Borealis to House Sensitive Metadata (Draft V2)

Introduction

Originally conceptualised and formalised by Harvard’s Institute of Quantitative Social Sciences ([IQSS](https://www.iq.harvard.edu/)), [*The Dataverse Projec*t](https://dataverse.org/about) is an open-source repository software initiative that had as its goal to facilitate the sharing, preservation, citation, exploration, and analysis of research data, all the while proliferating the web-based accessibility of virtual archives. Now adopted across the globe in 128 independent *Dataverse* [installations](https://dataverse.org/), the concept of the Dataverse has proved to be a powerful tool in developing a wide network of research data repositories, ultimately increasing the inter-institutional and international findability of research datasets.

One such Dataverse installation is [Borealis](https://borealisdata.ca/), the Canadian Dataverse Repository. Hosted by [Scholars Portal](https://scholarsportal.info/) at the University of Toronto, Borealis is a shared bilingual service provided in partnership with Canadian regional academic library consortia, institutions, research organisations, and the Digital Research Alliance of Canada. Following the overarching tenets of the Dataverse Project, Borealis is dedicated to providing a multi-disciplinary and safe research repositories for its participating [members and partners](https://borealisdata.ca/#partners).

Borealis holds datasets for participating member institutions and research organisations in their own institutionally curated repositories called “Dataverses”. In essence, Borealis is the repository of those institutional Dataverses. This set up facilitates the findability and access to data - some of which may be restricted and may require owner authorisation for access. That being said, such restricted data may be more difficult to find given their more discreet and sensitive nature. For reasons relating to the nature of certain datasets (such as potentially identifiable information within their confounds), some researchers may omit altogether any attempt at making their data findable.

There are, however, approaches by which researchers could make their data findable all the while preserving the ethical integrity of their restricted data; in other words, make their data findable, but not accessible – therefore opening a safe and responsible gateway towards ethical data reuse. The present document presents an introductory guide in ethical data findability, curatorship, and custodianship.

Findable Data Without Access

Some datasets cannot find their way to Borealis for reasons pertaining to their [Terms of Use](https://borealisdata.ca/termsofuse/). In summary (though an in depth understanding of the Terms of Use is necessary), Borealis users are barred from uploading any files containing information that could directly or indirectly identify a subject or participant. The upload of any such datasets would not only violate the Borealis Terms of Use, but would also likely violate Research Ethic Board (REB) guidelines concerning the dataset or the study at hand.

Although Borealis does not permit the upload any identifiable datasets, it does not bar users from uploading anonymized metadata (data about the data) for these datasets. For instance, a sensitive medical trial dataset containing identifiable information would be unacceptable for upload; however, metadata such as the questions posed to the participants, the type of instruments used for assessment, variable labels and related metrics would be permissible. That is, information ***about*** the study dataset could be deemed well within the boundaries of permissibility, but information ***about*** participants or testing locations would not. It logically follows from this that, given our goals of making restricted data findable with heavily anonymized metadata, borealis could be a central tool in our on-going endeavor.

Borealis Open Science and Transparency: Pros and Cons

Over the years, there has been a push by the Canadian Tri-Council Agencies[[1]](#footnote-1) to liberalise the access to research data, stating that “as publicly funded organizations, the agencies are strong advocates for making the results of the research they fund as accessible as possible” and that “they aspire to advance knowledge, avoid research duplication and encourage reuse, maximize research benefits to Canadians and showcase the accomplishments of Canadian researchers”[[2]](#footnote-2). Of course, not all data can be made readily available given their sensitive nature, but an effort should be made to make them as findable as possible all the while maintaining the privacy of the participants (which should be the foremost priority).

There are various benefits to using one’s respective institutional Dataverse via Borealis as a repository for sensitive data metadata.

1. Borealis is a safe and secure repository that houses continually institutionally curated Dataverses.
2. Making information about the study dataset available can initiate conversations between researchers for the reuse of data (provided interinstitutional REB approval) and for potential joint ventures.
3. Using Borealis also has the added benefit of linking a persistent identifier (in this case a Digital Object Identifier, or DOI) to the research data dataset metadata. The attribution of a persistent and unique DOI to the dataset metadata facilitates citation and subsequent findability.
4. Upon publishing in Borealis, dataset files are automatically formatted in Data Documentation Initiative ([DDI](https://ddialliance.org/)) standard metadata files. DDI standards ensure a high degree of consistency in machine readability, meaning that variables within dataset files can be found with search engines (therefore considerably increasing findability and potential reuse).

