the implementation phase and without the additional aid of grant support. Any financial support contributed by partner institutions (along with in-kind curator staff) will sustain a number of potential centralized services, including the hire of one full-time DCN Coordinator and annual DCN Curator training events (figure 11). Costs may be offset by potential revenue streams (figure 11), as fee-for-service users increase, and/or if the DCN becomes affiliated with a parent association to act as fiscal agent and cover some of the overhead burden.

The DCN planning phase team reviewed several governance documents of peer organizations, including the 2CUL Project, arXiv, DataOne, HathiTrust, Portage, and the Texas Digital Library, in order to draft a Memorandum of Understanding for partner institutions. Our DCN draft MOU anticipates the need for a governance body that advises on any major issues encountered by the Network staff. However, details for the makeup and responsibilities of this governing board will be determined in the Implementation phase of the DCN. An updated MOU will reflect any changes to the Network based on lessons learned from the Implementation phase and will be used to normalize and sustain operations of the DCN moving forward.

Figure 11: Central costs and potential revenue streams for the Data Curation Network

<table>
<tr><th colspan="2">Potential Central Costs</th></tr>
<tr><td>Human Resources
DCN Director
DCN Coordinator (1 FTE)
DCN Curators
Hire new curators in areas of need
Assessment specialist</td><td>Administrative
Business office functions (management, billing, etc.)
Telephony/virtual meeting support
Outreach and promotion
Legal support
Emergency/contingency fund</td></tr>
<tr><td>Technical
Registry of shared expertise, used as an exchange to track available capacity and match curators with data sets
Workflow/submission management system
Ticketing/tracking system
Technical support/developer for staging & submission systems and DCN web site</td><td>Events, Travel, Training
Annual meeting (event planning/food)
Training for new curators
Prof dev fund for curators (e.g., attend specialized training?)
Certification (levels of curation, expertise of curators, etc.)</td></tr>
<tr><th colspan="2">Potential Revenue Streams (future)</th></tr>
<tr><td colspan="2">Charge fees for curation services to institutional users
Make and sell a curation toolkit
Curation layer for publishers (e.g. PLOS), or general data repositories (Zenodo, figshare, etc.)
Data enhancement/transformation services (post-share, pre-reuse)
Training and consultation services (e.g., bootcamps, webinars) for institutions ramping up their local curation services.</td></tr>
</table>

5.3 Assessment Plan

The planning phase enabled our team to envision what metrics will be important to track to impact and success of the Data Curation Network over time. Therefore our assessment plan will require several key metrics to be tracked from the start of the implementation phase. This two-pronged approach tackles several things. First, we will closely monitor the number of datasets curated by the Network, the frequency of submission (high-volume time periods, etc.), and the variety and types of data (e.g., unique file formats and range of disciplines that utilize DCN services). An important factor in our assessment will be to track the effectiveness of data curation across the Network by tracking the time a dataset spends at each stage of our workflow (e.g., time from ingest to assignment, time with curator, time

with Local Curator before finalized, etc.). Building on our metric track during the planning phase, the DCN will track the overall level of data curation actions taken on each dataset. We will do this by documenting the level of curation needed for a dataset vs. level of curation taken and how well the finalized data scored on meeting FAIR standards. Figure 12 details the draft DCN "CURATE" procedures which include the steps: Check, Understand, Request, Augment, Transform and Evaluate.

Second, in addition to the above metrics, we plan to monitor *overall* service impact in the following ways:

- *DCN curation services statistics indicate positive growth and capacity*: DCN staff will track metrics to demonstrate the number of datasets curated by the Network over time, the variety and types of data (e.g., unique domains and disciplines that utilize DCN services), and the growth in capacity for data curation services as new Partners join the Network.
- *Data curated by the DCN are more valuable*: The DCN staff will monitor data sets curated by the Network and track the number of downloads, alternative-metrics (such as tweets), and the acknowledgement of DCN by external stakeholders (such as funders & publishers recommending DCN and invitations to participate in policy/standards development).
- *Researcher satisfaction & engagement*: The DCN will send and track responses to post-curation satisfaction surveys by users of DCN services. We will research and assess trust markers for reuse of DCN datasets and track researcher attitudes toward data curation activities building on our prior research.
- *Impact on curation community beyond DCN*: The DCN will become a leaders in data curation best practices and track our impact through our DCN website analytics, the number of peers using DCN educational materials/adopting DCN curation standards, and by giving recognition to the staff who played a role in curating a DCN data set.

