ISSN 2161-3974

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Journal of eScience Librarianship
putting the pieces together: theory and practice

eScience in Action

Data Curation through Catalogs: A Repository-Independent Model for Data Discovery

Helenmary Sheridan[1], Anthony J. Dellureficio[2], Melissa A. Ratajeski[1], Sara Mannheimer[3], and Terrie R. Wheeler[4]

[1] University of Pittsburgh, Pittsburgh, PA, USA
[2] Memorial Sloan Kettering Cancer Center, New York, NY, USA
[3] Montana State University, Bozeman, MT, USA
[4] Weill Cornell Medicine, New York, NY, USA

Abstract

Institutional data repositories are the acknowledged gold standard for data curation platforms in academic libraries. But not every institution can sustain a repository, and not every dataset can be archived due to legal, ethical, or authorial constraints. Data catalogs—metadata-only indices of research data that provide detailed access instructions and conditions for use—are one potential solution, and may be especially suitable for "challenging" datasets. This article presents the strengths of data catalogs for increasing the discoverability and accessibility of research data. The authors argue that data catalogs are a viable alternative or complement to data repositories, and provide examples from their institutions' experiences to show how their data catalogs address specific curatorial requirements. The article also reports on the development of a community of practice for data catalogs and data discovery initiatives.

Correspondence: Terrie R. Wheeler: tew2004@med.cornell.edu

Received: April 7, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Disclosures: The authors report no conflict of interest.

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Introduction and situation of the problem

Over the past several years, the number of academic libraries offering research data curation services has grown (Hudson-Vitale et al. 2017). Often this service coincides with the library hosting a standalone data repository (e.g., the Illinois Data Bank at the University of Illinois at Urbana-Champaign (University of Illinois 2021)) or an institutional repository that accepts data (Hudson-Vitale et al. 2017). However, a repository is not requisite for robust data curation services. A growing number of libraries have opted to implement data catalogs instead of or in addition to repositories to maximize discovery of their researchers' datasets.

Defined broadly, a data catalog is a curated collection of metadata records that describe and point to data products of interest. Data catalogs do not archive research datasets. Instead, they focus on increasing the discoverability of those datasets, "discoverable" being synonymous with "findable"—the F in the FAIR Data Principles (Wilkinson et al. 2016). But they are not aggregators like Google Dataset Search, which harvests and displays metadata directly from the web (Noy 2020): data catalog records are created, checked, and updated by professionals (frequently data librarians) through multi-step workflows that ensure their metadata is accurate and understandable to both humans and machines. By focusing solely on curating datasets for online discovery, data catalogs support the discovery, citation, and reuse of research data.

The reasons why libraries may develop a data catalog are myriad. Motivations from the authors' experiences include:

Capacity

- Storing and preserving datasets in all their different formats and sizes is not feasible for many academic libraries that may still wish to enable access to data. Data catalogs may be more sustainable.

Uncovering datasets that are otherwise hidden

- Data catalogs can describe high-value datasets that, due to access protocols, are not available anywhere else. For example, the NYU Data Catalog provides the authoritative description and access point for the protected Neurological Emergencies Outcomes at NYU (NEON) dataset (New York University 2021).
- Data catalogs can act as the point of access for datasets whose creators are reluctant to share their data in a public repository but willing to share upon request. The conversation with authors required to describe these private datasets also affords data catalogers an opportunity to suggest improvements to the data package, like the creation of READMEs.
- They can facilitate access to secured data by describing data governance, providing access instructions, and linking to request forms.
- Data catalogs can support research transparency by providing metadata for publicly-funded sensitive data that cannot be put online in full.

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Fostering collaboration in scientific and institutional communities

- Some data catalogs describe externally-created datasets that are widely used for secondary analysis (e.g., large national surveys) and add value by naming a "local expert" who is willing to act as a contact/collaborator/ mentor for institutional researchers (Read et al. 2015).
- Data catalogs can serve as a source for educational or training material, especially if an institutional author is available to answer questions about the data. The metadata fields in a data catalog can be used to parse out large datasets suitable for training algorithms, for example, or for practicing analysis using particular scripting languages.
- Data catalogs can complement community infrastructure that a library may already support. Institutional members of a community repository, e.g. Dryad Data Repository, can continue to deposit data there while the data catalog enhances its discoverability.
- Data catalogs allow datasets to be archived where they are most likely to be found and used. Data are more likely to be cited if they are archived in a disciplinary repository and indexed in multiple locations (Mannheimer, Sterman, and Borda 2016); a catalog record for a dataset hosted elsewhere contributes an additional index point while making the institution's relationship to the dataset explicit.
- Libraries can co-locate all datasets created at the institution in one institutional data catalog, which serves as an institutional data inventory or marketplace.

