# Dataverse Curation Guide

October 1, 2021

*The Dataverse Curation Guide Working Group*

Prepared by the Curation Expert Group Dataverse Curation Guide Working Group for the New Digital Research Infrastructure Organization

NDRIO | NOIRN
New Digital Research Infrastructure Organization | Nouvelle organisation d'infrastructure de recherche numérique

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

# Table of Contents

# 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.

This guide provides step-by-step instructions for curating new datasets deposited in Dataverse. Data curation is the active management of research data as it is created, maintained, used, archived, shared, and reused.[2] The guide is framed around the acronym CURATION to provide an easy reminder for curators, especially those starting out, of the main steps in the curation process. This framework is adapted from the Data Curation Network's CURATED steps[3] for use in a bilingual context and is intended to outline and provide guidance on curation best practices in Dataverse. Data curation is not always a linear process; the type of data you are working with, your institutional policies or practices, your comfort level with curation, and the amount of time the researcher is able to dedicate to the process may require you to skip steps or complete the steps in a different order. You may also need to circle back and complete some of the steps a second time. The level of curation your institution can offer, given competing priorities and number of staff dedicated to the curation service, may also determine how many curation steps you can complete.

Our Guide acknowledges 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, and institutional strategic direction. As a result, the Guide has been developed with flexibility in mind. It can be used by new or experienced curators within academic institutions of all sizes, and it can be adapted by institutions to support local policies and procedures. Three common data curation service scenarios are defined below, followed by the three possible levels of curation that we used to frame this guide.

[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/wp-content/uploads/2019/09/Curation_Primer_Aug2019_EN.pdf

[3] https://datacurationnetwork.org/outputs/workflows/ and https://doi.org/10.2218/ijdc.v13i1.616

## Possible Curation Service Scenarios

| | |
|---|---|
| **Unmediated Curation** | There is no intervention from the RDM service. 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. The researcher submits data to their dataverse (or dataset). Depending on local policy, the dataset is either flagged for review by the institutional Dataverse administrators, or the depositor requests 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. |

*Table 1: Possible Curation Service Scenarios.*

## Levels of Curation (Or, How to Use the CURATION Checklist)

The CURATION checklist is divided into three levels: Level 1, Level 2 and Level 3. Level one is the basic required information that should be completed to publish a dataset in Dataverse. Depending on the level of service that your institution can provide,[4] you may be able to complete some of the items in levels 2 and 3. A description of each level is below:

| | |
|---|---|
| **Level 1** | The minimum steps required to successfully publish in Dataverse and make the dataset findable, e.g., 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 dataset 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. |

*Table 2: Levels of Curation*

[4] For more information on levels of curation and how they were applied in practice at two institutions, please see Lafferty-Hess, Sophia, Julie Rudder, Moira Downey, Susan Ivey, Jennifer Darragh, and Rebekah Kati. "Conceptualizing Data Curation Activities Within Two Academic Libraries." Journal of Librarianship and Scholarly Communication 8, no. 1 (July 19, 2020): eP2347. https://doi.org/10.7710/2162-3309.2347.

# Complementary Guides

This Curation Guide should be used alongside several 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.

### Dataverse North Metadata Best Practices Guide[5]

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[6]

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**.[7]

### Data Curation Network (DCN) Curation Workflow[8]

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 Network Data Primers[9]

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.

---

[5] http://hdl.handle.net/2429/73609

[6] https://dataverse.scholarsportal.info/guides/en/latest/user/

[7] https://learn.scholarsportal.info/all-guides/dataverse/

[8] https://datacurationnetwork.org/outputs/workflows/

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

## The CURATION Guide

| | |
|---|---|
| **C** | Check<br>Ensure that all the data and metadata components required to successfully publish the dataset are present and in working order. |
| **U** | Understand<br>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. |
| **R** | Recommend<br>Request additional information from the depositor or suggest changes to the metadata and files that will improve findability and usability of the data in accordance with the FAIR principles. |
| **A** | Augment<br>Enhance the submission to facilitate discoverability and usability. |
| **T** | Transform<br>Ensure the dataset is using as many open and common formats as possible. |
| **I** | Include<br>Facilitate the reuse, proper attribution, and credit of data by including relevant persistent IDs and appropriate licencing information. |
| **O** | Optimize<br>Evaluate the overall FAIRness of the dataset and take steps to optimize the findability, accessibility, interoperability and reusability of the data. |
| **N** | Note Down<br>Ensure that you have made an accurate, written record of your curation work. |

