ISSN 2161-3974

JeSLIB 2021; 10(3): e1207
https://doi.org/10.7191/jeslib.2021.1207

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

Full-Length Paper

Introducing the Qualitative Data Repository's *Curation Handbook*

Robert Demgenski[1], Sebastian Karcher[1], Dessi Kirilova[1], and Nic Weber[1,2]

[1] Qualitative Data Repository, Syracuse University, Syracuse, NY, USA
[2] University of Washington, Seattle, WA, USA

Abstract

In this short practice paper, we introduce the public version of the Qualitative Data Repository's (QDR) *Curation Handbook*. The *Handbook* documents and structures curation practices at QDR. We describe the background and genesis of the *Handbook* and highlight some of its key content.

Correspondence: Sebastian Karcher: skarcher@syr.edu

Received: April 8, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Copyright: © 2021 Demgenski et al. This is an open access article licensed under the terms of the Creative Commons Attribution License.

Data Availability: This article describes QDR's curation handbook, available under CC-BY license as Demgenski, Robert, Karcher, Sebastian, Kirilova, Dessi, and Weber, Nic. 2021. "QDR *Curation Handbook*." Syracuse University: Qualitative Data Repository. https://doi.org/10.5281/zenodo.4672678

Disclosures: The authors report no conflict of interest.

Introduction

Launched in 2014, the Qualitative Data Repository (QDR) has been at the forefront of the open science movement with a focus on enabling appropriate and ethical data sharing in qualitative and multi-method research across the social sciences. QDR's research program attempts to break new ground in the field of data curation by developing tools, guidance, and resources – often addressing the particular demands of transparency in qualitative research. QDR is located at Syracuse University's Maxwell School of Citizenship and Public Affairs, and maintains a variety of research infrastructures to curate, publish, and serve qualitative researchers. This includes both a technical infrastructure for data deposit and preservation based on the Dataverse repository software, as well as a human infrastructure of curators who perform various tasks necessary for "purposeful work" (Palmer et al. 2013) with data. The curation services offered by QDR are aimed at preparing data projects that other scholars can discover, evaluate, and responsibly re-use for secondary research as well as teaching (Karcher et al. 2021).

Comprehensive and interactive curation has been at the core of the repository operations since the beginning, encompassing a wide variety of tasks. Some of these (e.g., DOI assignment or indexing for searchability) are not available to individual researchers, while others are things anyone can do in theory (e.g., developing and applying consistent file naming conventions within a project or documentation) even if researchers often do not or find unsatisfactory (Johnston et al. 2018).

In all cases, the goal of the curation staff in preparing a set of materials for publication is to work closely with depositors to guide them toward providing the most useful version of the research data they collected or produced. The repository's curation framework is one tailored to serve the specific user community the repository serves, but based on general principles of archival and information science. In practice it also requires general understanding of the project's domain and methods, a critical eye toward ethical and legal commitments that data sharing might impinge on, unwavering attention to detail, and generally takes place over at least a few weeks. This necessitates coordination across the staff, adhering to a common set of instructions and meticulous record-keeping of what additional decisions were made or which prescribed steps were not relevant in a given project.

QDR decided to create a comprehensive *Curation Handbook* to support and document its internal operations in early 2020. Based on seven years of experience in curating qualitative data, the *Handbook* records, in detail, how QDR has adopted, developed, implemented, and adapted data curation standards and practices for qualitative data. It aims to cover the entire data curation lifecycle, from an initial consultation between repository staff and depositors to post-publication processes and the long-term preservation and dissemination of data (see Figure 1).

Figure 1: The *Curation Handbook*'s table of contents.

In an unintentional yet foreseeable way, the *Handbook* has relevance to all three "legs" of productive data curation: organizational infrastructure, technological infrastructure, and requisite resources. (McGovern 2007, referring to digital curation; see Palmer et al. 2013 on the confluence and overlap of these terms.)