Figure 12: Draft procedures checklist of DCN CURATE steps and FAIRness scorecard

CURATE Actions	**Curation Checklist**
Check data files and read documentation ● Review the content of the data files (e.g., open and run the files or code). ● Verify all metadata provided by the author and review the available documentation.	❏ Files open as expected ❏ Issues _____ ❏ Code runs as expected ❏ Produces minor errors ❏ Does not run and/or produces many errors ❏ Metadata quality is rich, accurate, and complete ❏ Metadata has issues _____ ❏ Documentation Type (circle) Readme / Codebook / Data Dictionary / Other: _____ ❏ Missing/None ❏ Needs work
Understand the data (or try to) ● Check for quality assurance and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns. ● Try to detect and extract any "hidden documentation" inherent to the data files that may facilitate reuse. ● Determine if the documentation of the data is sufficient for a user with similar qualifications to the author's to understand and reuse the data. If not, recommend or create additional documentation (e.g., a readme.txt template).	*Varies based on file formats and subject domain. For example….* Tabular Data Questions (Microsoft Excel) ❏ Organization of data well-structured ❏ Not rectangular ❏ Split tables into separate tabs ❏ Headers/codes clearly defined ❏ Define headers ❏ Clarify codes used _____ ❏ Clarify use of "blanks" ❏ Clarify units of measurement ❏ Quality control clearly defined ❏ Unclear quality control ❏ Update/add Methodology
Request missing information or changes ● Generate a list of questions for the data author to fix any errors or issues.	*Narrative describing the concerns, issues, and needed improvements to the data submission*
Augment the submission ● Enhance metadata to best facilitate discoverability. ● Create and apply metadata for the data record, including descriptive keywords. ● When appropriate, structure and present metadata in domain-specific	❏ Discoverability sufficient ❏ Recommend (circle one) full-text index / file compression / file reorder / file descriptions / zip Other _____ ❏ Keywords Sufficient ❏ Suggestions _____ ❏ Linkages Sufficient

schemas to facilitate interoperability with other systems.	❏ Link to Report/Paper ❏ Link to related data sets ❏ Link to source data ❏ Link to other _____
Transform file formats ● Identify specialized file formats and their restrictions (e.g., Is the software freely available? Link to it or archive it alongside the data). ● Transform files into open, non-proprietary file formats that broaden the potential audience for reuse and ensure that preservation actions might be taken by the repository in later steps. Retain original files if data transfer is not perfect.	❏ Prefered file formats in use ❏ Recommend conversion from _____ to _____ ❏ Retain original formats ❏ Software needed readily available ❏ Unclear version of software ❏ Unclear software used ❏ Visualization of data easily accessible ❏ Recommend graphical representation _____ ❏ Recommend web-accessible surrogate _____
Evaluate and rate the overall data record for FAIRness.* ● Score the dataset and recommend ways to increase the FAIRness of the data and become "DCN approved."	Findable - ❏ Metadata exceeds author/ title/ date, ❏ Unique PID (DOI, Handle, PURL, etc.). ❏ Discoverable via web search engines like Google. Accessible - ❏ Retrievable via a standard protocol (e.g., HTTP). ❏ Free, open (e.g., download link). Interoperable - ❏ Metadata formatted in a standard schema (e.g., Dublin Core). ❏ Metadata provided in machine-readable format (OAI feed). Reusable - ❏ Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author. ❏ Clear indicators of who created, owns, and stewards the data. ❏ Data are released with clear data usage terms (e.g., a CC License).

* Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

6.0 Acknowledgements

The Data Curation Network project was supported by the Alfred P. Sloan funded grant project "Planning the Data Curation Network," that ran from May 2016–June 2017. We invite feedback and suggestions for improvement to make this approach as useful to community as possible. Please send comments to the authors at dcn-team@googlegroups.com.

Research Releases from the DCN Planning Phase:

1. Data Curation Network Project homepage, https://sites.google.com/site/datacurationnetwork.
2. Johnston, L. R., Carlson, J., Hswe, P., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. K., & Stewart, C. (2017), Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services, *Journal of eScience Librarianship 6*(1): e1102. DOI:10.7191/jeslib.2017.1102.
3. Definitions of Data Curation Activities used by the Data Curation Network. (2016). http://hdl.handle.net/11299/188638.
4. Results of the Fall 2016 Researcher Engagement Sessions: Data Curation Network Special Report (March 2017) with links to supplemental files: Data Tables, Worksheet Protocol, and Master Card Desk used in Researcher Engagement Focus Groups, http://hdl.handle.net/11299/188641.
5. Results of the Fall 2016 Data Curation Pilot: Data Curation Network Special Report. (March 2017). http://hdl.handle.net/11299/188640.
6. Imker, H., Hudson-Vitale, C., Johnston, L. R., Carlson, J., Kozlowski, W., Olendorf, R. K., & Stewart, C. (2017), "SPEC Kit #354: Data Curation," Association of Research Libraries (ARL). May 2017. http://hdl.handle.net/11299/188643.

Bibliography

Bardyn, T. P., Resnick, T., & Camina, S. K. (2012). Translational researchers' perceptions of data management practices and data curation needs: findings from a focus group in an academic health sciences library. *Journal of Web Librarianship, 6*(4), 274-287. DOI:10.1080/19322909.2012.730375.

Beagrie, N., & Houghton J.W. (2014) The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres, Jisc. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf.

Bloom, T., Dallmeier-Tiessen, S., Murphy, F., Austin, C. C., Whyte, A., Tedds, J., Nurnberger, A., Raymond, L., Stockhause, M., Vardigan, M., & Clarke, T. (2015). Workflows for Research Data Publishing: Models and Key Components. International Journal on Digital Libraries-Research Data Publishing Special, (27). https://www.rd-alliance.org/system/files/Workflows_for_Research_Data_Publishing-_Models_and_Key_Components_submitted.pdf.

