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 has 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). Researchers from non-member universities can create a Borealis account, but may be restricted from accessing or uploading data. Interested members can consult their data librarian to inform them of their interest in Borealis.

Borealis holds datasets for participating member institutions and research organisations in their own institutionally curated repositories called “Dataverses”. In essence, Borealis is the service that hosts and allows these Dataverses to be searched. This 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 sensitive nature. For reasons relating to the nature of certain datasets (such as potentially identifiable information), some researchers may omit altogether any attempt at making their data findable in digital repositories.

There are, however, approaches by which researchers could make their data findable while preserving the confidentiality of their restricted data; in other words, make their research metadata findable, but make the actual research data inaccessible without explicit authorisation – therefore opening a safe and responsible gateway towards ethical data reuse. The present document provides an introductory guide in ethical metadata findability, curatorship, and custodianship.

Findable Data Without Access

Some datasets cannot be fully deposited into 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 promises made during the consent process used to collect the dataset at hand.

Although Borealis does not permit the upload any non-anonymized 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 other related metrics would be permissible. That is, information ***about the study*** 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 the goal of making restricted data findable with anonymized metadata, Borealis could be a central tool.

Borealis Open Science and Transparency: Pros and Cons

Over the years, there has been a push by the Tri-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, though it follows from the above statements that 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 an institutional Dataverse via Borealis as a repository for sensitive data metadata.

1. Making information about the study dataset available can initiate conversations between researchers for the reuse of data and for potential joint ventures.
2. 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 metadata which facilitates citation and subsequent findability.
3. Upon publication in Borealis, a machine readable metadata file for the dataset is automatically created. This metadata file ensures that variables within dataset files can be found with the Borealis search engine (therefore considerably increasing findability and potential reuse).

Of course, there may be certain qualms about using a national repository such as Borealis:

1. Being a national service, Borealis data is often scraped by academic colleagues in other countries – which may not always be a good thing.
2. Given the sensitivity of some data, we may not always want it to be “findable”.

It is important to note that, in the present case, 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 is a secure 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 by then using the scroll down menu to select the desired institution . From there, users can login with their institution credentials. Alternatively, users can navigate to the [members and partners](https://borealisdata.ca/#partners) web page, and selecting their institution. An in-depth guide on logging in 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, see their institutional Dataverse repository. There are various paths through which users can go about depositing their data and/or metadata. Depending on their institutional policies, should be able to create their own Dataverse within their institution’s Dataverse repository and begin the sharing process. An in-depth guide can be found in the [Collections](https://learn.scholarsportal.info/all-guides/borealis/collections/) Scholars Portal learn Borealis website.

What Can and “Cannot” Go in Borealis Dataverses

As previously mentioned, some datasets cannot be uploaded on any Borealis repositories (institutional or otherwise) if they violate the [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. 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.).
3. Do I have to required copyrights to publish the metadata? (who owns this data and metadata?)

While individual variables may seem too ambiguous to identify someone, they can always be used in concert with other variables or information to potentially identify study participants [flag to add the SD toolkit data de-identification material]. 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, but not fully deleted. This means that it cannot be removed without approval from the highest authority in the Borealis institutional hierarchy.

All this being taken into consideration, institutional Dataverses can take in any file type. Given the nature of the sensitive data and metadata being discussed, file formats such as a .csv, .xlsx, or .sav are optimal; this is discussed in greater details in the following section.

Pragmatics of Metadata Findability in Borealis

Assuming that the proposed deposit is within the Terms of Use, 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). This is precisely where we want to make an important decision: what level of metadata quality do we want to create and make findable?

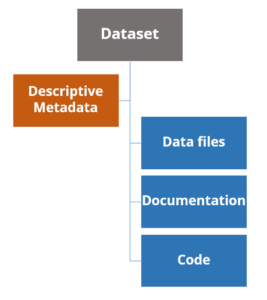
Here we will discuss three tiers, each building upon the previous, of metadata upload quality: good – better – best. With each new tier, the metadata becomes better contextualized, and more findable. The tier at which potential depositors decide to stop may largely be a function of the sensitivity level of their restricted metadata.

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

Step 1: Creating a “Dataset”

In Borealis, a dataset refers to the full contents of a deposit, not the data itself. In the context of sensitive data, you might think of this as an “entry” or a “record” where information about the dataset that isn’t being deposited is created. In Borealis lingo, a “dataset” is not only the raw data, but also the metadata that describes it, any code that goes alongside it, and any documentation that helps make sense of it. For restricted data, the “dataset” would refer to all the items in the below diagram, but *could* exclude any of the blue items as these might not be able to be deposited. For the remainder of this document, we will refer to a “Borealis dataset” as some combination of the components in figure 1 to distinguish them from a “dataset” (of raw or processed data) that fits the conventional understanding of the term.