The QDR *Curation Handbook* captures all these core aspects of the repository's work: it is an attempt most clearly to reflect the *organization* of complex and interrelated processes in a coherent work whole; it interfaces with newly developed *technological tools* that automate the most repetitive and laborious steps of qualitative data curation; and it indirectly serves to conserve and maximize the labor and financial *resources* of the institution.

We are now sharing a public version of this *Handbook* (Demgenski et al. 2021).[1] It differs from QDR's internal *Curation Handbook* only in the absence of internal administrative notes and in format—the internal *Handbook* is a Google document,

1 Available from Zenodo: https://doi.org/10.5281/zenodo.4672678

meant to continuously evolve and be improved upon as we encounter new scenarios and find ways to improve existing workflows or incorporate evolving standards. The shared version of the *Handbook* is a snapshot of our processes at the time of publication.

In the remainder of this note, we describe the *Handbook*'s general objectives, both for internal purposes and for this published version and its role in our ongoing effort to provide the highest quality of data curation services in an efficient, sustainable, and cost-effective manner. We conclude by highlighting three of the key elements of QDR's data processing as documented in the *Curation Handbook*—the accompanying GitHub-based tracking system for curation tasks, our use of scripting and automation in the curation workflow, and how we handle data with various types of restrictions.

General Objectives of the *Curation Handbook*

We outline below the initial objectives in developing the *Handbook* as well as what we hope to achieve with the published version.

Internal Objectives

QDR has faced two broad challenges since its inception—one inherent to its mission and the other of organizational nature. QDR operates not only in the context of a nascent open science movement, but focusing on an area—qualitative data curation—with little previous work (especially in the US). As a result, QDR has had few precedents to learn from or adopt, not only in terms of curation standards but specific practices—the nuts and bolts of curation operations (Elman and Kapiszewski 2014; Karcher et al. 2016).[2] Over time, QDR's staff developed expert knowledge accumulated through experience, research, and interaction with community stakeholders. As this body of knowledge and routine practices have become more complex, the need for consolidation and codification has increased. The second challenge is born out of QDR's organizational structure, with many permanent staff involved in curation in a part-time capacity and graduate assistants (GAs), who perform large parts of the hands-on curation work, being subject to regular turnover. The latter poses a particular problem in terms of knowledge loss the organization incurs with each departing GA and the coinciding need for resource-intensive training periods for new GAs, issues compounded the more sophisticated curation processes become.

With the creation of the *Curation Handbook*, we attempted to support QDR in facing both those challenges by achieving four internal objectives.

1. Consolidate the body of curation knowledge accumulated over time into one document to support standardization and codification of QDR's curation practices.

2 The largest existing collection of shared qualitative data, and a significant source of initial expertise and guidance for QDR, is the UK Data Service, which began archiving qualitative data in the 1990s (see, e.g., Corti 2000, 2006).

2. Increase organizational knowledge retention by structuring the *Handbook* in such a manner that, even as it standardizes procedures, it remains highly flexible for further improvements.
3. Serve as a training tool for new GAs by covering the entire curation process in such detail that one could, with limited or no prior experience in data curation, curate most qualitative data projects relying on the Curation *Handbook*, with minimal outside assistance.
4. Serve as a curation tool that remains useful even for experienced data curators and can be referred back to continuously.

External Objectives

While it was initially developed to serve internal operations exclusively, we believe there is significant value in sharing this published version of the *Handbook*. The purpose here is threefold:

1. Provide an additional layer of transparency to QDR's internal operations, inviting scrutiny and any suggestions to improve our processes.
2. Serve as a resource for qualitative researchers interested in what qualitative data sharing and qualitative data curation "in practice" entails, assisting them as they consider the best ways to manage their data.
3. Contribute to the pool of knowledge for the growing community of qualitative data curators, in the hope of generating discussions and knowledge-sharing to improve qualitative curation standards and optimize practices. It thus complements recently published "Data Curation Primers" for qualitative data (Corral 2019; Hadley 2019; Castillo, Coates, and Narlock 2020).