Complying with data governance requirements for sensitive data

- A data catalog can be set up to capture who is authorized to access confidential data according to data governance. Weill Cornell Medicine, for example, has integrated its catalog with its Data Core (secure data enclave) management system (Oxley 2020). This allows the institution to rapidly review requests for access and monitor changes in authorization for projects and datasets. By developing a data catalog that focuses on governance metadata, the organization has advanced its ethical responsibilities towards the handling of confidential data by both enhancing visibility of government constraints, and increasing the research value obtained from those data sets.
- Data catalogs can provide an audit trail for datasets containing clinical patient data. Weill Cornell's data-enclave integration tracks each dataset's initial registration, the purposes for which it is being used, who accesses it, and the conditions of users' authorization, helping to ensure patient confidentiality (Oxley et al. 2018).

Integrating with existing library and campus infrastructure

- In situations where researchers have deposited data products in multiple locations, data catalogs can pull together related material in one record. This also applies to data products hosted in campus enterprise services. A

hypothetical complex record could include structured metadata describing and linking to:

  - Dataset (processed data) in the organization's institutional repository
  - Large dataset (raw data) in the organization's Globus instance or other high-volume file transfer platform
  - Analysis code in Github
  - Registered protocol at protocols.io.

- Data catalogs provide an additional discovery layer for institutional repositories, which were viewed favorably by users in one study for their preservation functions but less so for "searchability" (Shen 2017, 120).
- Data catalog records may be indexed in a general library catalog by using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to extend a discovery system's ability to surface datasets.

Specifics of data curation activities within catalogs

As nexuses for discovery, data catalogs describe data products with standards-compliant metadata that creates linkages among related records and serves structured data to search engines. They are searchable and browsable, although the "how" varies according to each platform's architecture. In the Data Discovery Collaboration (DDC), to which the authors and their institutions belong (Data Discovery Collaboration 2021), defined metadata fields include but are not limited to: dataset title(s), creator(s), data and resource type, description, keywords, file size, software and equipment used to collect or analyze the data, funding information, and associated publications. Metadata entries are drawn from controlled vocabularies and URI-referenced whenever possible. Crucially, records include instructions for accessing the data, which may involve submitting an application or contacting the author. Institutions vary in the specific metadata schemas and standards implemented in their catalogs, but Datacite, DATS, Dublin Core, and schema.org are represented among the catalogs in the DDC.

The focused scope of a data catalog allows data catalogers to act nimbly, beginning with identifying data to describe. Information about a dataset that is a candidate for cataloging may enter an institution's workflow in any of three general ways (Ratajeski et al. 2019):

a) Submitted by the dataset creator, either because the researcher discovered the catalog independently or because s/he was solicited by a cataloger (e.g., through a "cold email" asking whether they had any datasets they would be willing to share, which may uncover datasets otherwise unavailable.)

b) Identified and cataloged by a cataloger, with varying degrees of participation from the dataset creator. A dataset deposited to a major

repository with full documentation may require very little, if any, further information from the author, while a dataset identified only through a paper's "data available upon request" statement may require much more communication.

c) Automated by using strategies such as harvesting data repository APIs, then enhanced by a data cataloger. Examples include Montana State University's Dataset Search, which finds datasets by institutional authors in external data repositories and presents them to data catalogers for manual review and record creation (Mannheimer et al. 2021).

Path a) resembles the workflow of many repositories, particularly large-scale and domain-nonspecific repositories, where staff may promote their repository's services but rarely make specific collection requests. Paths b) and c), in contrast, take a proactive approach to discovering datasets. Path b) casts the data cataloger in the role of the eventual data consumer, faced with the task of finding relevant data like the proverbial needle in a haystack—but with expert searching skills to help. Path c) lessens that burden with automation, providing the cataloger with a list of likely candidates (of datasets, publications with data availability statements, or simply known data-producing authors) winnowed from systems like faculty information systems, REDCap reports, or PubMed article alerts.

An examination of a data catalog's activities using the Data Curation Network's C-U-R-A-T-E-D framework (Data Curation Network 2018) will illustrate where the catalog's curation energies go. In this framework, each of the letters in the word 'CURATED' stand for a step in the DCN's data curation process. Many of the steps are analogous to steps in the data catalog curation process, with a key difference in focus. Data Curation Network stewards aim to improve the quality of a dataset and its accompanying metadata for submission to a repository. Data catalog stewards focus almost entirely on improving or creating entirely new metadata, as "hidden" datasets (e.g., available only upon request from the author) rarely have author-supplied metadata already.