Figure 1: The contents of a Borealis "dataset"



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

By default, deposits 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 Domain. That being said, depositors can input their own copyright guidelines to establish restrictions on how their deposit 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).

Depositors are not limited to Creative Commons attributions, and can create their own custom terms of use. Given the nature of the submitted files, data, and metadata it is advised that potential depositors consider how they would like to see the data be used, and establish solid grounds for restricting usage.

Step 3: Citation Metadata

Users must now create a title, description, select a subject category, and identify keywords for their Borealis dataset. This step is essential for base level findability. All of this metadata will subsequently be automatically formatted 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 deposit. Once this is completed users may Save their progress and proceed to the next step. ***File uploads aren’t part of tier 1. If you have files to deposit as part of your Borealis dataset, guidance is available in the higher tiers.***

Step 4: Field Specific Metadata

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

1. Geospatial Metadata (geographical region data that pertains to the deposit);
2. Social Sciences and Humanities Metadata (information pertaining to the sampling and data collection methods typical of social sciences);
3. Astronomy and Astrophysics Metadata (metadata pertaining to celestial objects and their observation/measurements);
4. Life Sciences Metadata (information relating to the biological substrate and data collection methods typical of life sciences);
5. Journal Metadata (metadata relating to related newspaper or academic journal publications);
6. Computational Workflow Metadata (information describing the type of work softwares used to make or document the data itself).

These field specific metadata are not mandatory, though they can significantly help in making the dataset more 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. 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**

At this point (end of tier 1), your Borealis dataset will have a persistent identifier (in this case a DOI), citation metadata (title, keywords, subject category), and field-specific metadata to facilitate findability.

**Is moving forward to tier 2 the right thing for you?:**

Users that wish to further contextualise their study data in a responsible manner should consider moving forward to tier 2. Other users, however, may deem that tier 1 is sufficiently contextualised for their purposes (for instance, divulsion of any more information may not be in their best interest or that of their participants). If depositors are unsure, they can come back later after publishing to expand the deposit.

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

Step 5: Supporting Documentation

This step does not necessarily make the Borealis 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 contents of the restricted dataset are. The submission of supporting documentation will provide a more in-depth description of what the dataset itself contains. This can be particularly useful if the collected data was not used or referred to in depth in a published research article. As such, tier 2 provides a human element that will help inform other researchers of the nature of the dataset.

In this case, a simple .txt, .docx, .pdf or other text format is sufficient to provide context for the dataset. This context file can be used to explain the goal or rationale of an original study, collected data, related publications, etc. It is important to note that, despite not being machine readable, these files should in no way divulge any information about participants. An in-depth upload guide can be found on the Borealis [Upload](https://learn.scholarsportal.info/all-guides/borealis/files/#Uploading-Files-to-a-Dataset) guide.

Tier 3: Tabular File Upload

Step 6: Thinking About Tabular File Preparation

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 7: Creating Tabular File

Figure 2 A decision tree diagram for depositing tabular information in a Borealis dataset

Borealis has a “[tabular ingest](https://learn.scholarsportal.info/all-guides/borealis/files/#Tabular-Ingest)” program that converts any files set up in a tabular format (.xlsx, .csv, .sav, .dta, etc.) into a standardised .tab file. This ingestion process has the distinct advantage of making file variables machine readable, and therefore findable by search engines. Submitting anonymized metadata files under a tabular format (such as the ones listed above) can allow other users to find the sensitive data’s study variables. While these variables will not contain any identifiable data, they can hold important information such as the study population, interviewer questions, literal questions, and/or additional notes that can be added with Borealis’ Data Explorer tool after the initial upload (although these will not be machine readable, they may still be useful to colleagues opening the files and investigating on their own).

The present example uses excel – though the same rules apply for other data software that can create tabular formatted files (Jasp, Google Sheets, Numbers, etc.). Regardless of the complexity of the file, there are a few rules that must be followed for proper results:

1. Each uploaded files should be made out of one sheet only (the tabular ingest function will only ingest the first sheet). If several sheets need to be uploaded, then they must be split into different files.
2. The metadata should be organised in a column format, with the first row being the variable name.