Process Optimization in Qualitative Data Curation

The *Handbook* encompasses the entire curation process—everything from templates for communicating with depositors, code scripts for software-assisted data curation, sensitive data handling, copyright review, workflow instructions, file-level and project-level metadata, data publication and post-publication tasks. In taking this comprehensive approach, we want to ensure that the curation process is both effective (i.e., achieving the desired level of data curation standard) and efficient (i.e., working as sustainably as possible).

Doing Qualitative Data Curation Right

The *Handbook* details how QDR aims to render each data project as close as possible to the ideal of the F.A.I.R data principles (Wilkinson et al. 2016) and orients its curation toward long-term preservation and enhancing reuse possibilities. Basic standard procedures are set. For instance, each project undergoes, in consultation with the depositor(s), an ethical and legal review to

ensure that the data can be shared in the first place, and whether special procedures or restrictions need to be implemented, such as reviewing de-identified human-participant data for disclosure risks, evaluating the copyright status of data, or restricting access to the data (sections 2.3 and 3.5). Scanned textual documents undergo OCR (Optical Character Recognition) to enable full-text search (section 2.8). All files curated by QDR are examined for bit-level integrity, converted to appropriate archival formats when necessary (sections 2.4. and 2.9), and assigned file-level metadata (section 2.7).

In addition to these highly standardized procedures, qualitative data curation includes a myriad of peculiarities that do not easily lend themselves to standardized approaches that can be brought up to scale. QDR continues to receive projects requiring the formulation of new policies and procedural or technological innovations—whether related to data sensitivity, copyright compliance and other legal considerations, data formats, or other issues. Yet, amidst all these differences, we believe the *Handbook* identifies enough common denominators to ensure that, for the vast majority of projects we receive, the curation process is kept on the "right" track.

Doing Qualitative Data Curation Efficiently

In order to deliver on the promise of long-term preservation, QDR also needs to ensure sustainability in the curation process. The *Handbook* includes a variety of procedures developed over time that enable us to reduce the amount of resources required to curate data projects, shorten project turn-around time, reduce the risk of errors, and enable us to curate large projects as well, with over a thousand data files (e.g., Loyle et al. 2018; Trachtenberg 2020). This is primarily done with the aid of software and scripts, both external and developed in-house (outlined in section 2.1 and discussed further below), but also with the organization and standardization of workflows broken down into repeatable tasks.

Key Features of the QDR *Curation Handbook*

Spanning almost 40 pages (in addition to accompanying software packages and scripts) and detailed descriptions of QDR workflows, even a summary of the *Handbook*'s content would exceed the length of a short introduction. Instead, we highlight here three of its key features that we believe best showcase QDR's approach to curating qualitative data.

GitHub-based Checklists

Checklists are widely used tools to handle complex tasks ranging from aviation, to surgery, and construction (Gawande 2010). Data curation includes a fairly large number of semi-standardized tasks, often performed by a team, and therefore lends itself well to a checklist, and several such checklists exist (e.g. DCC 2009, DCN 2018, Karvovskaya 2019). QDR uses a set of task-specific checklists for the key components of the curation process: initial assessment, metadata and

JeSLIB 2021; 10(3): e1207
https://doi.org/10.7191/jeslib.2021.1207

documentation, file processing, and publication. Each checklist is an issue on the GitHub platform, generated from templates in a private repository and added to a project board. The Kanban-style project board (see Figure 2) provides a quick overview of the project status for the curation team. The individual issues hold, in addition to the checklist items, any additional observations, communication, and decisions made during curation, and thus serve as both a record of curation activities and a point of reference for other curators working on a project. The four issues and the project board for a new data project are created automatically using the `dvcurator` R package at the beginning of curation (see section 2.2 of the *Handbook*). The checklists follow the same logic as the *Curation Handbook*: details on most individual items can be found in the *Handbook*.

