ISSN 2161-3974

JeSLIB 2021; 10(3): e1222
https://doi.org/10.7191/jeslib.2021.1222

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

Editorial

Introduction to the Special *JeSLIB* Issue on Data Curation in Practice

Cynthia Hudson Vitale, *Director, Scholars and Scholarship*
Association of Research Libraries, Washington, DC, USA

Jake R. Carlson, *Director of Deep Blue Repository and Research Data Services*
University of Michigan - Ann Arbor, MI, USA

Hannah Hadley, *Data Curation Education Librarian*
The Pennsylvania State University, University Park, PA, USA

Lisa Johnston, *Research Data Management/Curation Lead and Co-Director of the University Digital Conservancy*
University of Minnesota - Twin Cities, MN, USA

Focus

Research data curation is a set of scientific communication processes and activities that support the ethical reuse of research data and uphold research integrity. Data curators act as key collaborators with researchers to enrich the scholarly value and potential impact of their data through preparing it to be shared with others and preserved for the long term. This special issue focuses on practical data curation workflows and tools that have been developed and implemented within data repositories, scholarly societies, research projects, and academic institutions.

Correspondence: Cynthia Hudson Vitale: cvitale@arl.org

Received: July 27, 2021 **Accepted**: August 2, 2021 **Published**: August 11, 2021

Disclosures: The authors report no conflict of interest.

JeSLIB 2021; 10(3): e1222
https://doi.org/10.7191/jeslib.2021.1222

The last 10 to 20 years have seen a significant shift in how research is designed, recorded, and eventually shared to a global community of scholars. This, in turn, has impacted how academic libraries and archives support research locally. Institutions, through their libraries in particular, have invested in staffing, infrastructure, and workflows for the management, stewardship, curation, and preservation of research data created or reused on our academic campuses. One particularly important area of research data services that has gained momentum is research data curation.

Research data curation is a set of scientific communication processes and activities that support the ethical reuse of research data and uphold research integrity. Data curators act as key collaborators with researchers to enrich the scholarly value and potential impact of their data through preparing it to be shared with others and preserved for the long term. Curators possess expertise in a variety of areas, including disciplinary standards and practices, file formats, and the organization of information, to name a few.

However, the expertise needed to curate the wide variety and complexity of research outputs produced within an institution is broader than any one individual or small team could reasonably address. In 2016, six librarians got together to explore how they could work in closer collaboration with one another to strengthen each of their program's research data curation services. The result was the Data Curation Network (DCN), a radical collaboration across institutions to extend access to the advanced expertise of data curators at one member institution to all of its members. The DCN was initially funded by the Alfred P Sloan Foundation and most recently has transitioned to a sustainable member-driven organization with 14 member institutions and over 50 data curators within its active pool. Similar projects are being implemented worldwide, such as the Canadian Portage Network, whose bi-lingual guide for curation in Dataverse has built upon the CURATE(D) protocol created by the DCN. An introduction to this guide is included in our issue and illustrates the power of radical collaborations as a means of extending our shared understanding of curation processes.

The Education initiative within the DCN launched the development and delivery of a series of Specialized Data Curation Workshops with the help of funders, such as the Institute of Museum and Library Services. Participants in these sessions formed working groups to create Data Curation Primers, concise resources that provide essentials for curating specialized data formats or topic-specific considerations for research data. The shared experience of attending these trainings and collaborating on the creation of Primers has created an extended community of interest and expertise in data curation. The formation of this community has led to talks on how to further stimulate discussion and collaboration for curatorial technique and eventually resulted in our joint effort with JeSLIB to create this special issue on data curation practices.

Included within this issue are the introductions of two Data Curation Primers. Authors of a Neuroimaging Data Curation Primer provide a video presentation for

the resource that discusses curation of functional magnetic resonance images that are saved in either DICOM or NIfTI formats. The resource was created by health science librarians and informationists as part of a working group that formed from the Data Curation Network's IMLS funded Specialized Data Curation Workshops and was published earlier this year. The video shows an example workflow and gives highlights from the resource. Additionally, a newly completed Plain Text and Character Encoding Primer takes a direct approach to ensuring that textual encoding errors are well understood and resolvable. It also takes the editorial stance that "plain text isn't so plain." This primer exemplifies the DCN's intent to have the Primers serve as practical guides and actionable tools for data curators to use in their work.