Consultative Committee for Space Data Systems. (2011). Audit and Certification of Trustworthy Digital Repositories, Recommended Practice, CCSDS 652.0-M-1, Magenta Book, Issue 1 Washington, DC: CCSDS Secretariat. http://public.ccsds.org/publications/archive/652x0m1.pdf.

Dempsey, L., Malpas, C., & Lavoie, B. (2014). Collection directions: the evolution of library collections and collecting. *portal: Libraries and the Academy, 14*(3), 393-423. DOI:10.1353/pla.2014.0013.

Dunning, A., de Smaele, M., and Böhmer, J. (2017, January 31). Are the FAIR Data Principles fair?. Zenodo. DOI:10.5281/zenodo.321423.

Ember, C., R. Hanisch, G. Alter, H., Berman, F., Hedstrom, M., & Vardigan, M. (2013). Sustaining domain repositories for digital data: A white paper. http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf.

Erway, R. (2012). Swatting the Long Tail of Digital Media: A Call for Collaboration. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-08.pdf.

Fecher, B., Friesike, S., Hebing, M., & Linek, S. (2017). A reputation economy: how individual reward considerations trump systemic arguments for open access to data. Palgrave Communications. 3 (17051). doi:10.1057/palcomms.2017.51.

Holdren, J. P. (2013). Increasing access to the results of federally funded scientific research. Office of Science and Technology Policy, Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Ithaka S+R. (2016). A Guide to the Best Revenue Models and Funding Sources for Your Digital Resources. http://www.sr.ithaka.org/publications/a-guide-to-the-best-revenue-models-and-funding-sources-for-your-digital-resources/.

Johnston, L. R. (2014). A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot. University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/162338.

Johnston, L. R., Carlson, J., Hswe, P., Hudson-Vitale, C., Imker, H., Kozlowski, W.,. Olendorf, R. K., & Stewart, C. (2017). Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. *Journal of eScience Librarianship 6*(1): e1102. DOI:10.7191/jeslib.2017.1102.

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., … Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine, 23*, 3-4. DOI:10.1045/march2017-kindling.

Kitchin, R., Collins, S., & Frost, D. (2015). Funding models for Open Access Repositories. Maynooth: Maynooth University. Dublin:the Royal Irish Academy and Trinity College Dublin. DOI:10.3318/DRI.2015.4.

Kollen, C., Kouper, I., Ishida, M., Williams, S., & Fear, K. (2017). Research Data Services Maturity in Academic Libraries. In L. R. Johnston (Ed). *Curating Research Data Volume One: Practical Strategies for Your Digital Repository* (33-60). Chicago, IL: American College & Research Libraries, American Library Association. http://hdl.handle.net/10150/622168.

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. DOI:10.1371/journal.pone.0173987.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., & King, J. L. (2016). Liberating field science samples and data. *Science, 351*(6277), 1024-1026, DOI:10.1126/science.aad7048.

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

National Research Council, Committee on Future Career Opportunities and Educational Requirements for Digital Curation, Board on Research Data and Information, Policy and Global Affairs. (2015) *Preparing the Workforce for Digital Curation.* Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=18590.

Nilsen, K. (2017). "Beyond Cost Recovery: Revenue Models and Practices for Data Repositories in Academia." In L. R. Johnston. Curating Research Data Volume One: Practical Strategies for Your Digital Repository. ACRL: Chicago, IL. p 193-211,

http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf.

Roche, D. G., Kruuk, L. E., Lanfear, R., & Binning, S. A. (2015). Public data archiving in ecology and evolution: how well are we doing?. PLoS Biol, 13(11), e1002295. DOI:10.1371/journal.pbio.1002295.

Smith, R., & Roberts, I. (2016). Time for sharing data to become routine: the seven excuses for not doing so are all invalid. F1000Research, 5. DOI:10.12688/f1000research.8422.1.

Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and Data Support Services: A Study of ARL Member Institutions. Association of Research Libraries. http://www.arl.org/storage/documents/publications/escience-report-2010.pdf.

Stodden, V., Guo, P., & Ma, Z. (2012, September). How journals are adopting open data and code policies. In *The First Global Thematic IASC Conference on the Knowledge Commons: Governing Pooled Knowledge Resources*. https://pdfs.semanticscholar.org/fde2/8f99bc049044c8191abdbcead9d396668028.pdf.

Stuart-Stubbs, B. (1975). An Historical Look at Resource Sharing. *Library Trends, 23*, 4, p. 649-64. http://hdl.handle.net/2142/6812.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101. DOI:10.1371/journal.pone.0021101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One, 10*(8), e0134826. DOI:10.1371/journal.pone.0134826.

Weber, D. (1976). A Century of Cooperative Programs Among Academic Libraries. *College & Research Libraries, 37*(3), 205-221. DOI:10.5860/crl_37_03_205.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*. DOI:10.1038/sdata.2016.18 (2016).

York, J., Gutmann, M., & Berman, F. (2016). What Do We Know About The Stewardship Gap? University of Michigan Deep Blue. http://hdl.handle.net/2027.42/122726.

