ISSN 2161-3974

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

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

Full-Length Paper

Creating Guidance for Canadian Dataverse Curators: Portage Network's Dataverse Curation Guide

Alexandra Cooper[1], Michael Steeleworthy[2], Ève Paquette-Bigras[3], Erin Clary[4], Erin MacPherson[5], Louise Gillis[5], and Jason Brodeur[6]

[1] Queen's University, Kingston, ON, Canada
[2] Wilfrid Laurier University, Waterloo, ON, Canada
[3] Université de Montréal, Montreal, QC, Canada
[4] Canadian Association of Research Libraries, Portage Network, Ottawa, ON, Canada
[5] Dalhousie University, Halifax, NS, Canada
[6] McMaster University, Hamilton, ON, Canada

Abstract

Purpose: This paper introduces the Portage Network's Dataverse Curation Guide and the new bilingual curation framework developed to support it.

Brief Description: Canadian academic institutions and national organizations have been building infrastructure, staffing, and programming to support research data management. Amidst this work, a notable gap emerged between requirements for data curation in general repositories like Dataverse and the requisite workflows and guidance materials needed by curators to meet them. In response, Portage, a national network of data experts, organized a working group to develop a Dataverse curation guide built upon the Data Curation Network's CURATED workflow. To create a

Correspondence: Alexandra Cooper: coopera@queensu.ca

Received: March 31, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Disclosures: The authors report no conflict of interest.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Abstract Continued

bilingual resource, the original CURATE(D) acronym was modified to CURATION—which has the same meaning in both French and English—and steps were augmented with Dataverse-specific guidance and mapped to three conceptualized levels of curation to assist curators in prioritizing curation actions.

Methods: An environmental scan of relevant deposit and curation guidance materials from Canadian and international institutions identified the need for a comprehensive Dataverse Curation Guide, as most existing resources were either depositor-focused or contained only partial workflows. The resulting Guide synthesized these guidance materials into the CURATION steps and mapped actions to various theoretical levels of data repository services and levels of curation.

Resources: The following documents are supplemental to the Dataverse Curation Guide: the Portage Dataverse North Metadata Best Practices Guide, the Scholars Portal Dataverse Guide, and the Data Curation Network CURATED Workflow and Data Curation Primers.

Introduction

Data curation—the active management of research data as it is created, maintained, used, archived, shared, and reused—is a core component within the assemblage of infrastructure, processes, schemas, and curator expertise that supports best practices in Research Data Management (RDM).[1] The execution of a well-articulated data curation workflow can make good data *better* by expertly describing its contents, creating a coherent structure, providing meaningful documentation, enabling automation through code and syntax, and linking to other data and outputs.

Over the past half decade, Canadian academic institutions and programs have developed national RDM infrastructure and services. The work of organizations including the Canadian Association of Research Libraries' (CARL) Portage Network,[2] Compute Canada,[3] Research Data Canada (RDC),[4] and the Ontario Council of University Libraries' Scholars Portal[5] have expanded the capacity of academic institutions to organize and provide research data services such as data management planning and data deposit. At the time of writing, 58 Canadian institutions participate in Scholars Portal Dataverse (Portage Network 2020a) and many academic institutions in Canada today have RDM staff. However, in this buildup of infrastructure, staff, and programming, a gap between the improved facility to house properly curated data and the local institution's knowledge and capacity to conduct data curation itself became apparent (Clary et al. 2020).

In early 2020, the Portage Network's Curation Expert Group (CEG) convened a working group of data professionals from across Canada to help fill this gap. The *Dataverse Curation Guide Working Group (WG)* was tasked with three objectives (Portage 2020b):

- Engage with current Dataverse curators to solicit information related to current practices and guidance documents;
- Develop a set of basic documentation and reference materials that communicate best practices for curating datasets in Dataverse;
- Publish a Dataverse curation guide as a publicly available resource in both French and English.

This article presents the Dataverse Curation Guide and explains the context within which it was created, including the need for a bilingual document that can be used by new and experienced curators alike.

1 Research data management (RDM) refers to the processes applied through the lifecycle of a research project to guide the collection, documentation, storage, sharing and preservation of research data (Government of Canada 2018).

2 https://portagenetwork.ca

3 https://www.computecanada.ca

4 https://www.rdc-drc.ca

5 https://scholarsportal.info

Creating the Dataverse Curation Guide

Through 2020, our WG conducted an environmental scan of Dataverse-related deposit preparation and curation guidance materials from Canadian and international institutions. The purpose of this exercise was to identify existing material that might be applicable to a discipline-agnostic Dataverse curation guide, as well as define the breadth and scope of the gap our guidance should fill. In total, we compiled and analyzed 21 resources—11 from Canadian organizations and 10 from international ones. Curation guidance from each resource was categorized by activity and evaluated for its general applicability to curating data in Dataverse, as well as the intensity of the task. Our analyses revealed a scarcity of formal Dataverse curation guidance documents among Canadian institutions, and little consistency between the guidance documents that did exist. We also found that most material used by Canadian data professionals to curate in Dataverse took the form of deposit guides and partial workflows as opposed to curation advice and guidance on how to prepare deposited data for long-term stewardship.

To develop a comprehensive, bilingual curation guide suitable for discipline-agnostic Dataverse repositories, our group adapted the Data Curation Network's CURATE(D) workflow as a framework. This tool was first introduced to a large part of the Canadian curation community through DCN-led workshops at the Canadian Data Curation Forum in October, 2019.[6] The CURATE(D) workflow presents a standardized set of data curation steps in checklist format,[7] which enables the consistent treatment of submitted datasets across disciplines and repositories (Johnston et al. 2018). To fit the workflow within the Canadian context, our working group made two significant adaptations. First, to ensure our curation guide is fully bilingual and has utility for both English and French speaking curators, we modified the original CURATED acronym to CURATION, which has the same meaning in both languages (see Table 1). Although slight adjustments have been made to the workflow to accommodate this change, the original content and its intent remain in the new framework.