# Check

At the **Check**[10] step, confirm that all data and metadata components required by the system to successfully publish the deposit are present and supporting documentation is included. If possible, identify any characteristics that may require special consideration (e.g., data with disclosure risk, or data obtained from a third-party source).

## Level 1

| Yes | No | Some issues | N/A | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | The dataset has been submitted to the proper dataverse.<br><br>To evaluate whether the dataset is in the most appropriate dataverse, consider the following:<br>▪ Has the researcher (or their research group) previously created or submitted to a dataverse?<br>▪ Does the dataset require its own dataverse or is there an associated dataverse to which it belongs?<br>▪ 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.<br>▪ 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.<br>▪ Use the **Dataverse North Metadata Best Practices Guide**[11] to evaluate completeness and accuracy of required metadata fields (Title, Author Name, Contact Email, Description, and Subject). |

[10] 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.

[11] http://hdl.handle.net/2429/73609

# Understand

In the **Understand**[12] step, 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. You may also screen for disclosure risk, intellectual property rights infringements, and other tasks, dependent on your institution's policies and the level of curation service your repository provides.

For a thorough overview of steps you might take to understand the data and assess its completeness and usability, see Step 3.0 in *Curating Research Data Volume Two: A Handbook of Current Practice*.[13] The Data Curation Network's Curation Primers[14] are another excellent resource with guidance and best practice advice for curating specific file types. They include information about tools for file review and specify information that will be necessary to ensure the usability of the data over time.

## 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[15] and University of British Columbia's Creating a README for your Dataset: Quick Guide[16] 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. |

[12] 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.

[13] https://hdl.handle.net/11299/185335

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

[15] https://data.research.cornell.edu/content/readme

[16] https://doi.org/10.5281/zenodo.4058971

| | | | | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | 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. |
| ❑ | ❑ | ❑ | ❑ | Description of the computing environment required to run any code that has been included (operating system, software packages and dependencies). |

### 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:<br>❑ Summary description<br>❑ File formats (flagged if proprietary)<br>❑ File size<br>❑ Path and/or tree structure[17] |

[17] Consider using tools to automate this output:

- Windows: use the TREE function from the command line.
- Mac: use Homebrew to install the tree package.
- Linux: install and use the tree command.

| | | | | |
|---|---|---|---|---|
| | | | | ❑ A checksum that can be used to verify data integrity (e.g., md5, sha256)[18] |

### 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,[19] are general best practices, and conventions for specific data types or discipline may differ. For further information, the University of Ottawa's guide to file naming and organization of data[20] may also be helpful.

| 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 suggest an alternate solution for publishing and archiving software and/or review the code and associated documentation alongside the data deposit.

[18] Consider using tools to automate this output:
- Windows: use the program HASHMYFILES.
- Mac: use the function *md5* or *shasum -a 256* on the command line.
- Linux: use the function *md5sum* or *sha256sum* on the command line.