Author Bios

Lisa R. Johnston is the Research Data Management/Curation Lead at the University of Minnesota Twin Cities Libraries. Johnston coordinates the library's efforts around research data management and leads a team of five data curation experts for curating research data in the Data Repository for the University of Minnesota (DRUM). Since 2012, Johnston has also served as the co-director of the the University Digital Conservancy, the University of Minnesota's institutional repository for research and publications. Johnston has authored numerous publications, and most recently edited and authored the two-volume *Curating Research Data: Practical Strategies for Your Digital Repository* (2017) published by the American College and Research Libraries Press.

Jake Carlson is the Research Data Services Manager at the University of Michigan Library. Carlson oversees the creation, implementation and operation of Research Data Services (RDS) at the Library, which includes the development of the library-based data repository, Deep Blue Data, launched in 2016. Carlson is a primary architect of the Data Curation Profile Toolkit (http://datacurationprofiles.org) and the PI of the Data Information Literacy project (http://datainfolit.org). He is the co-editor of *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers* (2015, Purdue University Press) and the author of numerous articles on roles for librarians in managing and curating research data.

Cynthia Hudson Vitale is the Data Services Coordinator at Washington University in St. Louis Libraries. In this position, Hudson-Vitale leads research data services and curation efforts for the Libraries. Since coming into this role in 2012, Hudson-Vitale has worked on several funded faculty projects to facilitate data sharing

and interoperability, while also providing scaleable curation services for the entire University population. Hudson-Vitale currently serves as the Visiting Program Officer for SHARE with the Association of Research Libraries.

Heidi Imker is the director of the Research Data Service (RDS) at the University of Illinois at Urbana-Champaign. Imker came to the University Library in 2014 to lead the Illinois RDS, a campus-wide initiative that provides the Illinois research community with the expertise, tools, and infrastructure necessary to manage and steward research data. Prior to this position, Imker was the Executive Director of a large scale collaborative grant funded by NIH, called the Enzyme Function Initiative. There Imker was the co-director of the Data Core which aimed to manage, disseminate, and integrate research data produced by 15 different research groups across the disciplines of microbiology, metabolomics, molecular biology, structural biology, enzymology, and computational biology.

Wendy Kozlowski is the Data Curation Specialist at Cornell University. Kozlowski is coordinator of the Cornell Research Data Management Services Group, a cross-campus, collaborative organization that provides data management services to faculty, staff and students throughout the entire research process. Operating within Cornell University Library's Metadata Services group and as part of the library's institutional repository (eCommons) administrative team, Kozlowski is the point person for both repository-wide and scientific metadata, and works with subject liaisons to curate data sets deposited into eCommons. Kozlowski has a B.A. in biology and a M.S. in ecology, and spent 19 years in biology and oceanography research, working on multidisciplinary data sets and with teams from numerous institutions both in and outside the United States.

Robert Olendorf is the Research Data Librarian at Penn State University. Olendorf chairs the committee to develop comprehensive data services at the library and build collaborative relationships with other data service providers at the university. Since 2015 Olendorf has worked with primarily science faculty and students to help them better manage and curate their data often in collaboration with other campus partners. Olendorf is also the product owner of the Penn State

institutional repository, ScholarSphere. Prior to Penn State Olendorf worked at Los Alamos National Laboratory and University of New Mexico. Prior to life as a librarian, Olendorf was an evolutionary biologist focused on the evolution of cooperative behavior and also sexual selection usually working in cross disciplinary groups among multiple institutions. This work incorporated a variety of data including field experiments and observations, high performance computing simulations and large molecular genetic data sets.

Claire Stewart is the Associate University Librarian for Research and Learning at the University of Minnesota. Prior to arriving at Minnesota in 2015, Stewart held several positions at Northwestern University over a 21-year period, including directing the Center for Scholarly Communication and Digital Curation and serving as Head of Digital Collections. At Northwestern, Stewart served as campus lead for repository services and e-science, directing the creation of an E-Science Working Group and data management services as a collaboration between the office for research, information technology, and the library. At the University of Minnesota, Stewart is a member of the Libraries senior leadership team and co-sponsor of the Data Management and Curation Initiative. She directs the Libraries' education and research support programs, leading staff who provide general and specialized support, including GIS, digital humanities, and data management and curation services.

Appendix A: Roles and Responsibilities of Key DCN Staff

Each operational role in the Data Curation Network will have key responsibilities.

DCN Coordinator: This individual, centrally funded through the DCN, oversees the daily operations of the Network, tracks and monitors all datasets that flow through the Network, and assigns incoming data sets to the appropriate DCN Curator. DCN Coordinator responsibilities include:

- Take necessary action when a new data submission enters the Network through the use of a tracking tool (see Appendix D) and assign submission to appropriate and/or best fit curator.
- Inspect incoming data submissions (review the files and metadata) and if needed, run reports (identify finder, bulk extractor) for risk management and file inventory, file validation
- Create and manage a knowledge base and email templates of typical curator recommendations. Maintain best practice handouts and guidelines and encourage actions that have worked in the past.
- Closely monitor curation assignments and triage assignment of new data submissions to appropriate curator. Provide quality assurance of DCN curation activities and not let data submissions "fall through the cracks."
- Be the point of contact between the Local Curator and the DCN Curator. Resolves questions or connects DCN curator to additional support or resources if needed.
- Reviews finalized data sets curated by the DCN to determine if the needed actions were taken and if they meet the curation standards set by the Network (e.g, issue DCN "Badge" or track in a public-facing directory of DCN curated data sets).
- Keep records of the curation work done through the DCN. Monitors how often data of various types are curated by the Network and notifies DCN Board of any heavy use of the Network for any particular data type or institution. Tracks metrics to support evaluation of the services provided and inform the DCN Representatives of areas of strength and weaknesses.
- Lead regular DCN Curator check-in meetings (virtual conference calls).