6 https://data-curation.github.io

7 The original CURATE steps are available in Johnston et al. (2018, Appendix), and the workflow, now updated to CURATED, is available online (Data Curation Network. n.d.).

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Table 1: CURATION model in English and French.

	English	French
C	**C**heck	**C**onsulter
U	**U**nderstand	**U**n peu plus en profondeur
R	**R**ecommend improvements	**R**ecommander
A	**A**ugment	**A**méliorer
T	**T**ransform	**T**ransposer
I	**I**nclude persistent IDs and a reuse licence/agreement	**I**nclure ID pérennes et les licences/ ententes de réutilisation
O	**O**ptimize for FAIRness	**O**ptimiser selon les principes FAIR
N	**N**ote down curation activities	**N**oter les actions réalisées

Secondly, to accommodate a variety of service scenarios at academic institutions across Canada and variations in institutional curation capacity and curator experience, we have modeled the Guide on three potential service scenarios (see Table 2), and three different levels of curation (see Table 3).

Table 2: Potential Dataverse curation scenarios for academic institutions.

Unmediated curation	There is no intervention from the RDM service provider. The researcher creates their own dataverse and dataset, submits their data and publishes it.
Semi-mediated curation	The RDM service creates a dataverse or starts a dataset deposit and assigns a role to the researcher, and then the researcher submits data. Depending on local policy, the dataset is either flagged for review by the institutional dataverse administrators, or the depositor may request to have the dataset reviewed by the data management team before or after it is published.
Mediated curation	Data is submitted by the researcher to the RDM service. The RDM service creates the dataverse (or dataset) and the data is curated by the library and published once approved by the researcher.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Table 3: Conceptualized levels of data curation used in the Dataverse Curation Guide.[8,9]

Level 1	The minimum steps required to successfully publish in Dataverse and make the dataset findable, i.e., the dataset has been submitted to the proper dataverse and required metadata fields are accurate.
Level 2	Activities that enhance the discoverability of datasets and help ensure their usability over time. E.g., recommended metadata fields are populated and the dataset includes sufficient documentation to allow a user with a similar background to understand the data and open and use the files.
Level 3	Intensive curation actions intended to prepare datasets for preservation and improve the chances that data and code can be used to reproduce or replicate an associated study. For example, supporting documentation is enhanced, the content of files and code are reviewed, and data files are transformed into formats suitable for long-term preservation.

There is no "one size fits all" model to data curation. The level and quality of curation is dependent on local resourcing, capacity, policies, priorities, expertise, and institutional strategic direction. As a result, this Guide has been developed with flexibility in mind.[10] It can be used by new or experienced curators within academic institutions of varying sizes, and it can be adapted by institutions to meet the needs of local policies and procedures. We consider our Curation Guide, and its companion Quick Reference Guide [Appendix 1 (English), Appendix 2 (French)], to be "living documents," and they will be available in GitHub to facilitate collaboration and reuse.

Complementary Guides

This Curation Guide should be used alongside a number of other data curation resources, including Dataverse-specific guidance. The complementary guides listed below go into greater depth on the particulars of the Dataverse platform, considerations for specific file types, and best practice guidance for metadata in Dataverse. They will be particularly useful for Level 2 and 3 curation steps or as a reference when working with unfamiliar data types.

8 For a comprehensive list of curation tasks and the assigned level, please see the Dataverse CURATION Quick Reference Guide / CURATION dans Dataverse : Guide éclair (Appendix 1).

9 For more information on levels of curation and how these were applied in practice at two institutions, please see Lafferty-Hess et al. (2020)

10 The levels of curation presented in this Guide do not map perfectly to the curation service scenarios. For example, one repository may have a fully mediated curation service, but only have time to complete Level 1 checks on each submission. Another repository may have a semi-mediated service with a curator who conducts all Level 2 checks and one check from Level 3 after the dataset is published.

Dataverse North Metadata Best Practices Guide[11]

This bilingual guide provides definitions for each metadata field in Dataverse, along with examples and tips to ensure appropriate values are included and formatted correctly. It distinguishes between required, recommended and optional metadata fields and can be used to facilitate rich description beyond the scope of what the Curation Guide has covered.

Scholars Portal Dataverse Guide[12]

This bilingual data deposit guide provides an overview of the steps required to deposit a dataset in Scholars Portal Dataverse, the largest Dataverse instance in Canada. It provides information on how to use Dataverse, and how to create and edit datasets in Dataverse. For advanced guidance, see the Harvard Dataverse Project User Guide.[13]

Data Curation Network (DCN) Curation Workflow[14]

This guide provides a series of steps and checklists that walk through the curation process in a standardized manner, regardless of the data type or repository platform. The CURATED framework was adapted for this guide as CURATION.

Data Curation Primers[15]

Developed by the curation community with guidance from the Data Curation Network, these Primers are a collection of resources that provide in-depth advice for curating various types of data. The Primers provide additional guidance on file formats, software requirements, and address questions the other guides may not cover.

Dataverse Curation Guide—CURATION steps

Check / Consulter

At the **Check** step,[16] confirm that all data and metadata components required by the system to successfully publish the deposit are present. If possible, identify any characteristics that may require special consideration (e.g., data with disclosure risk, or data obtained from a third-party source).