Similarly, this issue includes "Data Curation in Practice: Extract Tabular Data from PDF Files Using a Data Analytics Tool" describing a useful curation tool and its application. The authors take a practical approach in describing a feature available in a Microsoft application (Microsoft Power BI) that could be applied to data curation work. The paper presents the tool and a technique for curators to use when extracting tabular data from a PDF document. PDFs are ubiquitous and they often contain data that are hard to access as "data." The description focuses on explaining a specific functionality, conveying what the feature does, how to use it and what to pay attention to in the process.

Data curation is a key component in enabling research reproducibility. To that end, data curators are seeking out tools and methods for addressing computational reproducibility. "Computational Reproducibility: A Practical Framework for Data Curators" presents a framework for computational reproducibility that is scaled or tiered based on expertise or capacity of the institution and curator. Enabling research reproducibility can be a daunting prospect for curators, particularly those that are early in their careers. The authors stress the importance of simply getting started with considering reproducibility as a part of curation work and to not be constrained by needs of perfection.

The experience of an individual curator is given in "An Insider's Take on Data Curation: Context, Quality, and Efficiency" providing a first-person perspective on curating research data and the importance the enhancements support. A curators' responsibility for ensuring the proper application of curation workflows are further explained in the article, which is inclusive of protecting the safety and privacy of respondents and assuring the research data are of high quality.

Workflow descriptions are further given in examples for new ways of curating. "Implementing and Managing a Data Curation Workflow in the Cloud" discusses a curation workflow for an institution using Figshare as their repository platform. The infrastructure is described in the context of the curatorial work. Human components of curation are highlighted and contrasted with what is automated. Two other papers explore curation throughout the creation of research outputs. In "(Hyper)active Data Curation: A Video Case Study from Behavioral Science," the Databrary repository describes infrastructure and tools for its curation services,

which were developed in coordination with the PLAY Project, a research initiative to actively curate video data. Similarly, "Active Curation of Large Longitudinal Surveys: A Case Study" presents a curation example involving longitudinal surveys. This manuscript describes what the authors call "active curation" which is defined as taking actions to improve the utility of the data throughout the active stages of the data lifecycle.

Curation workflows are also the subject of "Responding to Reality: Evolving Curation Practices and Infrastructure at the University of Illinois at Urbana-Champaign." In this paper, the UIUC Research Data Service team analyzed their service metrics as part of a five-year review of their program. They found that data curation is the most valued service offered by their program and break down some of the challenges and successes of this work. The article goes into detail about their curation workflow and how they tackled issues such as large file transfer, cloud services, and encouraging researcher collaboration in the curation process. Additionally, this program is placed in the context of the relationships with campus units, specifically the Provosts' office. Overall, this is a useful case study of a mature curation service provided by an academic library.

Robert Demgenski, Sebastian Karcher, Dessi Kirilova, and Nic Weber from the Qualitative Data Repository (QDR) provide an example of developing service maturity through the creation of their handbook. They explain their rationale for developing a handbook, describing their needs in developing and operating QDR's curation services and the challenges they face in doing so. They then describe the content of the handbook and how it is used within QDR. The needs and challenges described by QDR are broad enough to be relatable to other curation service providers.

Lastly, the article, "Data Curation through Catalogs: A Repository-Independent Model for Data Discovery" convincingly argues for data catalogs as a core service in a research data program, after all, not every institution can sustain a data repository, and not every dataset can be archived due to legal or technical barriers. Curation of a data catalog involves author self deposit, but also curator selection or ingest through APIs. The authors map these actions to the CURATE(D) framework used by DCN curators in guiding their work. Finally, the authors discuss the community of practice that has been created for data catalog institutions. This article broadens the understanding of data curation to be inclusive of data discovery and access via data catalogs.

In putting together this special issue, we wanted to bring together an international community of data curators and services to describe their curatorial practices. We hope that in doing so, attention and recognition can be brought to this important component of data services. Additionally, we can learn much from one another and keep the door open to future collaborations that will benefit the global reach of open science.