[19] https://libguides.uvic.ca/ld.php?content_id=35154390

[20] https://biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data

| 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?<br>▪ If yes, consider asking the researcher to archive it in the Software Heritage archive[21] or Zenodo[22] and link to it from the dataset metadata record.<br>▪ 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 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:<br>▪ The code contains header information such as: author, version number, filename, license<br>▪ The function or purpose of the code is clear (from the Readme and/or embedded description)<br>▪ If applicable, the depositor included information about how to run the code<br>▪ The required software packages and dependencies are listed<br>▪ Comments are concise and clear and describe the intention of the line(s) of code that follow, OR<br>▪ The code itself is expressive (can be understood by humans and machines) |
| ❑ | ❑ | ❑ | ❑ | The Readme or a header in the code itself includes information about:<br>▪ The license.<br>▪ The developer's name and contact information.<br>▪ The version, and the date the code was last modified and/or run.<br>▪ Any sources the code (or part of the code) was derived from.<br>▪ Information about required input and expected output.<br>▪ Instructions on how to install or use the code. If there are multiple script files, the order in which they are run should be clear.<br>▪ Required software packages and dependencies.<br>▪ Information about the environment in which the code was developed and/or can be used. |

[21] https://www.softwareheritage.org/save-and-reference-research-software/
[22] https://guides.github.com/activities/citable-code/

### Submission contains potential sensitivities

**Note**: Most instances of Dataverse allow researchers to restrict access to datasets, either temporarily or in perpetuity; however, 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 Dataverse.

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 human participants data, 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 researchers to mitigate disclosure risk. If you do so, reiterate to the depositor 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:

- Government of Canada Panel on Research Ethics: Guidance on Depositing Existing Data in Public Repositories[23]
- Portage Network Sensitive Data Toolkit for Researchers
  - Part 1 – Glossary of Terms for Sensitive Data Used for Research Purposes[24]
  - Part 2 – Human Participant Research Data Risk Matrix[25]
  - Part 3 – Research Data Management Language for Informed Consent[26]
- Data Curation Network Data Curation Primers
  - Curation of Data Collected by Informed Consent[27]
  - Human Participants Data Essentials[28]

[23] https://ethics.gc.ca/eng/depositing_depots.html
[24] https://doi.org/10.5281/zenodo.4060158
[25] https://doi.org/10.5281/zenodo.4060448
[26] https://doi.org/10.5281/zenodo.4060460
[27] https://hdl.handle.net/11299/218838
[28] https://hdl.handle.net/11299/216579

- Portage Network De-Identification Guidance[29]
- Global Biodiversity Information Facility: Current Best Practices for Generalizing Sensitive Species Occurrence Data[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).

| Yes | No | Unsure | N/A | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | Dataset includes information about:<br>▪ Human participants (interviews, survey responses, biomedical, health-related data, etc.)<br>▪ Indigenous topics/subjects<br>▪ Minors or people unable to provide informed consent<br>▪ Vulnerable species<br>▪ Protected/private property<br>▪ Illegal/offending content<br>▪ 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:<br>▪ Consent form<br>▪ Participation agreement<br>▪ Data sharing agreement<br>▪ 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 human subjects data. |

[29] https://portagenetwork.ca/tools-and-resources/rdm-guidance-for-covid-19/de-identification-guidance/
[30] https://doi.org/10.15468/doc-5jp4-5g10

### 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 your campus to determine how your organization's policies regarding third-party intellectual property and rights affect deposit into your repositories.

| Yes | No | Unsure | N/A | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | Dataset contains proprietary or restricted information.<br>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

In the **Recommend** step, you may request additional information from the researcher and suggest changes to metadata and files that will improve the findability and usability of the dataset in accordance with the FAIR principles. 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 likely 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 this deposit may consider them and 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. *Curating Research Data Volume Two: A Handbook of Current Practice*,[31] section 3.5 includes a sample message and more information to help you get started.[32]

[31] https://hdl.handle.net/11299/185335

[32] A variant of this template is available in Appendix 2.

# Augment

The **Augment** step is an opportunity to enhance the 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[33] 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 issues | 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) |

[33] http://hdl.handle.net/2429/73609

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

Objects published externally may provide additional context to the new deposit and 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. When possible, provide direct links out to associated publications, code, models, documentation, source data and other related resources.

| Yes | No | Some issues | 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

The goal of the **Transform** step is to produce datasets that use open and common formats when possible. Open formats will help guarantee 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 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.