There are certain qualms about using a national repository such as Borealis. Being a national service, Borealis data is often scraped by foreign academic colleagues in both peer and “near-peer” nations. Given the sensitivity of some data, how “findable” we want this data to be beyond the confounds of Canada remains a point of contention and will require a case by case assessment by the depositors’ institutions. However, as it will be outlined in the present document, the sensitive data itself is ***never*** made available on Borealis – only its metadata. This mitigates possible worries associated with potential bad actors; the present document therefore maintains that Borealis as a haven for sensitive data metadata findability.

Accessing Borealis

The institutional Dataverse of the depositor can be found by navigating to the Borealis [login page](https://borealisdata.ca/loginpage.xhtml), and then using the scroll down menu to select their institution to login via their institution SSO, or by navigating to the [members and partners](https://borealisdata.ca/#partners) web page, and selecting their respective institution. An in-depth guide on login and account creation procedures can be found on the [Scholars Portal learn Borealis](https://learn.scholarsportal.info/all-guides/borealis/user-accounts/) website.

Once logged-in, users should, by default, be on their institution Dataverse repository. There are various paths through which users can go about depositing their data and/or metadata. Depending on their institutional policies, they may be able to create their own Dataverse within their institution’s Dataverse repository. However not all user’s may have the required permissions to do so. In-depth instructions pertaining to the creation of proprietary Dataverses within institution’s Dataverse (and troubleshooting if the user does not have the required privileges) can be found in the [Collections](https://learn.scholarsportal.info/all-guides/borealis/collections/) Scholars Portal learn Borealis website.

Potential depositors should however consider whether or not this “Sensitive Metadata” Dataverse could be used by various other scholars to deposit their own metadata pertaining to restricted access datasets. If so, depositors could contact their institution Dataverse administrator to discuss the creation of such a Dataverse. Dataverse administrators can be contacted via the institution’s Dataverse page with the “Contact” button found on the top right of the Dataverse description.

What Can and “Cannot” Go in Borealis Dataverses

As previously mentioned, some datasets cannot be uploaded on any Borealis repositories (institutional or otherwise) if their violate their [Terms of Use](https://borealisdata.ca/termsofuse/). This is an important step in deciding whether or not the metadata can be uploaded in the repository. While this has also been previously discussed in the present document, here are two key questions to consider:

1. Do I have the authority to share this sensitive data or metadata (with or without restrictions? REB approval before or after study?)
2. Is my sensitive metadata free of any identifiable information? (all information that could be used to link it to someone are prohibited; to name a few, this includes variables such as height, weight, postal codes, diagnostics, etc.).

While some of these individual variables may seem too ambiguous to identify someone, they can always be used in conjunction with other variable information to triangulate where abouts or characteristics of study participants. It is important to note that data uploaded and then published on Borealis cannot be deleted – it can be [deaccessioned](https://learn.scholarsportal.info/all-guides/borealis/deleting-data/#Deaccessioning-a-Published-Dataset) by a super-administrator at Scholars Portal, but not fully deleted. This means that it cannot be removed without approval from the highest authority in the Borealis institutional hierarchy. This is not to scare away potential depositors from making their data findable, but simply to remind researchers of their due diligence when it comes to ethical custodianship of sensitive research data.

All this being taken into consideration, institutional Dataverses should be able to take in any file type. This includes video files, text curation files, audio files, statistical software files, and many more. Given the nature of the sensitive data and metadata being discussed, simple file formats, such as a .csv or .xlsx are optimal; this is discussed in greater details in the following section.

Pragmatics of Metadata Findability in Borealis

Assuming that everything aligns with Terms of Use and copyright, the upload process is relatively simple. Metadata in Borealis is generated both manually and automatically (assuming the uploaded file is tabular in nature such as .xlsx, .csv, .sav files – but not .RData or .json). This is precisely where we want to make an important decision: what level of metadata quality do we want to create and make findable?