11 In English http://hdl.handle.net/2429/73609 and en français http://hdl.handle.net/2429/73610

12 In English https://learn.scholarsportal.info/all-guides/dataverse and en français https://learn.scholarsportal.info/fr/guides/dataverse

13 https://guides.dataverse.org/en/5.3/user

14 https://datacurationnetwork.org/outputs/workflows

15 In PDF format https://hdl.handle.net/11299/202810 and in GitHub https://github.com/DataCurationNetwork/data-primers

16 The first two steps, Check and Understand, overlap. The tasks in the Check step are all Level 1 tasks that should be completed before publishing a dataset in Dataverse. The tasks in Understand are Level 2 and Level 3 tasks that may be completed in accordance with your institution's service level and policies.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Level 1

Yes	No	Some issues	N/A	
❑	❑	❑	❑	The dataset has been submitted to the proper dataverse. To evaluate whether the dataset is in the most appropriate dataverse, consider the following: • Has the researcher (or their research group) previously created or submitted to a dataverse? • Does the dataset require its own dataverse or is there an associated dataverse to which it belongs? • Does the dataset conform to the policies and submission standards associated with the specific dataverse?
❑	❑	❑	❑	The researcher has confirmed that the dataset is free of any licensing and intellectual property issues.
❑	❑	❑	❑	The researcher has confirmed that the dataset is free of any sensitive information (i.e., information that must be safeguarded against unwarranted access or disclosure).
❑	❑	❑	❑	Supporting documentation is included. For example, a codebook, data dictionary, methodology, Readme file, etc.
❑	❑	❑	❑	All files described in the documentation are included in the dataset.
❑	❑	❑	❑	Required metadata fields are accurate. Use the Dataverse North Metadata Best Practices Guide (Bascik et al. 2020) to evaluate the completeness and accuracy of the required metadata fields (Title, Author Name, Contact Email, Description, and Subject).

Understand / Un peu plus en profondeur

In the **Understand** step,[17] you should ensure the dataset is well-described and that end users will have a clear picture of what the data is and how it can be used. Review the metadata and documentation for thoroughness and clarity, create or recommend additional documentation if required, and check for usability issues such as missing data, code execution failures, ambiguous headings, and data presentation concerns. Depending on your institution's policies and the level of curation service your repository provides, you may also screen for disclosure risk, assess for intellectual property rights infringements, and other tasks.

17 The first two steps, Check and Understand, overlap. The tasks in the Check step are all Level 1 tasks that should be completed before publishing a dataset in Dataverse. The tasks in Understand are Level 2 and Level 3 tasks that may be completed in accordance with your institution's service level and policies.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

For a thorough overview of steps you might take to understand the data and assess its completeness and usability, see Johnston (2017, Step 3.0). The Data Curation Network's Data Curation Primers (2019) are another excellent resource with guidance and best practice advice for curating specific file types. They include information about tools for file review and a series of considerations and questions the curator should ask of the data.

Level 2

Supporting documentation is thorough, accurate, and complete

If the dataset has a Readme file, codebook, user manuals or other documentation, review it for accuracy and completeness. The documentation should provide contextual information about the dataset to increase its usability. If documentation is inadequate, work with the researcher to enhance existing documentation, or provide them with templates and other guidance to create documentation. Cornell University's "Guide to writing "readme" style metadata" (n.d.)[18] and University of British Columbia's "Creating a README for your Dataset: Quick Guide" (Brigham 2020)[19] are useful resources for both curators and researchers.

Yes	No	Some issues	N/A	Documentation includes...
❑	❑	❑	❑	Contextual information about the data (how the data was collected or generated, the goal of the research).
❑	❑	❑	❑	Description of file naming conventions and the structure of the files, if important.
❑	❑	❑	❑	Record of how the data were modified or processed.
❑	❑	❑	❑	Information about confidentiality and any restrictions placed on secondary use.
❑	❑	❑	❑	Names of labels and variables, information about allowable values and units of measure, codes and classifications, if applicable.
❑	❑	❑	❑	Explanations of codes and classifications.
❑	❑	❑	❑	Description of the computing environment required to run any code that has been included (operating system, software packages and dependencies).

18 https://data.research.cornell.edu/content/readme

19 In English https://doi.org/10.5281/zenodo.4058971 and en français https://doi.org/10.5281/zenodo.4058961

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Files open properly and contents appear as expected

Download the dataset and extract the contents of any archive file types (.zip, .tar, etc.). Review file content and address any issues with proprietary files. **Tip**: For unfamiliar file types, does the documentation provide guidance on what software was used to generate the file, or how it might be viewed? Try opening the file in a text editor as even binary files may have plain text headers with information about the instrument or software that generated the file.

If it is not feasible to open all files, check a subset that contains:

- At least one of every file type in the submitted dataset,
- Script files, code, and anything you suspect may be licensed,
- Any file you have reason to believe may include sensitive information.

Yes	No	Some issues	NA	
❑	❑	❑	❑	Files open as expected and archive file formats extract without issue (unknown software, incompatible versions, or no access to the software could be a reason for files not running).
❑	❑	❑	❑	If the files are not accessible, has the researcher provided a non-proprietary version of the files?
❑	❑	❑	❑	If non-proprietary files cannot be provided, does the Readme describe how the data were generated, and the software necessary to use the data?
❑	❑	❑	❑	File contents are consistent with expected structure and encoding.
❑	❑	❑	❑	Readme file describes the data files and includes the following: ❑ Summary description ❑ File formats (flagged if proprietary) ❑ File size ❑ Path and/or tree structure ❑ A checksum that can be used to verify data integrity (e.g., md5, sha256)

20 Consider using tools to automate this output. In Windows, use the TREE function from the command line. In Linux, install and use the tree command. On a Mac, use a package manager such as Homebrew to install the tree package.