Note that the details of each step below will vary among data catalog-maintaining institutions. Note too that since many data catalogers are data librarians, they may also have separate conversations with researchers about their datasets' quality, especially if they are submitting data to a repository to which a data catalog record might then link.

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Table 1: Comparison of data catalog curation activities and the Data Curation Network's C-U-R-A-T-E-D steps

DCN C-U-R-A-T-E-D Step	Data catalog activity
C: Check data files and read documentation	Somewhat similar. Catalogers read documentation and examine data files to create a high-quality metadata record; however, they do not check files for completion, quality, or file integrity
U: Understand the data	Somewhat similar. Catalogers try to understand the data enough to describe it, but do not comment on the data files unless also offering advice prior to submission to a repository
R: Request missing information or changes	Similar. Catalogers ask for more information to create a metadata record, and may suggest that the authors create documentation
A: Augment with metadata for findability	Very similar. Catalogers create descriptive metadata, incorporating author-supplied terms when possible, and source metadata from controlled vocabularies for interoperability
T: Transform file formats	Does not apply, although catalogers can make recommendations for the data stored elsewhere
E: Evaluate and rate for FAIRness	Does not apply; although catalog staff may have their own checklist for acceptable metadata records, they do not control the data themselves
D: Document throughout curation activities	Somewhat similar. Since no actual datasets are changing hands, submission agreements and chain-of-custody documentation are unnecessary, but institutions may have their own cataloging workflow requirements. The open-source code developed by NYU keeps a basic log of editing dates made to records, and catalogers have the option of adding detailed edit notes

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Setting standards for data catalogs

Traditional card catalogs and their online public access catalog successors, aiding in the discovery of individual items in a collection, have been a mainstay of libraries for over a century. Libraries that support data catalogs, however, are relatively new and few in number. In 2017, several academic health science libraries organized into a loose network called the Data Catalog Collaboration Project (DCCP) to offer support and community for those working toward the shared goal of enhancing the findability of datasets. Each member institution (some with funding from the Network of the National Library of Medicine) indexed their biomedical research data with local instances of an open-source data catalog platform created at the founding member institution, New York University (Lamb and Larson 2016). The DCCP brought a cross-institutional perspective to addressing usability, data sharing workflows, metadata, and outreach to improve data discovery.

At a February 2020 retreat, current and potential institutional members met to reassess the intent and priorities of the group. Among the outcomes were a new name, the Data Discovery Collaboration (DDC); the creation of a steering committee; inclusion of non-health sciences institutions; and a purposeful shift towards a broader, platform-agnostic approach towards data discovery that would be focused on developing standards and best practices. The Mission Statement of this reimagined organization reads: "To enhance discovery of data and other research products in order to maximize their value" (Data Discovery Collaboration 2021).

In its new iteration, the DDC enables data discovery in its broadest forms through a governance structure designed to encourage participation from member organizations through working groups, listserv discussions, and Slack channel conversation. Data catalogs are no longer a requirement for membership, but they remain a central topic in addition to metadata creation, interoperability, and code sharing. Table 2 summarizes the current member institutions and their data catalog platforms:

JeSLIB 2021; 10(3): e1203
https://doi.org/10.7191/jeslib.2021.1203

Table 2: DDC members and their data catalogs as of June 2021

Organization Name	Link to Catalog	Data Catalog Software
New York University, NYU Langone Health	NYU Data Catalog	Open-source code developed by NYU
Northwestern University, Feinberg School of Medicine, Galter Health Sciences Library & Learning Center	DigitalHub (integrated with IR; new service coming late 2021)	Fedora/Sufia (new service will run on Python/Flask)
Memorial Sloan Kettering Cancer Center	MSK Data Catalog	Open-source code developed by NYU
Montana State University	Dataset Search	Developed in-house
University of Maryland, Baltimore, Health Sciences and Human Services Library	UMB Data Catalog	Open-source code developed by NYU
University of Pittsburgh, Health Sciences Library System	Pitt Data Catalog	Open-source code developed by NYU
Wayne State University	Wayne State Data Catalog	Open-source code developed by NYU
Weill Cornell Medicine	WCM Data Catalog	Developed in-house
Zucker School of Medicine at Hofstra/Northwell	Hofstra/Northwell Data Catalog	Open-source code developed by NYU

The Data Discovery Collaboration welcomes involvement with both potential new members (individuals and institutions) and unaffiliated organizations who share its goal of supporting data reuse by increasing discoverability of data products. The collaboration's core groups are currently working on projects such as building out metadata elements for basic/bench science data; packaging the NYU Data Catalog code into a Docker container for easier installation; recruiting an advisory board to help the DDC navigate the technological and social aspects of data discovery; and sharing strategies for promoting data sharing and reuse within our institutions. To join the conversation, please contact any of the authors or visit the Data Discovery Collaboration website (https://datadiscoverycollaboration.org).