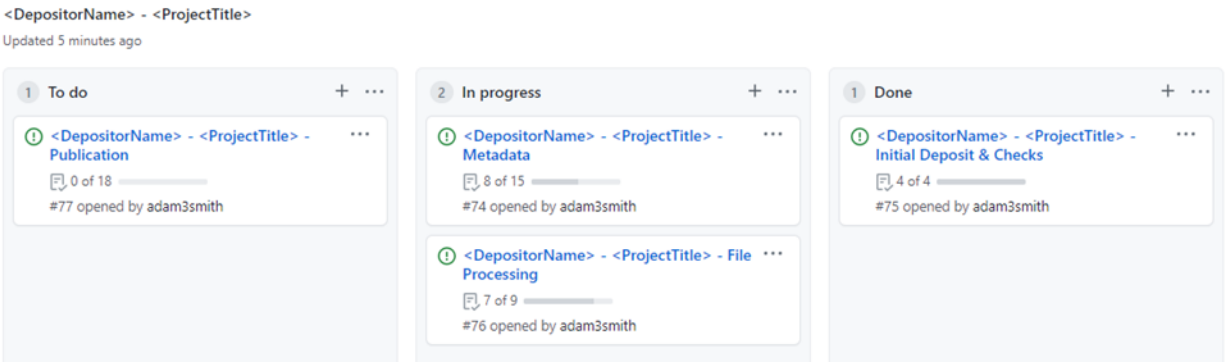
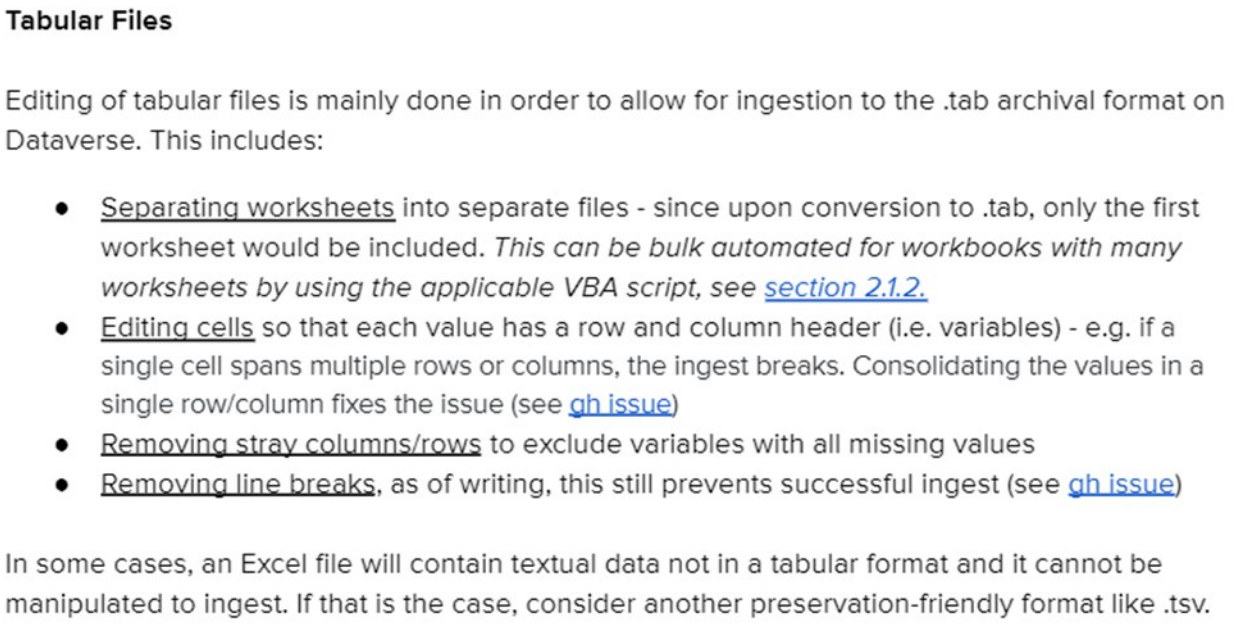


Figure 2: GitHub Project Board for a (fictive) QDR deposit in progress.