21 Consider using tools to automate this output. In Windows, use a program such as HASHMYFILES. In Linux, use the function md5sum or sha256sum on the command line. On a Mac, use the function md5 or shasum -a 256 on the command line.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Files and folders are named and structured appropriately

Ideally, files should be named and organized in a manner that is understandable and allows end users to easily navigate the contents of the dataset. The directory structure should be simple and directory names should clearly communicate their contents. The criteria below, provided by the University of Victoria (Khair 2020), are general best practices, and conventions for specific data types or disciplines may differ.

Yes	No	Some issues	NA	
❑	❑	❑	❑	Filenames are free of spaces or special characters and use underscores or hyphens as delimiters (e.g., Datacollection_20201009_v02.csv).
❑	❑	❑	❑	File and directory names are concise, consistent, and understandable.
❑	❑	❑	❑	Dates in filenames use consistent formatting (e.g., YYYYMMDD).
❑	❑	❑	❑	Filenames use leading zeros for version numbers (e.g., v_012).
❑	❑	❑	❑	Files are grouped in a logical folder structure and are not overly nested. Note: Folder structure may be dependent on the code, syntax, or output of a piece of software.

Level 3

Code is well commented and produces the expected results

The researcher may have included script files that were used to process or analyze the data, code that extends an existing model, executable files, or other software. While it may be beneficial to keep script files with the data, there are many purpose-built repository options for code and software that have robust version control systems and allow for ongoing development. You may review the code and associated documentation alongside the data deposit and/or suggest an alternate solution for publishing and archiving software.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Yes	No	Some issues	N/A	If the dataset includes software or code...
❑	❑	❑	❑	Is the code also available in GitLab, GitHub, Bitbucket, or another purpose-built repository? • If yes, consider asking the researcher to archive it in the Software Heritage archive[22] and link to it from the dataset metadata record. • If no, and the researcher has included more than data processing or analysis scripts (such as .r, .m, or .do files), consider suggesting a purpose-built repository.
❑	❑	❑	❑	Is the code (or part of the code) derived from another source?
❑	❑	❑	❑	If the code is derived from another source, does the original source code allow for redistribution? Is the current license compatible with the license of the original source?
❑	❑	❑	❑	If run, does the software code produce the expected results without error?
❑	❑	❑	❑	If an executable file is included, is the source code also available?
❑	❑	❑	❑	Code is well commented • The code contains header information such as: author, version number, filename, license • The function or purpose of the code is clear (from the Readme and/ or embedded description) • If applicable, the depositor included information about how to run the code • The required software packages and dependencies are listed • Comments are concise and clear and describe the intention of the line(s) of code that follow, and/or the code itself is expressive (can be understood by humans and machines)
❑	❑	❑	❑	The Readme or a header in the code itself includes information about: • The license • The developer's name and contact information • The version, and the date the code was last modified and/or run • Any sources the code (or part of the code) was derived from • Instructions on how to install or use the code. If there are multiple scripts, the order in which they are run should be clear. • Required software packages and dependencies • Information about the environment in which the code was developed and/or can be used • Information about required input and expected output

22 https://www.softwareheritage.org/save-and-reference-research-software

Submission contains potential sensitivities

Note: Most instances of Dataverse allow researchers to restrict access to datasets, either temporarily or in perpetuity; however, unless the repository servers are encrypted and the service explicitly accepts data that must be safeguarded against unauthorized access or disclosure, Dataverse is not an enclave suitable for sensitive data. Review your Dataverse policies and/or consult with your administrator to confirm what types of information are suitable for your repository.

Data with potential sensitivities should be flagged by the researcher prior to deposit, but this will not always be the case. The dataset description or the presence of participant consent forms or participation agreements may alert you to a dataset that warrants special consideration, but some types of sensitive data will not be obvious at the point of deposit. Data that may need extra consideration or care include data collected with Indigenous partners; data collected about or from Indigenous peoples and their land, water, resources and environment; traditional knowledge; data collected on private property; data with location information of vulnerable species or protected sites; proprietary data or data collected with an industry partner; data that are being reused or were otherwise provided by a third party; and other data where collection or publication are subject to a data sharing agreement.

You may wish to flag content that appears to be in violation of your repository terms of use or work with the depositor to mitigate disclosure risk. If you do so, reiterate that responsibility for compliance with your terms of use, participant consent, and data sharing agreements does lie with the research team.

Local considerations or policy may determine how you handle data with sensitivities, and you may not have the authority to make data sensitivity determinations without external guidance. There are many stakeholders who can provide support, including your university library, Research Ethics Board, Research Office, Indigenous Relations Office, Industry Liaison Office, and other departments associated with research support on your campus.

Sensitive data resources

- Portage Network Sensitive Data Toolkit for Researchers
  - Part 1 – Glossary of Terms for Sensitive Data Used for Research Purposes (Sensitive Data Expert Group 2020a)[23]
  - Part 2 – Human Participant Research Data Risk Matrix (Sensitive Data Expert Group 2020b)[24]

23 In English http://doi.org/10.5281/zenodo.4088946 and en français https://doi.org/10.5281/zenodo.4088985

24 In English http://doi.org/10.5281/zenodo.4088954 and en français https://doi.org/10.5281/zenodo.4107119

    - Part 3 – Research Data Management Language for Informed Consent (Sensitive Data Expert Group 2020c)[25]
    - Data Curation Network Primers
    - Consent Forms Data Curation Primer (Hunt, Hofelich Mohr, and Woodbrook 2021)[26]
    - Human Subjects Data Essentials Data Curation Primer (Darragh et al. 2020)[27]
- Portage Network De-Identification Guidance[28]
- UK Statistics Authority Ethical considerations in the use of geospatial data for research and statistics (Centre for Applied Data Ethics 2021)[29]
- Global Biodiversity Information Facility Current Best Practices for Generalizing Sensitive Species Occurrence Data (Chapman 2020)[30]

To determine if a dataset contains sensitivities or data sharing restrictions, the following approaches may be useful:

- Based on title, description, keywords, and documentation, identify whether there is potential for sensitive information to be disclosed in the dataset.
- Review consent forms for language that precludes data sharing.
- Inquire with the researcher about the presence of sensitive information. **Note**: if researchers are required to complete a dataset information form prior to submitting, this question can be included on the form.
- Open files and explore for sensitive content (e.g., precise location information for protected species and sites, direct and indirect personal identifiers, and other information in violation of your repository terms of use).