Local Curator: Each DCN user will designate a staff member who submits a dataset from their home institution to the Network. Local Curator responsibilities include:

- Determine which local data submissions should be sent to the DCN for review and uses discretion as to what extent they involve researchers in this decision.
- Perform a preliminary review of the data before submitting assignment to the Network to ensure submission meets local appraisal and selection guidelines. Ensures that the data do not contain any private or sensitive data that should not be released to a third party.
- If data are not publically available (post-ingest curation), moved a copy of the dataset files and metadata to a web-accessible shared storage account (e.g., Box.com).
- Is the primary contact for the local researcher and responsible for communicating all recommended changes made by the DCN Curator in ways that fit institutional culture (e.g., email list of curatorial changes or meet with researcher in person).
- Once data are curated by the Network, responsible for completing deposit for access whether to a local institutional data repository or for a disciplinary data repository.
- Once data are finalized, notify the DCN Coordinator for review and lists any known limitations that prevented recommended changes from happening locally.
- Responsible for any local storage, preservation, and access needs of the data going forward.

DCN Curators: Each partner institution will contribute 1-2 data curation staff (at 5%-10% FTE) to provide expert curatorial services for the Network. DCN Curator responsibilities include:

- Participate in regular (virtual) check-in meetings with DCN Curators.
- Participate in annual training for DCN Curators which will preferably be held in-person to build relationships and ensure strong communications channels across the Network.
- Contribute to a knowledge base of curation procedures, standards, and frequently asked questions or situations with guidance on how to address them.
- Take necessary actions to curate data assigned to them, including the following CURATE steps:
    - **C**heck files and read documentation.
    - **U**nderstand the data (or try to), if not...
    - **R**equest missing information or changes.
    - **A**ugment metadata for findability.
    - **T**ransform file formats for reuse.
    - **E**valuate for FAIRness.
- Document their work and the changes made to the data set in ways that could be included in a provenance log for the data set by the Local Curator.
- Complete data curation assignments with high level of professionalism and in a timely fashion. Track progress in the shared tracking tool / or in close communication with the DCN Coordinator.

DCN Representatives: Each partner institution will select one DCN Representative to participate in the Network as the institutional representative. DCN Representatives are also the DCN planning phase collaborators and authors of this report. Responsibilities include:

- Is the primary point of contact for all DCN updates and responsible for addressing staffing and performance issues with respect to the DCN Curators from that institution.
- Represents their local institution and participates in DCN governance activities.
- Keeps up to date with DCN procedures, policies, MOU updates, and general issues regarding the Network.
- Represents the DCN at conferences and other professional development opportunities to promote the Network.
- May also hold the role of DCN Curator and/or Local Curator (when sending data sets from home institution to the Network).

DCN Lead Representative: A DCN Lead Representative, based at the lead institution (currently the University of Minnesota), will provide overall direction, outreach, and marketing for the Network. DCN Lead Representative responsibilities include:

- Serves as lead to all DCN operations and service.
- Provides overall leadership for DCN procedures and policy implementation.
- Responsible for leading annual DCN meetings and governance events.
- Markets the DCN to potential new partners and provides outreach and communications to the broader stakeholder community (e.g., Deans at DCN affiliated institutions, etc.).
- Communicates performance metrics to DCN staff and stakeholders.
- Supervises the DCN Coordinator (prefer that both staff are based at the same institution).
- This role could rotate to other DCN Representatives.

Appendix B: Draft Memorandum of Understanding for Institutional Partners

Background: The DCN planning phase team reviewed several governance documents of peer organizations, including the MOU's from the 2CUL Project, arXiv, DataOne, HathiTrust, Portage, and the Texas Digital Library. Our DCN draft MOU (figure 13) anticipates the need for a governance body that advises on any major issues encountered by the Network staff. However, details for the makeup and responsibilities of this governing board will be determined in the Implementation phase of the DCN.

Figure 13: Draft MOU for the Data Curation Network partner institutions

Subject to discussion and change during the first two years of the implementation phase.

Introduction

Research data curation is a costly process in terms of staffing, especially because it often requires specialized expertise in a particular domain. It can be difficult or impossible for an institution to maintain adequate staff to curate the variety of data that might be created by the institution. The Data Curation Network (DCN) is a collaborative staffing model created to facilitate the curation of research data across the Network by using the expertise of staff at each member institution to fill the gaps that might be found at any particular institution and also to provide exchange of knowledge between the institutions.

To help alleviate this problem, [Insert member institution name here] will join the DCN as a [member level] to share expertise and staffing. This partnership will provide [insert member institutions] with access to the expertise and resources of the DCN while also benefitting from the existing and future pool of resources of the DNC.

Definitions

Research Data: the recorded factual material commonly accepted in the research community as necessary to validate research findings. The DCN takes a very inclusive view of what constitutes research data, but excludes journal articles, white papers and other material that is primarily an interpretation of research data.