Scripts and Automation

The deposit and sharing of qualitative data is comparatively novel, and only few of QDR's depositors have any experience sharing qualitative data prior to depositing with QDR. Additionally, concerns about confidentiality and ethics of sharing human participant data feature heavily in debates about qualitative transparency (Bishop 2009; Kapiszewski and Wood 2021; Yardley et al. 2014). Close scrutiny of data for possible (inadvertent) violations of ethics and confidentiality is thus an essential part of curating qualitative data. QDR's work is labor-intensive. At the same time, labor is expensive and, as any data repository, QDR faces economic constraints (see, e.g., Eschenfelder and Shankar 2017; OECD 2017).

Without compromising on the human element of curation, which will remain indispensable for qualitative data, we seek to automate labor-intensive and repetitive tasks as much as possible in what we have termed "human-in-the-loop curation" (Weber, Karcher, and Myers 2020). The *Handbook* contains references to a number of scripts and automation tools, including the `dvcurator` R package we are developing in-house, tools to facilitate renaming files and file metatags, VBA (Visual Basic for Applications) scripts to work with Excel files, and command-line scripts (see section 2.1).



Tabular Files

Editing of tabular files is mainly done in order to allow for ingestion to the .tab archival format on Dataverse. This includes:

- Separating worksheets into separate files - since upon conversion to .tab, only the first worksheet would be included. *This can be bulk automated for workbooks with many worksheets by using the applicable VBA script, see section 2.1.2.*
- Editing cells so that each value has a row and column header (i.e. variables) - e.g. if a single cell spans multiple rows or columns, the ingest breaks. Consolidating the values in a single row/column fixes the issue (see gh issue)
- Removing stray columns/rows to exclude variables with all missing values
- Removing line breaks, as of writing, this still prevents successful ingest (see gh issue)

In some cases, an Excel file will contain textual data not in a tabular format and it cannot be manipulated to ingest. If that is the case, consider another preservation-friendly format like .tsv.

Figure 3: Automation solutions are referred to and linked throughout the *Handbook*—an example from section 2.5. File Editing.

The Diversity of Qualitative Data

QDR maintains a list of 29 different types of qualitative data (https://qdr.syr.edu/content/types-qualitative-data) that are likely to be deposited by users. Differences in data types concern the formats of data (text, video, audio, images), the methodologies and epistemologies of depositors, and the types of constraints on the publication of the data. The *Curation Handbook* seeks to provide a framework with enough flexibility to accommodate the richness of qualitative data deposited by researchers, including with respect to constraints. Section 3.5 addresses some of the different access conditions that may be used for data. That includes different levels of controlled access for sensitive human participants data, which typically are assigned once at publication and remain static, but also conditions for which further curation work is expected due to scheduled change of status, such as embargos, both for first use and for material under copyright set to enter the public domain.

Conclusion

QDR's *Curation Handbook* is constantly evolving to add additional checks, improve workflows, or accommodate new forms of data or deposits. In this practice paper, we briefly described the institutional needs and intellectual rationale that led to the *Handbook*'s creation, as well as key features of its first iteration. More broadly, this document illustrates an important way in which a relatively new data organization with a deep focus on curation matures and addresses operational challenges. After more than one year of intensive use internally, we believe the *Handbook* has