25 In English http://doi.org/10.5281/zenodo.4107178 and en français https://doi.org/10.5281/zenodo.4107185

26 https://hdl.handle.net/11299/218838

27 https://hdl.handle.net/11299/216579

28 In English https://portagenetwork.ca/tools-and-resources/rdm-guidance-for-covid-19/de-identification-guidance and en français https://portagenetwork.ca/fr/outils-et-ressources/directives-de-la-gdr-sur-la-covid-19/directives-sur-la-depersonnalisation-des-donnees

29 https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-geospatial-data-for-research-and-statistics

30 https://doi.org/10.15468/doc-5jp4-5g10

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Yes	No	Unsure	N/A	
❑	❑	❑	❑	Dataset includes information about: • Human participants (interviews, survey responses, biomedical, health-related data, etc.) • Indigenous topics/subjects • Minors or people unable to provide informed consent • Vulnerable species • Protected/private property • Illegal/offending content • Any other content that could be potentially sensitive or obviously violates your repository's terms of use.
❑	❑	❑	❑	If "Yes" to above, is at least one the following included: • Consent form • Participation agreement • Data sharing agreement • Research Ethics Board approved exception where seeking informed consent is impracticable (Panel on Research Ethics 2021)[31] • Any other document outlining permission to make data available for secondary use.
❑	❑	❑	❑	The consent form or data sharing agreement includes language that allows the researcher to share the data in a repository.
❑	❑	❑	❑	If the data can only be shared under a specific set of circumstances, the researcher has selected the appropriate mechanisms to restrict access. NOTE: Dataverse servers are not encrypted and should not be used to store identifiable

Submission contains data or code from third party sources

Datasets should be inspected for data or code from third-party sources to verify that researchers have the proper rights or permissions to share data and that proper attribution has been provided. Although resources may be free to access, view, or use, it does not necessarily follow that they are free to redistribute. Consult with your copyright office or specialists on campus to determine how your organization's policies regarding third-party intellectual property and rights affect deposit into your repositories.

31 https://ethics.gc.ca/eng/depositing_depots.html

Yes	No	Unsure	N/A	
❑	❑	❑	❑	Dataset contains proprietary or restricted information. Example: commercially licensed or proprietary data or code, or third-party data that are only accessible by registering or logging in.
❑	❑	❑	❑	Third-party data or code has been properly cited and the original terms of use allow for redistribution.
❑	❑	❑	❑	A data sharing agreement is referenced or included and allows for information to be redistributed.

Recommend / Recommander

In the **Recommend** step, you may request additional information from the researcher and suggest changes to metadata and files to improve the findability and usability of the dataset. Recommend is listed as a Level 1 curation task. At Level 1, it may not be necessary to request changes to every deposit, but you will need to reach out if the minimum steps required to publish a dataset in Dataverse are not met. If you complete tasks in Level 2 and Level 3, you will need to reach out to researchers more frequently. Be prepared to manage expectations as researchers may not be familiar with the required or recommended practices in your repository, or aware that your repository offers curation services. Some researchers will be more amenable to recommendations or changes than others, however, even those researchers who cannot apply your recommendations to the current deposit may use them to improve the next deposit.

Level 1

- You may find it helpful to flag questions about metadata or data as you work through the other steps in the CURATION framework.
- In the end, your requests should be prioritized, and the difference between what is required to publish the dataset versus what would be nice to have should be clear.
- If the deposit would benefit from significant revision, an in-person discussion may be preferable to email.
- If you have a set of guidelines to help researchers during the deposit process, consider referring to them as a starting point in your conversation.
- To save yourself time, and to ensure consistency across deposits, consider using templates as a starting point for your messages. Johnston (2017, section 3.5)[32] includes a sample message and more information to help you get started.

32 https://ethics.gc.ca/eng/depositing_depots.html

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Augment / Améliorer

The **Augment** step is an opportunity to enhance dataset documentation and metadata to further facilitate discoverability and usability. This includes any steps you can take to improve documentation and metadata beyond what was done in the Check and Understand steps. The Dataverse North Metadata Best Practices Guide (Bascik et al. 2020) includes extensive guidance that can be used to enrich the metadata record.

Level 2

Metadata is rich, accurate, and complete

Tip: Keyword and Topic Classification terms provide additional information about the dataset that situate it within a field of study, and aid in classification, indexing and discovery.

Yes	No	Some	NA	
❑	❑	❑	❑	The dataset has a descriptive title
❑	❑	❑	❑	Terms and acronyms are defined
❑	❑	❑	❑	The metadata is free of jargon (for human readability) and symbols (for machine readability)
❑	❑	❑	❑	The language in descriptive metadata fields is precise and specific
❑	❑	❑	❑	The ***Keyword*** field includes terms that describe the dataset
❑	❑	❑	❑	The ***Topic Classification*** field includes terms that describe the dataset
❑	❑	❑	❑	Each term has been assigned to its own field (i.e., one term per box)
❑	❑	❑	❑	Terms are provided from a controlled vocabulary (i.e., ***Vocabulary*** and ***Vocabulary URL*** subfields are completed)