Curation: the processes and activities related to the organization and integration of data collected from various sources, annotation and documentation of the data and the publication and presentation of the data such that the value of the data is maintained and remains available over time.

Policy and Procedures

[Insert member institution name here] will be able to submit data sets to the DCN for curation when needed using the workflow published at [link here]at a rate of [insert maximum number of data sets] or more as capacity allows. Likewise, [insert institution name here], will be available to curate data that corresponds to their stated areas of expertise at a rate of at least [insert percentage and hours] of curator time. [Insert member institution here] will be responsible for the initial collection of content and metadata from the data depositors. The DCN and its member institutions will further curate the data as described in the published SOP [link here]. The DCN will send all communications through the [insert institution name here]'s representative contact person(s).

The DCN governing board lead [insert name and email] is the primary contact for questions about governance and policy. The DCN Coordinator {insert name and email] is the primary contact for submitting data sets for curation, receiving data sets for curation and questions about curation.

[Insert institution name here] retains all rights to the content curated by the Network for the [institution name] and likewise makes no claims against data it curates. The intellectual property rights for any training, education or outreach materials created by [institutions name] for the DCN will be retained by the DCN under CC4-Attribution license.

By signing the agreement [insert institution name] agrees to an annual membership fee of [insert amount or delete line if its in kind] and expects to donate approximately [insert FTE] FTE in curatorial effort to the DCN. The DCN will provide access to the expertise of the Network, additional training and support as required.

[insert institution] may withdraw from the DCN at any time although any annual dues would be forfeit. [insert institution] will no longer have any obligation to curate research data, nor will it have access to the DCNs curatorial services. The DCN retains the right to suspend member institutions in the case that they do not fulfill their obligations as outlined above.

Updates to the MOU
The DCN retains the right to modify aspects of the MOU not directly affecting the cost of the partnership to either the DCN or its partners. This includes changes to SOPs and workflows, recommended best practices and other aspects of the partnership that pertain to maintaining the quality of the curation without affecting the cost to either the DCN or the partner institutions.

The DCN may ask for changes to in kind compensation or membership dues as reviewed on an annual basis, but changes must be agreed to and signed off on by both parties.

To maintain fairness and sustainable, the DCN may, on an annual basis, review member institutions to ensure that their obligations are being met. In cases where it is decided that obligations have not been met the DCN retains the right to ask for changes or if needed to terminate the partnership.

Conclusion
It is the hope of the DCN that this MOU will help the partners understand the nature of their partnership with the DCN. In addition, it should provide partner institutions the mechanisms for communicating and collaborating with the DCN.

Appendix C: Draft DCN Workflows for DCN Curators

This section first provides an overview (figure 14) and then details the workflow steps (figure 15) drafted for the various roles in the DCN. Once implemented, DCN curators and representatives will be expected to communicate on a regular, ongoing basis (e.g., bi-weekly conference calls) in order to share out on curation assignments and make adjustments and changes to the workflow as new situations arise.

Figure 14: Swimlane diagram of the roles and steps involved with the DCN workflow