## 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… |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | Does the dataset contain proprietary file formats or formats that are not suitable for preservation? For guidance, see DataverseNO[34] preferred files formats or Guide concernant les formats recommandés par Bibliothèque et Archives nationales du Québec[35] [in French]. |
| ❑ | ❑ | ❑ | ❑ | Can proprietary files be transformed to non-proprietary formats without data loss? |
| ❑ | ❑ | ❑ | ❑ | If files cannot be transformed, does the documentation describe:<br>❑ The file format and contents of the files?<br>❑ The software or instrument that generated the files?<br>❑ The software necessary to view or use the files?<br>❑ 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). |

[34] https://site.uit.no/dataverseno/deposit/prepare/

[35] https://www.banq.qc.ca/documents/archives/archivistique_gestion/publications_proposees/Guide-formats-BAnQ_Final.pdf [in French]

# Include

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 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, need to be added manually.

A license will describe the acceptable uses of the published dataset. The default license in Dataverse is the CC0 public domain dedication,[36] but you may need to adjust this. Ideally the license selected for the dataset will meet 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[37] and GitHub's Choose an Open Source License,[38] and resources such as the Open Source Initiative,[39] the Free Software Foundation[40] (i.e, the GNU license), and the Open Knowledge Foundation[41] can help select an appropriate license. Dataverse has also published a template that may be useful for creating a data use agreement.[42] Please be aware that Creative Commons recommends against using their licenses for code or software.[43]

## 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. |

[36] https://creativecommons.org/publicdomain/zero/1.0/
[37] https://creativecommons.org/choose/
[38] https://choosealicense.com/
[39] https://opensource.org/licenses
[40] https://www.gnu.org/licenses/license-list.en.html
[41] https://opendatacommons.org/
[42] https://dataverse.org/best-practices/sample-dua
[43] https://creativecommons.org/faq/#can-i-apply-a-creative-commons-license-to-software

| | | | | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | The system-assigned data DOI is registered and resolves to the dataset landing page (if curation takes place before publication, you will not be able to confirm this until after curation is published). |

## 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.

| Yes | No | Unsure | N/A | |
|---|---|---|---|---|
| ❑ | ❑ | ❑ | ❑ | Has the researcher selected a license already?<br>❑ 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?<br>❑ 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?<br>❑ If yes, consult with the depositor to suggest a more appropriate license.<br>❑ If no, you can move forward with the license the depositor selected. |
| ❑ | ❑ | ❑ | ❑ | Is proper attribution provided for data obtained from third-party sources or derived from existing datasets? |
| ❑ | ❑ | ❑ | ❑ | Readme includes a preferred citation for the deposited dataset. |

# Optimize

At the **Optimize** step, the overall FAIRness[44] 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[45] (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.
- The CSIRO 5-star Data Rating Tool[46] is a self-assessment rating scheme to evaluate FAIRness. The assessment criteria are adapted from the OzNome Data Ratings criteria[47], which provide specific metrics for achieving FAIR and trusted data.
- The Australian Research Data Commons FAIR Self-Assessment Tool[48] 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."

[44] https://doi.org/10.1038/sdata.2016.18
[45] https://satifyd.dans.knaw.nl/
[46] https://doi.org/10.4225/08/5a12348f8567b
[47] https://confluence.csiro.au/display/OZNOME/Data+ratings
[48] https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/

# Note Down

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 may 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 *Curating Research Data Volume Two: A Handbook of Current Practice*,[49] section 3.2.

## 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:<br>❑ DOI, internal item ID, or another way to link the log file to the dataset<br>❑ Any potential issues uncovered as part of the curation process<br>❑ Questions and high-level change recommendations for the researcher<br>❑ A high-level summary of the researcher's response to requests and questions<br>❑ Any changes made to the dataset during the curation process, including:<br>❑ Changes made to existing files or documentation<br>❑ The names of any files added or removed from dataset, and why<br>❑ A list of metadata terms that were changed, added or removed |

[49] https://hdl.handle.net/11299/185335

# References


Australian Research Data Commons. n.d. FAIR Self-Assessment Tool. Accessed on 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, John Huck, 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.