JeSLIB 2021; 10(3): e1207
https://doi.org/10.7191/jeslib.2021.1207

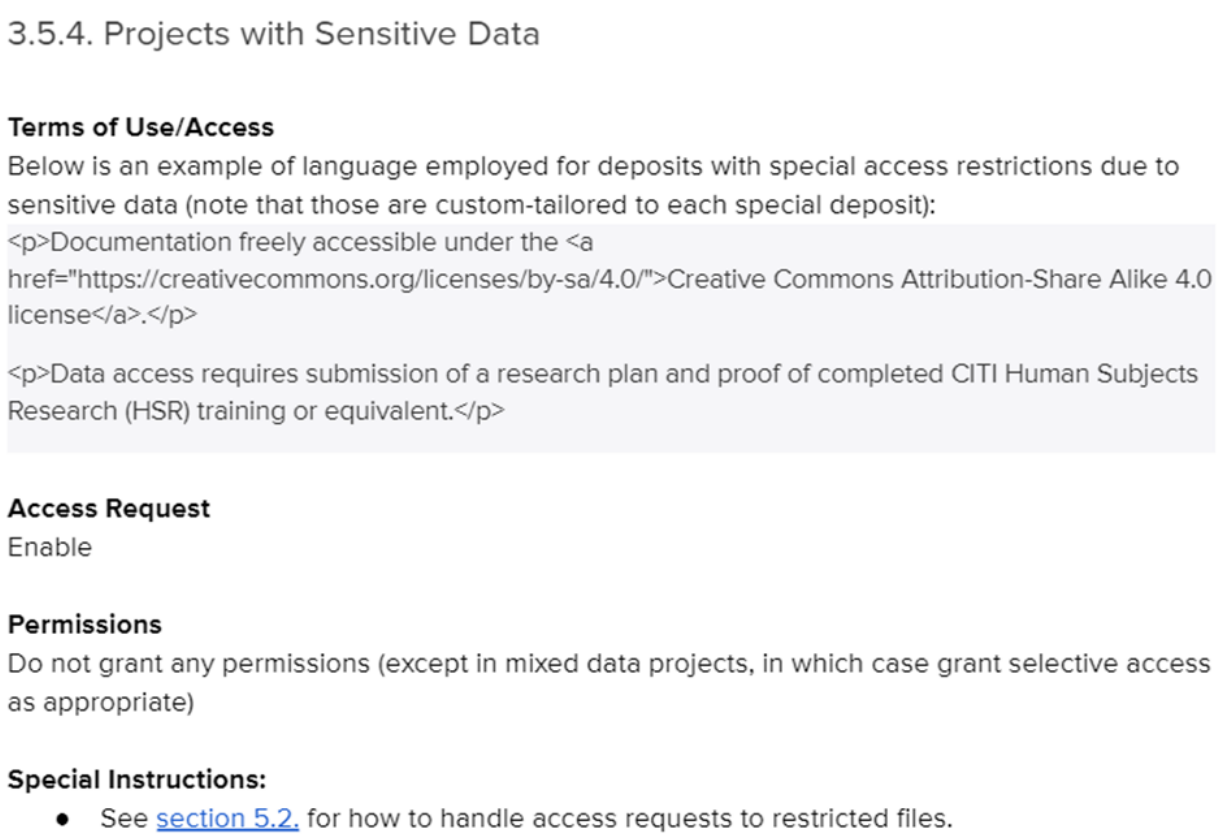


Figure 4: An example of instructions for special access restrictions from section 3.5.4., Projects with Sensitive Data.

reached a stage of development and reliability that make it a useful tool for the data community at large, too. We believe that the approach laid out in the *Handbook* can serve as an example for a variety of institutions that perform human-facilitated curation of digital artifacts beyond the qualitative and multi-method data in which QDR specializes. It outlines processes and principles for maximizing finite resources and enhancing the consistency and sustainability of their services. We also hope that, in the spirit of open science, this publication will spark discussion about good practices and may lead us to revise, expand, and improve this *Handbook*—and the curation practices it describes.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1823950.

Data Availability

This article describes QDR's curation handbook, available under CC-BY license as Demgenski, Robert, Karcher, Sebastian, Kirilova, Dessi, and Weber, Nic. 2021.

"QDR *Curation Handbook*." Syracuse University: Qualitative Data Repository. https://doi.org/10.5281/zenodo.4672678

References

Bishop, Libby. 2009. "Ethical Sharing and Reuse of Qualitative Data." *Australian Journal of Social Issues* 44(3): 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Castillo, Diana, Heather Coates, and Mikala Narlock. 2020. "Qualitative Data Primer." Data Curation Network. https://github.com/DataCurationNetwork/data-primers

Corral, Margarita. 2019. "ATLAS.Ti Data Curation Primer." Data Curation Network. https://github.com/DataCurationNetwork/data-primers

Corti, Louise. 2000. "Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research—The International Picture of an Emerging Culture." Forum *Qualitative Sozialforschung / Forum: Qualitative Social Research* 1(3): Text. Archive. Re-Analysis. https://doi.org/10.17169/fqs-1.3.1019