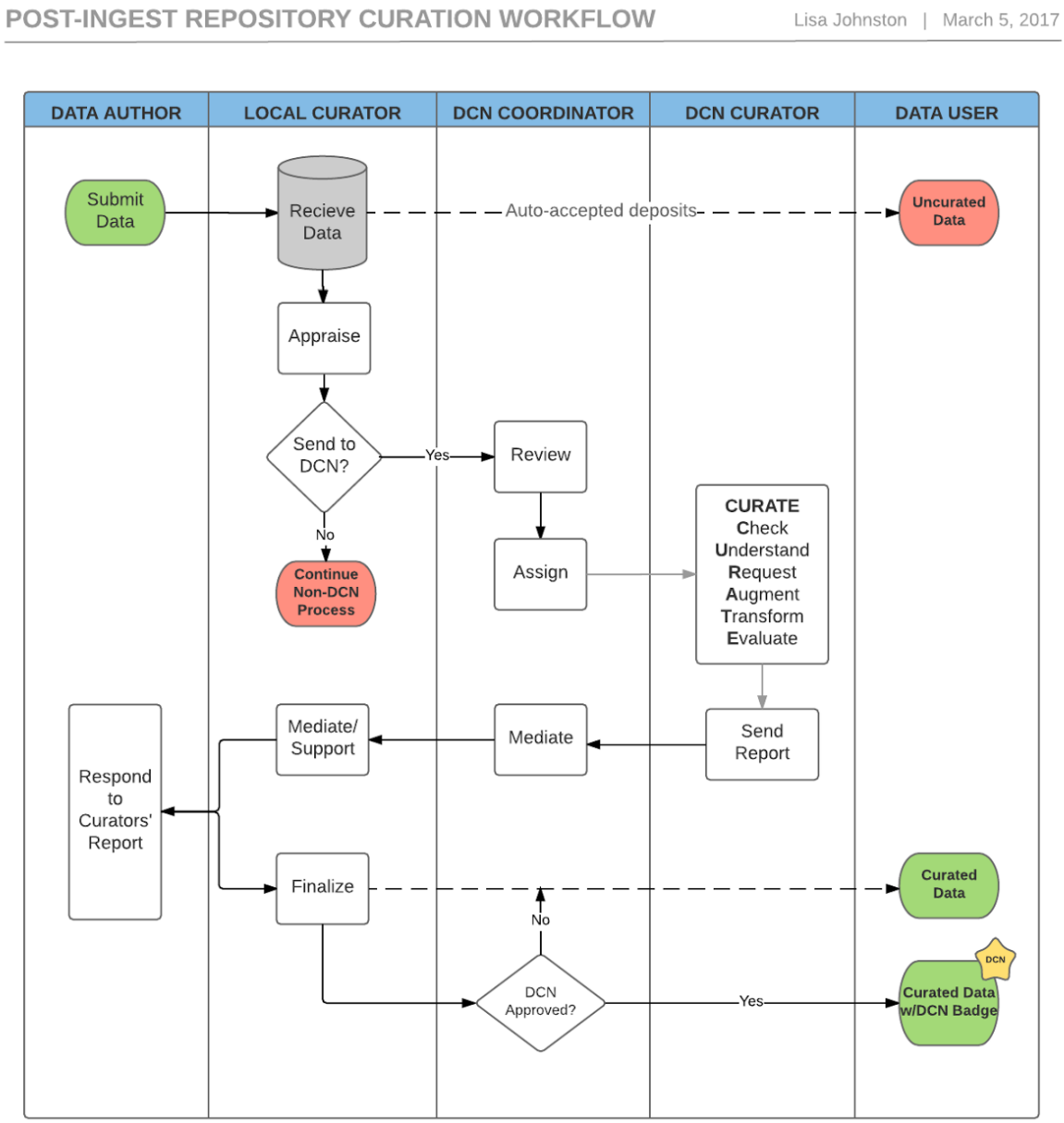


Figure 15: Detailed workflow steps in the DCN Model

Post-Ingest Curation Scenario: Data Intended for an Open Access Institutional Repository			
#	**Step**	**Roles and Responsibilities**	**Data Curation Activities**
1	Submit Data	Role: Local Researcher Action: Self-deposits dataset to local open access data repository. Responsibilities: ● Understand and agree to the terms of deposit into the repository (e.g., sign a deposit agreement). ● Follow any policies involving legally protected or restricted use data (e.g., deidentify potentially disclosive information prior to deposit). ● Provide metadata and/or documentation pertaining to the data files at a level appropriate for reuse (e.g., complete a submission form).	Authentication Deidentification Deposit agreement Metadata Contextualization
2	Appraise	Role: Local Curator Action: Appraise the data submission and determines if the data should be submitted to the DCN for curation. Responsibilities: ● Determine that the local repository is the appropriate home for this data (e.g., the data meets all collection policies, risks). ● Facilitate the ingest of the data files in a secure manner that protects the integrity and authenticity of the data (e.g., generate file checksums). ● Store the ingested files securely in a well-configured (in terms of hardware and software) archival storage environment. ● Organize and rename the files to optimize their meaning, and display them in a way that might facilitate reuse. ● Generate and maintain a persistent identifier (e.g., a DataCite DOI) to the data. ● Starts a new ticket for the DCN with either a link to the publicly accessible version in the repository or provides access to a copy of the data in a shared, secure location.	Appraisal/Selection Persistent Identifier Risk Management Chain of Custody Arrangement and description Transfer to DCN
3	Review + Assign	Role: DCN Coordinator Action: Reviews submission and assigns to appropriate DCN curator based on file format, discipline expertise, and other factors such as availability. Responsibilities:	File Inventory File Validation Link Checking Virus Scan Expertise Match

		• Inventory the submission and document the number, file types, and file sizes of the data. • Identify any missing, duplicate, or corrupt (e.g., unable to open) files. Red flag issues. • Determine if any additional information or files need to be acquired from the author before assigning to a curator. • Assign the submission to the appropriate data curator based on subject and format expertise required and availability.	
4	*CURATE (see steps)*	Role: DCN Curators Action: Performs a timely review of the data and deliver a report of the recommended actions needed for the data to become DCN-approved. Responsibilities: Perform and document each C-U-R-A-T-E step. **CURATE** steps are **C**heck files and read documentation. **U**nderstand the data (or try to), if not... **R**equest missing information or changes. **A**ugment metadata for findability. **T**ransform file formats for reuse. **E**valuate for FAIRness.	Curation Log Working Copy Inspect Files Inspect Metadata Documentation Create Metadata Quality Assurance Code Review File Format Transformations
5	Mediate	Role: DCN Coordinator Action: Mediates recommendations identified by the DCN Curator to the Local Curator. Responsibilities: • Tracks/monitors the submission review process. • Maintains email templates on typical actions/best practices. • Updates knowledge base as needed.	Communications with Local Curator
6	Support	Role: Local Curator Action: Works with researcher to address any changes, augmentations, or corrections to the data (in person, via email, etc.). Responsibilities: Level of local support will vary.	Communications with Author
7	Response	Role: Data Author Action: Respond to any curatorial issues and submits any files or changes to Local Curator as needed. Responsibilities: • Perform any changes and/or corrections to the data files and documentation. • Transfer the processed data files and documentation back to the repository.	Documentation Metadata Quality Assurance -Interoperability -Data Cleaning -Restructure File Format Transformations