Bibliothèque et Archives nationales du Québec. 2020. Guide concernant les formats recommandés par BAnQ. Accessed on June 2, 2021. https://www.banq.qc.ca/documents/archives/archivistique_gestion/publications_proposees/Guide-formats-BAnQ_Final.pdf. [In French].

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

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

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

Creative Commons. n.d.-a. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Accessed on June 9, 2021. https://creativecommons.org/publicdomain/zero/1.0/.

Creative Commons. n.d.-b. Choose a License. Accessed on June 2, 2021. https://chooser-beta.creativecommons.org/.

Creative Commons. n.d.-c. Frequently Asked Questions: Can I apply a Creative Commons License to Software? Accessed on September 8, 2021. https://creativecommons.org/faq/.

Curation Expert Group. 2019. Primer: Data Curation. Portage Network. Accessed on September 8, 2021. https://portagenetwork.ca/wp-content/uploads/2019/09/Curation_Primer_Aug2019_EN.pdf.

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. 2019. Data Curation Primers. [Collection]. University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/202810.

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

Dataverse Project. n.d. Sample Data Usage Agreement. Accessed on 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 on March 20, 2021. https://satifyd.dans.knaw.nl/.
Free Software Foundation. 2021. GNU Operating System: Various Licenses and Comments about Them. Accessed June 16, 2021. https://www.gnu.org/licenses/license-list.en.html.
GitHub. n.d.-a. Choose an Open Source License. Last modified March 23, 2021. https://choosealicense.com/.

GitHub. n.d.-b. Making your code citable: GitHub Guides. Accessed on September 7, 2021. https://guides.github.com/activities/citable-code/

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

Government of Canada. 2021. Guidance on Depositing Existing Data in Public Repositories. Panel on Research Ethics. Last modified May 25, 2021. https://ethics.gc.ca/eng/depositing_depots.html.

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, 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 on March 28, 2021. https://opendatacommons.org/.

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

Portage Network COVID-19 Working Group. 2020. De-Identification Guidance. Accessed on 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 on March 28, 2021.

https://learn.scholarsportal.info/all-guides/dataverse/.

Sensitive Data Expert Group. 2020-a. 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.

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

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

Software Heritage. n.d. Save Research Software. Accessed on 8 February 2021. https://www.softwareheritage.org/save-and-reference-research-software/.

University of Ottawa Library. Research Data Management Team. n.d. File naming and organization of data. Accessed on June 9, 2021. https://biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, … Barend Mons. 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 on 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.

# Appendix 1 – Examples of Curated Datasets in Dataverse

This appendix contains examples of curated datasets published in Scholars Portal Dataverse. The first set of examples show the minimum level of curation required to deposit data in Dataverse. The next set of examples address different elements of curation and provide examples of good curation practice.

## Level 1

Level 1 curation steps prepare datasets for publication in Dataverse. It is the minimum level of curation required to successfully publish in Dataverse and to make the dataset findable. Level 1 steps include confirmation that the required metadata fields are complete and accurate. Readme files and file-level curation may occur.

- Ipsos Canadian Public Affairs Dataverse - https://dataverse.scholarsportal.info/dataverse/Ipsos
- Ipsos Global @dvisor Dataverse (University of Toronto) - https://dataverse.scholarsportal.info/dataverse/Ipsos_UofT
- Winseck, Dwayne, 2020, "Canadian Media Concentration Research Project Dataset 2019", https://doi.org/10.5683/SP2/YSZMOQ, Scholars Portal Dataverse, V1
- Hamarat, Yaprak, 2019, "Données photographiques pour la thèse de doctorat intitulée « L'esthétique de l'engagement écologique : l'impensé des politiques environnementales »", https://doi.org/10.5683/SP2/IS0SWH, Scholars Portal Dataverse, V1

## General Examples

These datasets illustrate elements of well-curated datasets.