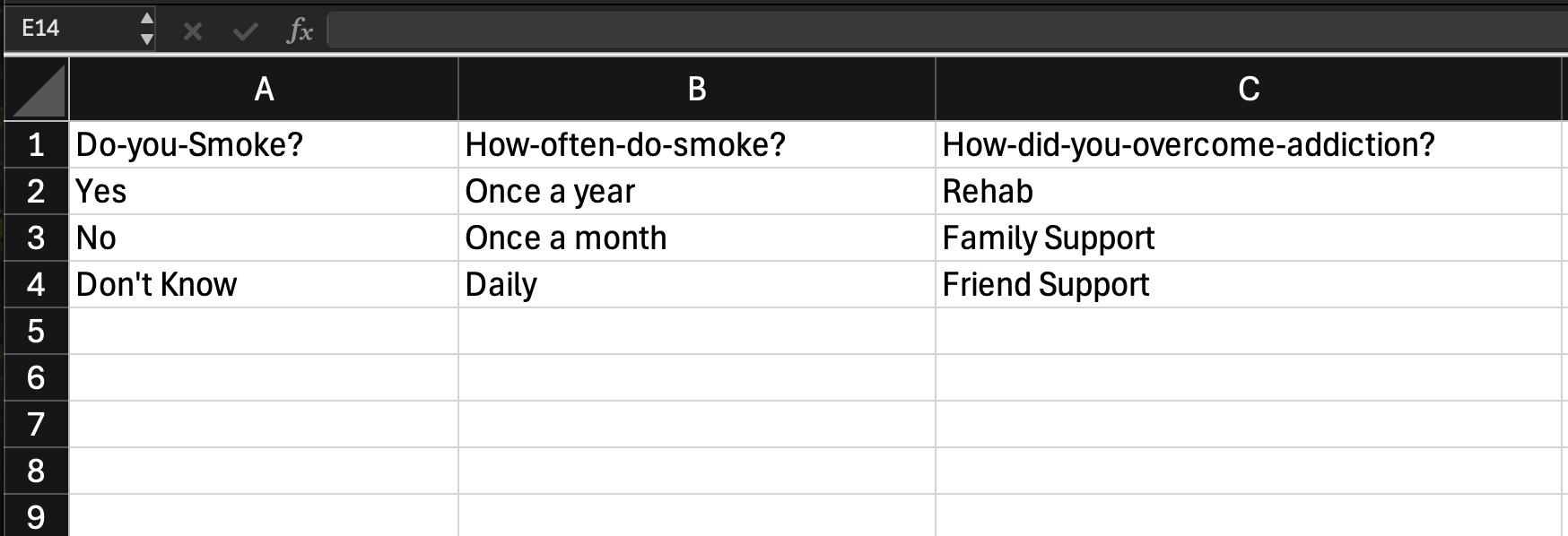
To facilitate machine readability post ingestion, users ***MUST*** replace spaces with either dashes/hyphens (-) or the HTML code for spaces ( &#32; ) in their variable names. During the ingestion process, all spaces and underscores ( \_ ) are removed, merging multiword variable names into a singular string of letters, which prevents any benefits in findability from uploading a file. Dashes/hyphens are strongly recommended as it is more user friendly (and human readable) than the HTML code for spaces. As seen in the above example, the file creation process is relatively simple. The variable names are in the first row, and the available response options (assuming it is a survey) can be listed below. Notice that there are no dashes/hyphens in the categorical options – this is because these categories, as previously mentioned, are not ingested when uploading an excel sheet, as such they do not need to be machine readable (though they can be in an .sav file).

Figure 3A sample tabular file for ingestion

Step 8: Uploading Metadata File and Curation.