———. 2006. "Qualitative Archiving and Data Sharing: Extending the Reach and Impact of Qualitative Data." *IASSIST Quarterly* 29(3): 8. https://doi.org/10.29173/iq370

DCC. 2009. "Curation Checklists." Digital Curation Centre. https://www.dcc.ac.uk/sites/default/files/Curation%20Checklists.pdf

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

Demgenski, Robert, Karcher, Sebastian, Kirilova, Dessi, and Weber, Nic. 2021. "QDR *Curation Handbook*." Syracuse University: Qualitative Data Repository. https://doi.org/10.5281/zenodo.4672678

Elman, Colin, and Diana Kapiszewski. 2014. "Data Access and Research Transparency in the Qualitative Tradition." *PS: Political Science & Politics* 47(01): 43–47. https://doi.org/10.1017/S1049096513001777

Eschenfelder, Kristin, and Kalpana Shankar. 2017. "Organizational Resilience in Data Archives: Three Case Studies in Social Science Data Archives." *Data Science Journal* 16(0): 12. https://doi.org/10.5334/dsj-2017-012

Hadley, Hannah. 2019. "NVivo Data Curation Primer." Data Curation Network. https://github.com/DataCurationNetwork/data-primers

Johnston, Lisa, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2018. "How Important Is Data Curation? Gaps and Opportunities for Academic Libraries." *Journal of Librarianship and Scholarly Communication* 6(1): eP2198. https://doi.org/10.7710/2162-3309.2198

Kapiszewski, Diana, and Elisabeth Jean Wood. 2021. "Ethics, Epistemology, and Openness in Research with Human Participants." *Perspectives on Politics* March: 1–17. https://doi.org/10.1017/S1537592720004703

Karcher, Sebastian, Dessislava Kirilova, and Nicholas Weber. 2016. "Beyond the Matrix: Repository Services for Qualitative Data." *IFLA Journal* 42(4): 292–302. https://doi.org/10.1177/0340035216672870

Karcher, Sebastian, Dessislava Kirilova, Christiane Pagé, and Nic Weber. 2021. "How Data Curation Enables Epistemically Responsible Reuse of Qualitative Data." *The Qualitative Report* 26(6): 1996–2010. https://doi.org/10.46743/2160-3715/2021.5012

Karvovskaya, Lena. 2019. "Data Curation Checklist YODA." https://www.uu.nl/sites/default/files/checklist_YODA_V5p.pdf

Loyle, Cyanne E.; Davenport, Christian; Sullivan, Christopher. 2018. "Association for Legal Justice (ALJ) Human Rights Testimony, Northern Ireland." Qualitative Data Repository. https://doi.org/10.5064/F6LHMHJR

McGovern, Nancy. 2007. "A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?" *RLG DigiNews* April 15, 2007. https://hdl.handle.net/2027.42/60441

OECD. 2017. "Business Models for Sustainable Research Data Repositories." *OECD Science, Technology and Industry Policy Papers*. Paris: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/302b12bb-en

Palmer, Carole M., Nicholas L. Weber, Trevor Muñoz, and Allen H. Renear. 2013. "Foundations of Data Curation: The Pedagogy and Practice of 'Purposeful Work' with Research Data." *Archive Journal* June. http://dev.archivejournal.net/?p=4819

Trachtenberg, Marc. 2020. "United States Cold War Documents." Qualitative Data Repository. https://doi.org/10.5064/F6T3PFTW

Weber, Nicholas, Sebastian Karcher, and James Myers. 2020. "Open Source Tools for Scaling Data Curation at QDR." *The Code4Lib Journal* 49(August). https://journal.code4lib.org/articles/15436

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan 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(1): 1–9. https://doi.org/10.1038/sdata.2016.18

Yardley, Sarah J., Kate M. Watts, Jennifer Pearson, and Jane C. Richardson. 2014. "Ethical Issues in the Reuse of Qualitative Data Perspectives from Literature, Practice, and Participants." *Qualitative Health Research* 24(1): 102–113. https://doi.org/10.1177/1049732313518373