Links to related publications, datasets, and other resources are included

When possible, provide direct links out to associated publications, code, models, documentation, source data and other related resources. Objects published externally provide additional context to the new deposit and may help the end user better understand the dataset. Some of these resources may be mentioned in the dataset description or Readme file, or they may have come to light during the dataset review process.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Yes	No	Some	NA	
❑	❑	❑	❑	***Related Publication*** field and subfields (Citation, ID Type, ID Number, URL) are complete
❑	❑	❑	❑	Any related datasets are noted in the ***Related Datasets*** field
❑	❑	❑	❑	Links to other published resources, such as associated models or code, documentation, survey instrument, study protocol, analysis plans, data management plan, etc. are added to ***Other References*** field
❑	❑	❑	❑	Grant Information field and subfields (***Grant Agency*** and ***Grant Number***) are complete
❑	❑	❑	❑	The ***Data Sources*** field is complete

Transform / Transposer

The goal of the **Transform** step is to produce datasets that use open and common formats. Open formats ensure the data are available to the widest possible audience, including those without access to specialized software, and that preservation actions can be taken in the future. There will be times when specialized formats are required (e.g., complex data types/structures that require specific formats to represent), but where alternatives exist, you may wish to consider transformation. Loss of data or information embedded in the data structure or metadata can be a worry. Whenever possible, ask the researcher to provide an alternate format, rather than transforming the files yourself, so they can confirm the result is accurate and complete.

Retaining the original files alongside the transformed files may be ideal because the original data are available in the form the researcher used and can be opened and manipulated in their original software. However, depending on overall submission size, the researcher may not wish to publish both the original, proprietary file types and an open-source alternative.

If the files are transformed, confirm file names reflect which files are original and which are transformed. If you or a preservationist are responsible for transformation, the original files should be retained in accordance with the repository's policy. If the files cannot be transformed, confirm that the Readme file describes the software needed to view or use the files and outlines the file contents.

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Level 3

File formats are open, or appropriately documented

Yes	No	Some issues	N/A	**Identify file formats and software used to create the files**…
❑	❑	❑	❑	Dataset contains non-proprietary file formats or formats that are preservation friendly. For guidance, see DataverseNO[33] preferred files formats.
❑	❑	❑	❑	Proprietary files can be transformed to non-proprietary formats without losing data.
❑	❑	❑	❑	If files **cannot** be transformed, does the documentation describe: ❑ The file format and contents of the files? ❑ The software or instrument that generated the files? ❑ The software necessary to view or use the files? ❑ A freeware option for accessing the files, if one is available?
❑	❑	❑	❑	Documentation describes data, formulas used, column headings, variable labels, etc. so data are usable.
❑	❑	❑	❑	Documentation is in a preferred format (e.g., PDF/A or .txt).

Include / Inclure

The goal of the **Include** step is to facilitate data reuse and promote proper attribution and credit. The dataset should include relevant persistent IDs and appropriate licensing information. Some of this information will likely be assigned automatically, for example, the repository will automatically assign a DOI and register it when the dataset is published. Other information, such as ORCID IDs, typically need to be added manually.

A license will describe the acceptable uses of the published dataset. Ideally, the selected license meets the researcher's needs without being overly restrictive. It is typically preferable to use an established license; however, a custom license or data sharing agreement may be necessary. A researcher's preference, the requirements for redistributing any data or code that was obtained or derived from a third-party source, and the requirements of any data sharing agreement the researcher may have entered into should all be considered before a license is selected. Tools, such as the Creative Commons License Chooser[34] and GitHub's Choose an Open Source License,[35] and resources such as the Open Source

34 https://chooser-beta.creativecommons.org
35 https://choosealicense.com

Initiative[36] and the Open Knowledge Foundation[37] can help select an appropriate license. Dataverse has also published a template that may be useful for creating a data use agreement (Dataverse Project n.d.).[38] Please be aware that Creative Commons recommends against using their licenses for code or software (n.d.a).

Level 2

Include persistent identifiers wherever possible

Yes	No	Some issues	N/A	
❑	❑	❑	❑	Dataset is part of a publication, or a supplement to another resource?
❑	❑	❑	❑	If yes, citation or persistent link is included in the documentation and metadata record. Verify link goes to the correct place.
❑	❑	❑	❑	Relevant author IDs, such as ORCIDs, are included.
❑	❑	❑	❑	Relevant funder/grant data are included.
❑	❑	❑	❑	The system-assigned data DOI is registered and resolves to the dataset landing page (if curation takes place before publication, you may not be able to confirm this until after curation is complete).

Level 3

Review the licensing and terms of use for the dataset

It is the responsibility of the researcher to select a license that is appropriate for the dataset they are publishing. If part of the dataset was provided by a third-party, obtained from an existing resource, or derived from an existing dataset, the newly selected license must comply with the terms of use assigned to the original resource. You may wish to discuss the items below with the researcher.

36 https://opensource.org/licenses
37 https://opendatacommons.org
38 https://dataverse.org/best-practices/sample-dua

JeSLIB 2021; 10(3): e1201
https://doi.org/10.7191/jeslib.2021.1201

Yes	No	Unsure	N/A	
❑	❑	❑	❑	Has the researcher selected a license already? If yes, which license or terms of use were selected _____
❑	❑	❑	❑	Does the deposit contain data or code obtained or derived from a third-party source? If yes, what license or terms of use were assigned to the original source? _____
❑	❑	❑	❑	Is the license selected by the depositor less restrictive than terms set by third-party source? If yes, consult with the depositor to suggest a more appropriate license. If no, you can move forward with the license the depositor selected.
❑	❑	❑	❑	Are data that were obtained from third-party sources or derived from existing datasets properly attributed?
❑	❑	❑	❑	Readme includes a preferred citation for the newly deposited dataset.