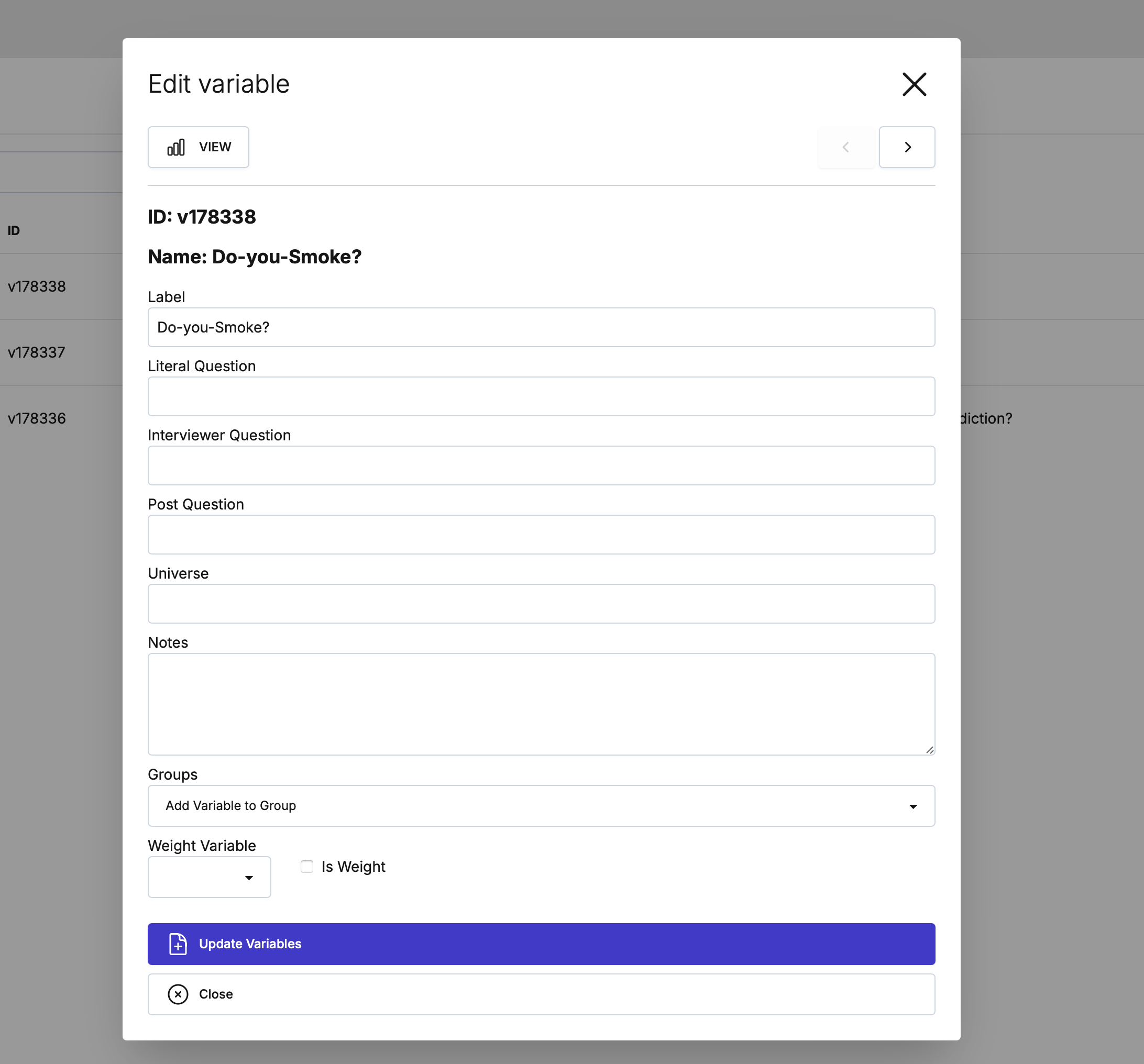
The upload process is quite similar to the one linked in Step 5. However, after uploading the file (and its ingestion), users can manually curate the metadata file using the “[Data Explorer](https://learn.scholarsportal.info/all-guides/odesi/working-with-data/)” tool integrated in Borealis. With this tool, users can add additional useful metadata to their ingested files. These are useful for other researchers seeking to better understand the variables, potential response types, study population/dataset universe, or relevant context related to the present metadata file (see image below). The [Data Explorer curation guide](https://learn.scholarsportal.info/all-guides/odesi/data-curation-in-data-explorer/) provides an in-depth look at how to add and modify variable level metadata with the tool.

Figure 4 Screenshot of the variable editor

Tier 4: Advanced Statistical Software Metadata Upload

Bonus Information

The chosen file format under which users submit their metadata may have advantages over other formats. For instance, more advanced file formats such as those used by SPSS (.sav) or R (.RData) can include additional metadata information that simpler file formats (.csv, .xlsx) cannot include. That is, Borealis DDI metadata outputs for .csv and .xlsx are limited to variable level metadata (such as the variable names in the first row), whereas more advanced .sav or .RData can include categorical information (such as the variable categories) in their DDI outputs. For example, if “Province” is a variable, then its categories would be “QC, ON, AB, BC, MB, NB, NL, SK, PE, and NS” – in this case .csv and .xlsx outputs will not include the individual provinces in their machine readable DDI outputs, but more advanced formats could.

It follows from this that there are very distinct advantages to using a statistical software such as SPSS to make the metadata even more findable than the approach discussed in Tier 3 above.

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)