Though less well-known than their repository relatives, data catalogs are a powerful tool for curating research data in the library. Their focus on data discovery makes them ideal candidates for a wide range of settings and purposes, as shown by just some of the diverse use cases presented in the Data Discovery

Collaboration: they can stand alone or be an add-on, describe data by or data for institutional researchers, and make data public or help keep it secured. As the community of practice grows, catalogs may evolve to meet needs yet unseen.

Acknowledgements

The authors would like to thank Ian Lamb for writing the software code for the original data catalog that was used by the Data Catalog Collaboration Project (DCCP) which became the DDC. The authors also thank Nicole Contaxis, whose leadership has enabled the DCCP to transition to the DDC.

References

Data Curation Network. 2018. "Checklist of CURATED Steps Performed by the Data Curation Network." http://z.umn.edu/curate

Data Discovery Collaboration. n.d. "The Data Discovery Collaboration." Accessed June 28, 2021. https://datadiscoverycollaboration.org

Hudson-Vitale, Cynthia, Heidi Imker, Lisa R. Johnston, Jake Carlson, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2017. *Spec Kit 354: Data Curation*. Washington, DC: Association of Research Libraries. https://doi.org/10.29242/spec.354

Lamb, Ian, and Catherine Larson. 2016. "Shining a Light on Scientific Data: Building a Data Catalog to Foster Data Sharing and Reuse." *Code{4}Lib* 32(April). https://journal.code4lib.org/articles/11421

Mannheimer, Sara, Jason Clark, Kyle Hagerman, Jakob Schultz, and James Espeland. 2021. "Dataset Search: A Lightweight, Community-Built Tool to Support Research Data Discovery." *Journal of eScience Librarianship* 10(1): e1189. https://doi.org/10.7191/jeslib.2021.1189

Mannheimer, Sara, Leila Belle Sterman, and Susan Borda. 2016. "Discovery and Reuse of Open Datasets: An Exploratory Study." *Journal of eScience Librarianship* 5(1): e1091. https://doi.org/10.7191/jeslib.2016.1091

Noy, Natasha. 2020. "Discovering Millions of Datasets on the Web." *The Keyword* January 23, 2020. https://blog.google/products/search/discovering-millions-datasets-web

NYU Data Catalog. n.d. "Neurological Emergencies Outcomes at NYU." Accessed June 28, 2021. https://datacatalog.med.nyu.edu/dataset/10330

Oxley, Peter R., John Ruffing, Thomas R. Campion, Jr., Terrie R. Wheeler, and Curtis M. Cole. 2018. "Design and Implementation of a Secure Computing Environment for Analysis of Sensitive Data at an Academic Medical Center." *AMIA Annual Symposium Proceedings* 2018: 857–866. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371349

Oxley, Peter R. 2020. "Data Core Web Manager." https://github.com/oxpeter/datacore_web_manager

Ratajeski, Melissa A., Katherine Akers, Nicole Contaxis, Megan D. Del Baglivo, Na Lin, Helenmary Sheridan, and Kevin Read. 2019. "Outreach Strategies and Researchers' Motivations for Sharing Data through a Data Catalog." Poster presented at *Medical Library Association Annual Conference*, Chicago, IL, 2019. http://jmla.mlanet.org/ojs/jmla/article/downloadSuppFile/897/1682

Shen, Yi. 2017. "Burgeoning Data Repository Systems, Characteristics, and Development Strategies: Insights of Natural Resources and Environmental Scientists." *Data and Information Management* 1(2): 115–123. https://doi.org/10.1515/dim-2017-0009

Read, Kevin, Jessica Athens, Ian Lamb, Joey Nicholson, Sushan Chin, Junchuan Xu, Neil Rambo, and Alisa Surkis. 2015. "Promoting Data Reuse and Collaboration at an Academic Medical Center." *International Journal of Digital Curation* 10(1): 260–267. https://doi.org/10.2218/ijdc.v10i1.366

University of Illinois. n.d. "Illinois Data Bank." Accessed June 28, 2021. https://databank.illinois.edu

Wilkinson, Mark D., Michel Dumontier, IJsbrand Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The Fair Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(160018). https://doi.org/10.1038/sdata.2016.18