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.

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). File Preparation

Borealis

The institutional Dataverse of the depositor can be found by navigating to the Borealis [login page](https://borealisdata.ca/loginpage.xhtml) or by navigating to the [members and partners](https://borealisdata.ca/#partners) web page, and selecting their respective institution.

file

Dataverse for Restricted Access Datasets

Before even talking about how to safely upload metadata pertaining to sensitive datasets, it is important to have an idea of how this would be done in the first place. The first step would be for the CRDCN to set up a Dataverse of their own within Borealis. It would require discussions with the Borealis team as to where this Dataverse would be set up. For instance, not all Dataverses are immediately visible on the main Borealis page; for the most part, only academic institutions are advertised on the main page. However, there are other repositories that are “hidden” (such as the Dataverse for Odesi, the Social Science data repository operated by Scholars Portal).

The CRDCN has two main options:

1. Create their own Dataverse (not my go to)
2. Odesi Dataverse (whose explicit purpose is to house public access Statistics Canada data)

There, the Dataverse is bilingual, well fleshed out, and continually curated (in part by yours truly, me and other staff members). That being said, I don’t speak on behalf of the service – though it nonetheless seems to be me to be a viable approach. Of course discussions would need to be held in order to hash out the details.

As it stands, Odesi has various Dataverses for a wide range of different data types. Given the nature of the sensitive Metadata and of the present project, not all would be suitable – nevertheless here is the list of existing Odesi Dataverses:

1. Census of Population
2. Statistics Canada Aggregate Data
3. Statistics Canada Public Use Microdata Files (PUMFs)
4. Canadian Public Opinion Polls
5. Statistics Canada Data Liberation Initiative (DLI) Licensed Data
6. **Other Canadian and Government Data**
7. International Data

As readers of the present document may have astutely noticed, (6) Other Canadian and Government Data was bolded to denote an emphasis. It is of my belief that this could be a viable repository for the present endeavour. So let us keep this Dataverse in mind and proceed to other points of discussion.

What Can and “Cannot” Go in Borealis Dataverses (Including Odesi)

Borealis does have extensive documentation about what can and cannot go into their data repositories (this is for very obvious legal reasons). I therefore recommend an in-depth read of [Borealis’ terms of use](https://borealisdata.ca/termsofuse/) page. For the sake of brevity, here are screenshots of the Terms of Use directly relevant to the subject at hand (though of course the entire Terms of Use should be carefully reviewed in order to build a convincing argument/case as to why our goals are justified, safe, and necessary for the preservation of metadata.

Our case has to be air-tight with a clear goal. Data uploaded to Borealis cannot be deleted – it can be [deaccessioned](https://learn.scholarsportal.info/all-guides/borealis/deleting-data/#Deaccessioning-a-Published-Dataset), but not fully deleted. Meaning that it cannot be removed without approval from the highest authority in the Borealis institutional hierarchy.

All this being taken into consideration, Borealis takes almost everything. As an example, a student (I hope not faculty) at Western has been uploading [Harry Potter self-insert fanvids](https://borealisdata.ca/dataverse/esapuridis) in their own Western Dataverse (which *technically* could be infringing on the fanvids’ original owners’ copyright and intellectual property rights, but I digress). The University of Toronto also has a collection of 240+ [pictures of icicles](https://borealisdata.ca/dataverse/Icicle_Atlas). My point is that anything goes in and is minted a DOI.

Copyright and Intellectual Property

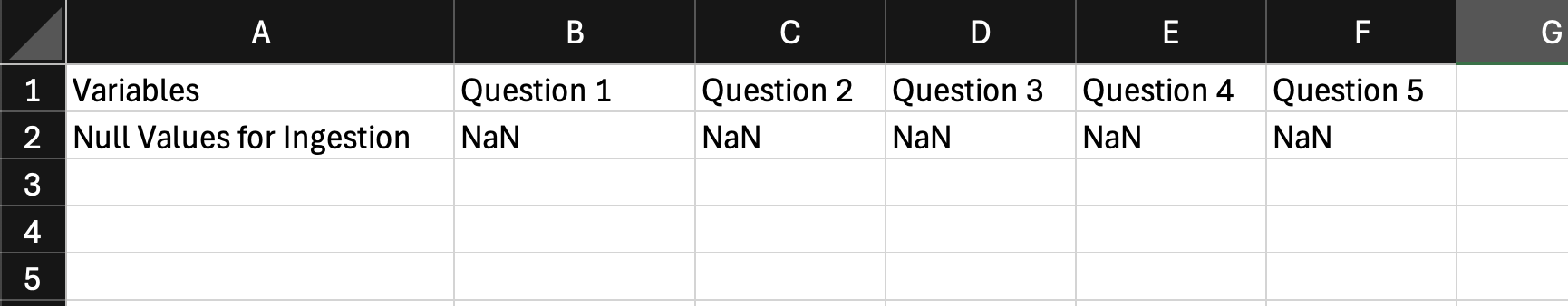
Copyright terms of service are of paramount importance on Borealis (for serious data, at least). To briefly come back to the Harry Potter fanvids mentioned above, they used “CC0 1.0” meaning that it is public domain - which it may or may not be (probably not), but by default datasets in Dataverse are given CC0 1.0 unless specified otherwise. The icicle images, on the other hand, uses “CC-BY 4.0”, which has more constraints on how the data can be used (see [Creative Commons attributions](https://creativecommons.org/licenses/by/4.0/)).

Statistics Canada datasets uploaded to the Odesi Dataverse use a custom template that go beyond Creative Commons rights to address the specific copyrights of Statistics Canada data. I’ll attach it to the appendices (Appendix A for English, Appendix B for French) of the present document. Point is, were we to upload metadata concerning sensitive datasets, we would not only need to strongly uphold Borealis’ Terms of Use, but also ensure that we do not put ourselves in any line of fire for any litigations (or adjacent actions) as this latter would prove problematic for both Borealis and the CRDCN. Provided that these terms are appropriately met, there should be no problem in preserving metadata on Borealis.

Pragmatics of Metadata Findability in Borealis

Assuming that everything aligns with Terms of Use and copyright laws, 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: do we want a file that is ingested, thus including variable level metadata in the DDI standard XML metadata file, or do we not want that?

It is of my belief that an ingested tabular formatted file would provide the best “findability” and preservation outcomes. To demonstrate this I have conducted a small trial in demo.borealisdata.ca (data here is prone to being deleted and is to test Borealis updates before it ends up in production).

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.

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)