- Margaret B. Harrison; Practice and Research in Nursing Group: Wound Care Collaborative; Ian Graham; E. Andrea Nelson; Elizabeth VanDenKerkhof; Karen Lorimer; Connie Harris; Meg Carley; The Canadian Bandaging Trial Group, 2013, "Practice and Research in Nursing (PRN) Wound Studies, 1999-2009 [Canada]", https://hdl.handle.net/10864/CORX8, Scholars Portal Dataverse, V6
    - This dataset provides an overview of the studies used to create the dataset and a detailed description of how the data was processed. The curation team worked with the data manager to ensure the confidentiality of participants was maintained while maximizing the portion of the dataset open for reuse.
- Agricultural and Forest Meteorology Group. Elora Research Station/Guelph Turfgrass Institute., 2018, "Weather records for the Elora Research Station, Elora, Ontario [Canada]: Meteorological data 2018", https://doi.org/10.5683/SP/RQRDSH, Scholars Portal Dataverse, V3.
    - This dataset includes a clear description that summarizes the data that were collected and anomalies that affected recording. It contains a well-structured tabular dataset and codebook.
- Paul, Jason; Baltzer, Jennifer L.; Kokelj, Steve V., 2020, "Near-surface permafrost ground ice characteristics and ecological and physical drivers of transient layer ice content in discontinuous permafrost near Yellowknife, NT", https://doi.org/10.5683/SP2/LX5IJN, Scholars Portal Dataverse, V1
    - The dataset description summarizes data collection, location, and method. Metadata are enriched with geospatial information for improved indexing and search. Individual files include descriptive metadata that identifies their content prior to download. All variables are defined within the spreadsheets. Files describe all variables in all spreadsheets. The Readme file

includes methodological information, contact information, and a file overview for reference. The researcher has selected a CC BY-NC-SA 4.0 license and end-users must agree to the Terms of Use before the data can be downloaded.

- Moubayed, Anna-Maria, 2018, "All About Eve and Other Stories: Representations of Eve in French Romanesque Sculpture (and more)", https://doi.org/10.5683/SP/2Y5C1X, Scholars Portal Dataverse, V1
    - An example of a dataset from the humanities. In addition to the data file, all of the field research notebooks are included, along with a link to an interactive map and the files needed to recreate the map.
- Gagné, Monique; Ward, W. Peter, 2019, "Birth weight and economic growth data sets, Utrecht Hospital, 1880-1940, [2012]", https://doi.org/10.5683/SP2/WNY6FG, Scholars Portal Dataverse, V1.
    - This publication contains the original dataset, which has been augmented with a codebook, a description of the record layout, and a data dictionary to facilitate reuse.

# Appendix 2 – Templates for Correspondence

The templates below can be used to correspond with depositors. You may need to modify a template to more accurately reflect the dataset you are curating, or you may wish to modify a template to better reflect your voice. In some cases it may be more appropriate to use the word "Question" instead of "Recommendation". Please note that text in brackets should be removed or replaced with details of the dataset you are curating before you send your message.

## Change request: submission not returned to depositor

***Subject line***: Recent Dataverse submission ["Shortened dataset title here"]

Dear [NAME],

Thank you for your recent submission ["Full title of the dataset"] to [Repository Name Here]. [You may want to include something positive! E.g., complement documentation, interesting study, etc.]. After we receive a dataset, we review it to ensure the datasets we host are as complete, accessible and understandable as possible. We have reviewed your submission and have the following recommendations for you:

Recommendation #1
Recommendation #2
Recommendation #3
Recommendation #4

If you are willing to make these changes please [XXX] [E.g., please email the revised README to us and we will use it to overwrite the existing version in your submission; OR, please confirm, and we will update the metadata and publish your submission].