Optimize / Optimiser

At the **Optimize** step, the overall FAIRness[39] of a dataset is formally evaluated, and you or the researcher may take steps to improve the Findability, Accessibility, Interoperability, and Reusability of the data. While many tasks in the CURATION framework are intended to improve FAIRness, this step is a formal assessment. You may complete this task for internal purposes, to improve your own practice, or you may use the results of the assessment to frame the requests you send to the researcher in the Recommend step.

Level 3

Evaluate the dataset and Optimize FAIRness

There are a number of tools available to evaluate how well a dataset adheres to the FAIR principles, including:

SATIFYD[40] (Self-Assessment Tool to Improve the FAIRness of Your Dataset), from the Data Archiving and Networked Services, rates FAIRness based on your responses to 12 questions and provides tips to improve the score (Fankhauser 2019).

39 For a description of the FAIR principles and their importance, see Wilkinson et al. (2016).
40 https://satifyd.dans.knaw.nl

The *CSIRO 5-star Data Rating Tool*[41] is a self-assessment rating scheme to evaluate FAIRness (Yu and Cox 2017). The assessment criteria are adapted from the OzNome Data Ratings criteria, which provide specific metrics for achieving FAIR and trusted data (Yu n.d.).

The Australian Research Data Commons *FAIR Self-Assessment Tool*[42] was designed for data librarians and IT staff. Answers are selected from a drop-down menu and overall FAIRness is adjusted in real-time via a green "progress bar."

Note Down / Noter les actions

The objective of the **Note Down** step is to assure an accurate, written record of your curation work. Certain actions and information may be recorded automatically in the metadata record (e.g., curator name, name of depositor, date the dataset was submitted, returned, or approved) while others should be recorded manually in a *curation log*, a standardized document that summarizes what changes have been made, by whom, and why. The format, content, and extent of the curation log may vary depending on your repository's record keeping requirements and the level of interaction you have with your researchers. Although it is listed as the last step in the curation process, it is helpful to create and populate the log as you move through the other steps, and then finalize it at the end of the curation process. More information about curation logs, including a template, can be found in Johnston (2017, section 3.2, 86).[43]

Level 2

Create a curation log to document your decisions and actions

Yes	No	N/A	
❑	❑	❑	A plain text log file has been created to document the following information: ❑ DOI, internal item ID, or another way to link the log file to the dataset ❑ Any potential issues uncovered as part of the curation process ❑ Questions and high-level change recommendations for the researcher ❑ A high-level summary of the researcher's response to requests and questions ❑ Any changes made to the dataset during the curation process, including: ❑ Changes made to existing files or documentation ❑ The names of any files added or removed from dataset, and why ❑ A list of metadata terms that were changed, added or removed

41 https://doi.org/10.4225/08/5a12348f8567b
42 https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool
43 https://hdl.handle.net/11299/185335

Conclusion

This Guide introduces the CURATION framework, a set of steps for curating new datasets deposited in Dataverse. The framework is adapted from the Data Curation Network's CURATE(D) steps (Johnston et al. 2018), for use in a bilingual context. Although the steps are presented in a particular order, we acknowledge that data curation is rarely a linear process. The tasks that are completed, and the order in which they are done will depend on a number of factors, including the institution's policies or practices; the number of dedicated curation staff and the amount of time they have available to review each new deposit; the type of data and the curator's comfort level with that data and the curation process; and the amount of time the researcher can dedicate to questions and revisions. Our Dataverse CURATION framework builds upon the CURATE(D) protocol and highlights opportunities for future internationalization and collaboration of the model. Its standard set of steps will directly support new and expert Dataverse curators as they walk through the curation process, and will facilitate the long-term stewardship of research data.

Acknowledgements

Canadian Association of Research Libraries—Association des bibliothèques de recherche du Canada (CARL-ABRC), Portage Newtork, Data Curation Network (DCN), Scholars Portal Dataverse, Canadian Dataverse Community for comments and feedback, and Meghan Goodchild and Lee Wilson for their help shaping the Guide.

Supplemental Content

Appendix 1 - Dataverse CURATION Quick Reference Guide (English)
Appendix 2 - CURATION dans Dataverse : Guide éclair (French)
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2021.1201 under "Additional Files".

References

Australian Research Data Commons. n.d. "FAIR Self-Assessment Tool." Accessed March 20, 2021. https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool

Bascik, Teresa, Philippe Boisvert, Alexandra Cooper, Martine Gagnon, Mark Goodwin, JohnHuck, Amber Leahey, Michael Steeleworthy, and Sally Taylor. 2020. "Dataverse North Metadata Best Practices Guide: Version 2.0." Vancouver: University of British Columbia Library. http://hdl.handle.net/2429/73609

Brigham, Doug. 2020. "Creating a README for Your Dataset: Quick guide." Zenodo. https://doi.org/10.5281/zenodo.4058971

Brodeur, Jay, Kaitlin Newson, and Sandra Sawchuk. 2019. "Canadian Data Curation Forum, CARL Portage." Accessed June 24, 2021. https://data-curation.github.io

Centre for Applied Data Ethics. 2021. "Ethical considerations in the use of geospatial data for research and statistics." London: UK Statistics Authority. Accessed June 15, 2021. https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-geospatial-data-for-research-and-statistics/pages/1

Chapman, Arthur D. 2020. "Current Best Practices for Generalizing Sensitive Species Occurrence Data." Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-5jp4-5g10

Clary, Erin, Jason Brodeur, Lee Wilson, Jeff Moon, and Shahira Khair. 2020. "Conceptualizing a National Approach to Data Curation Services in Canada." Zenodo. https://doi.org/10.5281/zenodo.3894935

Cornell University Research Data Management Service Group. n.d. "Guide to Writing "Readme" Style Metadata." Accessed March 20, 2021. https://data.research.cornell.edu/content/readme