Here we will discuss three hierarchal tiers, each building upon the prior, of metadata upload quality; a good – better – best tier hierarchy. With each increase in tier, the metadata becomes more findable and contextualised. The tier at which potential depositors decide to stop may largely be as a function of the sensitivity level of their restricted metadata.

Tier 1: Keywords and Descriptions – ***No File Upload***

Step 1: Creating a Dataset

Once in the appropriate Dataverse repository, users can easily create datasets (assuming they hold the appropriate privileges to do so – otherwise they should contact their institution Dataverse administrator if they require any help). An in-depth guide on how to create a dataset as a user can be found on Scholars Portal’ Borealis [Dataset creation](https://learn.scholarsportal.info/all-guides/borealis/datasets/) web page.

Step 2: Template Selection and Intellectual Property

Once the dataset creation process is initiated, users can select a dataset template (which vary between institutions and Dataverses). Selecting a template attributes copyright terms of use conditions.

By default, datasets in Dataverse are given [CC0 1.0](https://creativecommons.org/publicdomain/zero/1.0/legalcode.en) unless specified otherwise. This means that, by default, all uploaded data is considered Public Domaine. That being said, data depositors can input their own copyright guidelines to establish restrictions on how their uploaded data can be used. For instance, users can attribute [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/), which has more constraints on how the data can be used, how it must be cited, and whether or not it can be modified in subsequent iterations by different users (see [Creative Commons attributions](https://creativecommons.org/share-your-work/cclicenses/) for more examples).

Data depositors are not limited to Creative Commons attributions, and can create their own custom terms of use. Given the nature of the submitted data and metadata, it is advised the potential depositors ponder on how they would like to see their data be used, and establish a solid ground for restricted usage.

Step 3: Citation Metadata

Users must now create a title, description, select a subject category, and identify keywords for their dataset. This step is essential for base level dataset findability. All of this metadata will subsequently be auto-generated into a machine readable metadata file. It is advised to fill out as many as the metadata fields as possible to maximise findability, though not all will apply to every dataset. Once this is completed users may Save the Dataset and proceed to the next step. ***Do not upload any files yet***.

Step 4: Field Specific Metadata

After saving the dataset, depositors can go back and edit their metadata in newly available metadata sections (“Add + Edit Metadata” button). These sections include:

1. Geospatial Metadata
2. Social Sciences and Humanities Metadata
3. Astronomy and Astrophysics Metadata
4. Life Sciences Metadata
5. Journal Metadata
6. Computational Workflow Metadata

These field specific metadata are not mandatory, though they can significantly help in making in the dataset findable. Some of the field entries can help provide an idea of the non-published dataset (such as sample size, sampling procedure, type of research instrument, units of analysis, etc.) Note that not all fields will be useful, as some may divulge too much information about the dataset, which may not be desirable in some cases. As always, it is advised that depositors carefully consider what information they want to include in their metadata – this is usually on a “case by case” basis and will vary between depositors.

**Accomplished so far:**

* Creation of an empty dataset with a minted persistent identifier (in this case a DOI);
* Selection of a template/terms of use;
* Input of citation metadata (title, keywords, subject category);
* Input of additional field specific metadata to facilitate findability.

**Options to consider:**

Depositors now have two options: they can publish the dataset if they do not want to/cannot add files to the dataset, or they can pass onto Tier 2 to contextualise the dataset with supportive documentation. Depositors can also come back later after publishing to expand the dataset.

Tier 2: Non-Tabular File Upload (Metadata Contextualisation)

Step 5: Supporting Documentation

This step does not necessarily make the dataset more findable by means of search engines; that is, it will not add any substantial machine readable metadata. It will however contextualise the present metadata. While the manually added metadata allows for findability, it can be rather limited in describing what the dataset is a stand-in for.

In this case, a simple .txt, .docx, .pdf or other text format is sufficient to provide vague contextualisation for the dataset. This contextualising file can be used to explain the goal of the study at hand, rationale, collected data, etc. It is important to note that, despite not being machine readable, these files should in no way divulge any information about participants.

Although a relatively minor addition to the dataset, the addition of supporting documentation is an important step towards data reuse.