Please don't hesitate to contact us if you have any questions or concerns.
Sincerely,
[Your name here]

## Change request: submission returned to depositor for revision

***Subject line***: Recent Dataverse submission ["Shortened dataset title here"]

Dear [NAME],

Thank you for your recent submission ["Full title of the dataset"] to [Repository Name Here]. [You may want to include something positive you noticed about the dataset! E.g., complement documentation, interesting study, etc.]. After we receive a dataset, we review it to ensure the datasets we host are as complete, accessible and understandable as possible. We have reviewed your submission and have the following recommendations for you:

Recommendation #1

Recommendation #2
Recommendation #3
Recommendation #4

We have returned your submission so you may modify your files. To access it, please log in to [XXXX]. When your changes are complete, please verify that everything looks accurate and to resubmit your dataset.

If you have questions about any of our recommendations, or if you experience any difficulty accessing or editing your submission, please let us know.
Sincerely,
[Your name here]

# Appendix 3 – Additional Curation Resources

The NDRIO Portage Curation Expert Group (CEG) stewards the Curation Commons, a space for research data management and curation practitioners to share resources related to the practice of data curation. Its purpose is to help build data curation capacity and support the curation community in Canada and beyond.

Additional resources - related generally to curation and specifically to curation in Dataverse - have been indexed in the *Curation Commons.*

# Appendix 4 – Dataverse CURATION Quick Reference Guide

- **Level 1:** This level of curation prepares datasets to be successfully published in Dataverse and is the minimum level of curation required to make the dataset findable.
- **Level 2:** This level of curation enhances the discoverability and helps ensure the usability of datasets over time.
- **Level 3:** This level of curation prepares datasets for reproducibility, and preservation. This level includes:

| Letter | Definition | Major Tasks | Level 1 | Level 2 | Level 3 |
|---|---|---|---|---|---|
| C | Check<br>Ensure that all the data and metadata components required to successfully publish the dataset are present and in working order. | ❑ Dataset has been submitted to the proper dataverse | X | | |
| | | ❑ All files described in the documentation are included in the dataset | X | | |
| | | ❑ Required metadata fields are accurate | X | | |
| | | ❑ Supporting documentation is included | X | | |
| | | ❑ The researcher has confirmed that the data is free of any licensing and intellectual property issues | X | | |
| | | ❑ The researcher has confirmed that the data is free of identifying/sensitive information | X | | |
| U | Understand<br>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. | ❑ Supporting documentation is thorough, accurate, and complete | | X | |
| | | ❑ Files open properly and contents appear as expected | | X | |
| | | ❑ Files and folders are named and structured appropriately | | X | |
| | | ❑ Code is well commented and produces the expected results | | | X |

| | | | | | |
|---|---|---|---|---|---|
| | | ❑ Submission contains potential sensitivities | | | X |
| | | ❑ Submission contains data or code from third party sources | | | X |
| R | **Recommend**<br>Request additional information from the depositor or suggest changes to the metadata and files that will improve findability and usability of the data in accordance with the FAIR principles. | ❑ Prioritize your recommendations in a list to determine which requests are critical or actionable, and which requests you may be able to live with if they are not fulfilled | X | | |
| | | ❑ Reach out to the depositor with a clear request for information | X | | |
| A | **Augment**<br>Enhance the submission to facilitate discoverability and usability. | ❑ Metadata is rich, accurate, and complete | | X | |
| | | ❑ Links to related publications, datasets, and other resources are included | | X | |
| T | **Transform**<br>Ensure the dataset is using as many open and common formats as possible. | ❑ File formats are open, or appropriately documented | | | X |
| I | **Include**<br>Facilitate the reuse, proper attribution, and credit of data by | ❑ Include persistent identifiers wherever possible | | X | |

| | | | | | |
|---|---|---|---|---|---|
| | including relevant persistent IDs and appropriate licensing information. | ❑ Review the licensing and terms of use for the dataset | | | **X** |
| **O** | **Optimize**<br>Evaluate the overall FAIRness of the dataset and take steps to optimize the findability, accessibility, interoperability and reusability of the data. | ❑ Evaluate the dataset and Optimize FAIRness | | | **X** |
| **N** | **Note Down**<br>Ensure that you have made an accurate, written record of your curation work. | ❑ Create a curation log to document your decisions and actions | | **X** | |