Creative Commons. n.d.a "Frequently Asked Questions: Can I apply a Creative Commons License to Software?" Last modified August 28, 2020. https://creativecommons.org/faq

————. n.d.b "License Chooser." Last modified March 9, 2021. https://chooser-beta.creativecommons.org

Darragh, Jen, Alicia Hofelich Mohr, Shanda Hunt, Rachel Woodbrook, Dave Fearon, Jennifer Moore, and Hannah Hadley. 2020. "Human Subjects Data Essentials Data Curation Primer." Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/216579

Data Curation Network. n.d. "The DCN Curation Workflow." Accessed March 20, 2021. https://datacurationnetwork.org/outputs/workflows

————. 2019. "Data Curation Network Primers." [Collection]. Minneapolis: University of Minnesota Libraries Digital Conservancy. Last modified March 11, 2021. https://hdl.handle.net/11299/202810
Dataverse Project. n.d. "Sample Data Usage Agreement." Accessed March 28, 2021. https://dataverse.org/best-practices/sample-dua

DataverseNO. n.d. "Prepare Your Data: Preferred File Formats". Accessed on March 28, 2021. https://site.uit.no/dataverseno/deposit/prepare

Fankhauser, Eliane, Jerry de Vries, Nina Westzaan, and Vesa Åkerman. 2019. "SATIFYD: Self-Assessment Tool to Improve the FAIRness of Your Dataset." Data Archiving and Networked Services. Accessed March 20, 2021. https://satifyd.dans.knaw.nl

GitHub. n.d. "Choose an Open Source License." Last modified March 23, 2021. https://choosealicense.com

Government of Canada. Innovation, Science and Economic Development Canada. 2018. "Frequently Asked Questions Tri-Agency Research Data Management Policy." Last modified March 15, 2021. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97609.html#1d

Hunt, Shanda, Alicia Hofelich Mohr, and Rachel Woodbrook. 2021. "Consent Forms Data Curation Primer." Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/218838

Institute for Quantitative Social Science. n.d. "Dataverse Project: User Guide. Version 5.3." Last modified December 10, 2020. Cambridge: Harvard University. https://guides.dataverse.org/en/5.3/user

Johnston, Lisa R. 2017. "Curating Research Data Volume Two: A Handbook of Current Practice. Chicago: Association of College and Research Libraries." Retrieved from the University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/185335

Johnston, Lisa R., Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart, Mara Blake, Joel Herndon, J., Timothy M. Mcgeary, Elizabeth Hull, and Elizabeth Coburn. 2018. "Data Curation Network: A Cross-institutional Staffing Model for Curating Research Data." *International Journal of Digital Curation* 13(1): 125-140. https://doi.org/10.2218/ijdc.v13i1.616

Khair, Shahira. 2020. "Deposit Guidelines for UVic Dataverse." University of Victoria Libraries, Research Data Services. Last modified May 6, 2020. https://libguides.uvic.ca/ld.php?content_id=35154390

Lafferty-Hess, Sophia, Julie Rudder, Moira Downey, Susan Ivey, Jennifer Darragh, and Rebekah Kati. 2020. "Conceptualizing Data Curation Activities Within Two Academic Libraries." *Journal of Librarianship and Scholarly Communication* 8(1): eP2347. https://doi.org/10.7710/2162-3309.2347

Open Knowledge Foundation. n.d. "Open Data Commons: Legal Tools for Open Data." Accessed March 28, 2021. https://opendatacommons.org

Open Source Initiative. n.d. "Licenses & Standards." Accessed on March 28, 2021. https://opensource.org/licenses

Panel on Research Ethics. 2021. "Depositing Existing Data in Public Repositories: Guidance in Applying TCPS 2." Ottawa: Government of Canada. Accessed on June 15, 2021. https://ethics.gc.ca/eng/depositing_depots.html

Portage Network. 2020a. "CARL Portage Funds Scholars Portal Dataverse to Acquire Repository Storage for Canadian Research Data." Last modified November 16, 2020. https://portagenetwork.ca/news/carl-portage-funds-scholars-portal-dataverse-to-acquire-repository-storage-for-canadian-research-data

————. 2020b. "Curation Expert Group: Dataverse Curation Guide Working Group." Last modified January 27, 2021. https://portagenetwork.ca/network-of-experts/curation-expert-group

Portage Network COVID-19 Working Group. 2020. "De-Identification Guidance." Accessed March 29, 2021. https://portagenetwork.ca/tools-and-resources/rdm-guidance-for-covid-19/de-identification-guidance

Scholars Portal. n.d. "Scholars Portal Dataverse Guide." Accessed March 28, 2021. https://learn.scholarsportal.info/all-guides/dataverse

Sensitive Data Expert Group. 2020a. "Sensitive Data Toolkit for Researchers, Part 1 – Glossary of Terms for Sensitive Data Used for Research Purposes." Zenodo. http://doi.org/10.5281/zenodo.4088946

————. 2020b. "Sensitive Data Toolkit for Researchers, Part 2 – Human Participant Research Data Risk Matrix." Zenodo. http://doi.org/10.5281/zenodo.4088954

————. 2020c. "Sensitive Data Toolkit for Researchers, Part 3 – Research Data Management Language for Informed Consent." Zenodo. http://doi.org/10.5281/zenodo.4107178

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E. Bourne. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18

Yu, Jonathan, and Simon Cox. 2017. "5-Star Data Rating Tool, v4." The Commonwealth Scientific and Industrial Research Organisation. Software Collection. Accessed March 20, 2021. https://doi.org/10.4225/08/5a12348f8567b

Yu, Jonathan. n.d. "OzNome: Data Ratings." Last modified May 22, 2018. https://confluence.csiro.au/display/OZNOME/Data+ratings