Tier 3: Tabular File Upload

Decision tree depicting the steps for file preparation. The tree goes as follows:

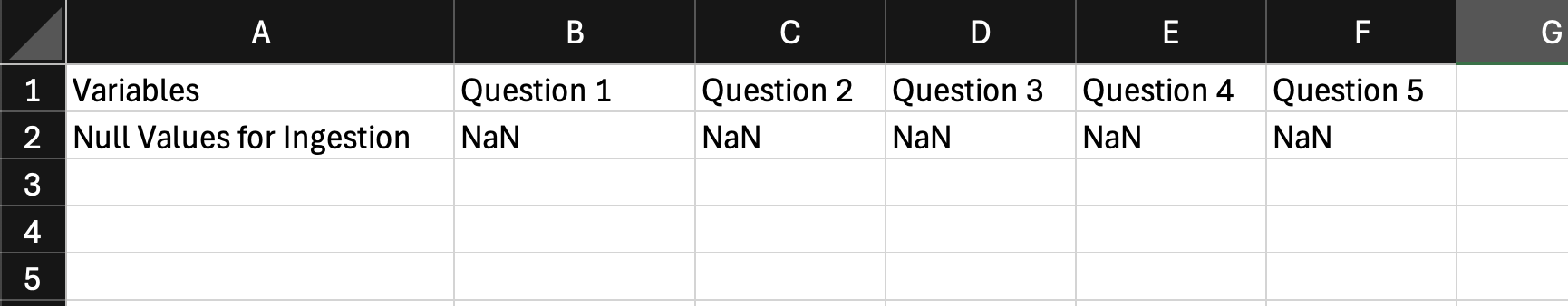
Does the study have REB approval for Data reuse? If no, then get REB approval before continuing. If yes, ask yourself: Can the data file be safely anonymized? (that is, can all potentially identifiable information be removed, such as locations, names, intersectionalities, etc.). 

The tree then splits in 2 major branches: No, and Yes.

Here is the « Yes » Branch. Ask yourself: is the data already anonymised? If yes, then you are ready to begin the upload process (with restrictions). If no, then make sure that all identifiable information be removed from the file before proceeding. 

Here is the « No » branch. If the data file cannot be safely anonymised, then the data file and study results should not be uploaded - however the metadata file could still be uploaded. Now ask yourself: can the metadata file be safely anonymized? If it is not, then do not pursue metadata sharing. If it is, then you are ready to create a file (which is discussed in the next section). Step 6: Thinking About Tabular File Preparation

Step 7: Creating Tabular File

Here is the dataset itself:

It really is as simple as it gets. The first row denotes the variables (imagine that each “Question” actually denotes a Variable of some sort). Note that the “Null” entries are necessary – they can just as well be interchanged for NaN or 0 or “Cheese” or any place holder. However, it is better to just use a string input NaN as it stops the software for calculating summary statistics that it would usually compute for data points in a column. Without these Null values (meaning the dataset is composed of a singular row), the file will not be ingested.

Now here is an example between a non-ingested DDI XML file, and one that is ingested, with this same .xlsx file (note that it would be the exact same outcome were it .csv, or any other – as long as it is not json or RData; this is an issue with Harvard, not Borealis per se).

Non-ingested output:

However if the file is ingested, it also has this following section in addition to the above metadata information (note that it is truncated on the right side and bottom):

This entire entry above is not included if the file is not ingested; this would mean that we would be missing out on a lot of machine readable data were we to not ingest the files. I have already written a python code to exclusively extract variable labels if ever this were needed.

We can also bring this up a notch in terms of metadata quality by including offered response types, but NOT who responded, and how they responded – though I still need to do some testing to see what works and what doesn’t.

Comment: So here’s where we might want the good-better-best. Good would be: you put some text content in borealis so that you get a PID and have something to point to (similar to what we’ve created for Lunaris). Then, better would be - some content/documentation is deposited alongside core metadata that will sort of describe vaguely what’s going on, then best, this XML variable level metadata with averages or not depending on whether this will disclose information (flag a link to k-anonymity code from the sensitive data toolkit) I will provide this after next draft.

Appendices

1. Canadian Institute of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC). [↑](#footnote-ref-1)
2. See <https://science.ised-isde.canada.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-statement-principles-digital-data-management> [↑](#footnote-ref-2)