8	Finalize	Role: Local curator Action: Finalize data submission. Responsibilities: ● Maintain integrity of the files and chain of custody throughout the curation process. ● Maintain all storage, access, dissemination, and preservation functions for the data going forward. ● Notify DCN Curator and DCN Coordinator of the final status of the data submission.	Secure Storage Terms of Use Rights Management Embargo Discovery Services -Full-Text Indexing -Metadata Brokerage -Use Analytics -Data Citation Versioning Succession Planning Tech. Monitoring and Refresh
9	DCN Approval	Role: DCN Coordinator Action: Review final data publication to determine if necessary actions were taken. If so, grants "DCN Approval." Responsibilities: ● Reviews FAIRness report and finalized dataset to determine if needed actions were taken. ● Certifies the data "DCN Approved" when applicable. ● Closes the ticket.	Final Inspection DCN Stamp of Approval

Appendix D: Functional Requirements for the DCN Tracking Form

The Data Curation Network will operate and function primarily through a tool or application that fulfills the requirements displayed in figure 16.

Figure 16: Functional requirements for a DCN tracking form

Project and project component features
1. Create 'tickets' for datasets requiring curation
2. Add and modify 'Templates' to the project or project components
3. Project/Ticket fields:
  a. Notes
  b. Subject - controlled vocabulary
  c. Data Format (excel, sql, geodatabase, text, matlab, etc.) - controlled vocabulary
  d. Institution - controlled vocabulary
  e. Contact information for data set creator (PI, email address)
  f. Status updates (Time stamps)
4. Assign individuals to projects or project components
5. Email and alerting at various points throughout the project (initiation, updates, closure)
	Users features/profiles:
	1. User login
	2. User types:
	  a. DCN Curator
	  b. Local Curator
	  c. DCN Representative
	3. Varying permission levels
	4. Profile fields:
	  a. subject or functional expertise
	  b. institution
	  c. existing projects in process
	  d. Projects completed

Templates for Curation Checklists
1. Multiple templates or checklists for curation activities or review
  a. General/first review checklist
  b. FAIR checklist
  c. Workflow checklist
  d. Excel data checklist
  e. SQL data checklist
  f. GIS data checklist
  g. Qualitative data checklist
  h. etc.
2. Multiple templates for email
  a. To Local Curator
	Infrastructure requirements
	1. Cross-institutional use
	2. Low-barrier learning curve
	3. Out-of-box functionality, or cloud-based product

Analysis of Available Options

There are numerous available tracking systems on the market, all having different strengths and weaknesses. Focusing specifically on workflow tracking software rather than project management, or IT service

management software, our initial evaluation shows that JIRA is a promising option for accomplishing many of our needed tasks. However, a combination of platforms to allow for both issue tracking and email integration may also have to be considered. A full analysis of software under consideration is shown in Figure 17.

Figure 17: Workflow and Issue Tracking Software and Tools Evaluation

Service/Tool	Additional Comments
Asana https://asana.com Free version supports just 15 users. $10/user/mo billed annually. Web-based.	Pros/Cons Does track workflow, but project management focused. Is possible to set up with little or no IT expertise needed on our part to get it set up.
Basecamp https://basecamp.com Web-based. $100/mo or $1000/year	Built for project management. Looks good in many ways, but no outside email integration.
Freshdesk https://freshdesk.com/ Not free.	All about "customer support".
JIRA https://www.atlassian.com/software/jira Hosted or server based. Not free.	Pros/Cons Super powerful, but has a bit of a learning curve. Works best when connected to Confluence, but that would incur additional costs. Has outgoing email integration/notification. Atlassian has cloud-hosting options that we could use.
OSF https://osf.io Free, open, web-based	Pros/Cons No email connectivity, but would potentially facilitate sharing of materials from existing storage locations.
OTRS https://www.otrs.com Web Based, Open Source	Full-service ITSM product, like Remedy. Probably more than we need.
RedLine http://cargocollective.com/superchen/filter/web-app/Redline-the-Visual-Bug-Tracker	Focus is on bug tracking in website development. Generates tickets and URLs for tracking.
Redmine http://www.redmine.org/ Open Source. Requires local install on any OS with Ruby on Rails.	Simple but complete features. Email integration. More focused on project management than customer service.
Remedy http://www.bmc.com/it-solutions/remedy-itsm.html Local install required	Not recommended for cross-institutional work; steep user learning curve, esp for non-full-time users
RT https://bestpractical.com/request-tracker/ Open source but not web-based.	Pros/Cons Will require a linux box and underlying relational database to run our own installation. Does track workflow.

Samanage https://www.samanage.com/ Not free	Very much designed to support IT service needs.
Trello https://trello.com/ Boards, Lists, Cards. Web-based.	Pros/Cons . People seem to love or hate it. May not do everything we need. Has email integration.
Waffle/Git https://waffle.io/ Github powered. Free.	Will really work only if we decide to keep everything else in GitHub as well.
Zapier https://zapier.com Web based, free version available.	Pros/Cons Has great app integration (eg. and email in GMail can trigger a file to be moved to Box and a message sent to Slack). Free version will not be adequate. Not sure it can do all our other required tasks.