ONDRI's Data Standards

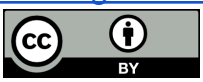
On behalf of the Neuroinformatics and Biostatistics (NIBS) platform

This document contains 7 (primary) sections. The data standards are a set of rules for the preparation of ONDRI data which harmonizes data for distribution and release. These standards help ensure compatibility across all platforms, cohorts, timepoints, and even projects. The standards ensure relative ease for combining, inspecting, and analyzing data. This ONDRI data standards document covers the following topics:

This document is subject to change as necessary. To cite this document:

The Ontario Neurodegenerative Disease Research Initiative's Neuroinformatics and Biostatistics team (2021). *ONDRI's Data Standards*. Retrieved from https://github.com/ondri-nibs/documentation

Version: 6.2.0.0



Data packages definitions

“Data package” is the general term used to define a set of required and recommended files for ONDRI data release. There are two primary data packages (1) tabular and (2) non-tabular. A third option exists and can only be used at the discretion and direction of NIBS (see **Other data packages**). We briefly describe and define the two primary data packages here and expand on all types in their respective sections.

1. *Tabular*: Data that can be contained within a single spreadsheet. Participants are listed down the rows and variables across the columns. Exists as two subtypes:

    a. *Wide tabular*: Each participant exists exactly zero (missing) or exactly one time down the rows. Examples: Demographics, MoCA. **Wide tabular is preferred and generally enforced by NIBS.**

    b. *Long tabular*: A participant can exist zero (missing), one, or many times down the rows (i.e., a repeated factor). Examples: Concomitant medications, medical history. **If you believe data should be long tabular, contact NIBS for consultation and approval.**

2. *Non-tabular:* data that cannot be contained within a single spreadsheet, that is, a file or multiple files per participant. Examples: participant accelerometer files collected by wearable technology, NIfTI neuroimaging files for resting state fMRI. Conceptually non-tabular have “wide” and “long” analogs. See **Non-tabular data packages** for details. Non-tabular data also have “wide” and “long” analogs to tabular data.

NOTE: The majority of foreseeable data delivered to NIBS for standards, outlier, and release processes will be, or should be formatted as wide and in most cases tabular (see **Common formatting and requirements** and **Tabular data packages**).

Common and general formatting, requirements, naming, and codes

Filename format and requirements

Generally, all filenames consist of the following format (some conditions and changes apply for *Non-tabular* and *Other* data packages; see those sections for details):

[ONDRI CODE]_[COHORT CODE]_[VISIT CODE]_[PLATFORM CODE]_[SUBPLATFORM CODE]_[DATA SET CODE]_[DATE OF RELEASE FOR CURATION]_[FILE TYPE].

Each released file name must be unique. Each partition of the above format is described below and separated by underscores ("_"). SUBPLATFORM CODE and DATA SET CODE are described last and in more detail. For more details, definitions, (current) valid codes, and examples see **Appendix A**.

- **ONDRI CODE:** ONDRI study to which these data belong.
- **COHORT CODE:** Cohort (disease) to which these data belong.
- **VISIT CODE:** Visit at which data were collected.
- **PLATFORM CODE:** Assessment platform to which the data belong.
- **SUBPLATFORM CODE:** Some platforms have multiple data files. For examples:
    - Sensor data (SNSR) has multiple sources of sensor data, as organized by device name: GNAC, BITF, NONW
    - Genomics (GNMC) has two sources of genetic data: NeuroX and ONDRISeq.
    - Neuroimaging (NIMG) has multiple data modalities and types, e.g., resting state fMRI, DTI, SABRE-LE estimates.
- **DATA SET CODE**: Some platforms have multiple data files. For examples:
    - Clinical (CLIN) will have multiple partitions that generally reflect certain types of data (e.g., demographics, MOCA, questionnaires, disease specific). These will be denoted by brief additional descriptors e.g., Neuropsychology (NPSY) and Neuroimaging's SABRE-LE (NIMG_SABR) each have two data sets: a "full" with comprehensive and "item level" data and a "minimum" which generally reflect summaries derived from the full data.
- **DATE OF RELEASE FOR CURATION:** The date a curator prepares, preprocess, and/or provides data with the intent that it goes through standards checks and outlier analyses.
- **FILE TYPE:** Labeled as DATA, DICT, README, METHODS, GLOSSARY or SUP_[ADDITIONAL INFORMATION].

NOTE: If a platform plans multiple partitions or versions of the data for release (e.g., full and minimum) please contact NIBS so that we can plan accordingly.

General requirements

- All text-based files (.csv and .txt) must be UTF-8 (preferred) or ASCII (acceptable) encoded. Other encoding types will be rejected. **NOTE:** Most simple text editors (e.g., NotePad++, gedit, TextEdit) allow users to export or “Save As...” with different encoding types.

- All document files (METHODS, GLOSSARY, occasionally SUP*) must be in PDF format. Other formats will be rejected.

- Do not include multiple versions of participant identifiers (e.g., original ID, platform-specific ID, and/or transfer IDs).

- Platforms can only include data collected by their own platform. For example: do not include age, sex, or cohort in data files, as these can be found in the Clinical (CLIN) data.

- Duplicate data should not exist within or across data packages. For examples: do not include weight in pounds and weight in kilograms. This does not apply to conversions of variables (e.g., raw to normalized). **NOTE:** in some cases there exists similar or related variables independently collected by multiple platforms. For example both clinical (CLIN) and neuropsychology (NPSY) have education variables that measure education in two different ways. Please contact NIBS if you have questions on possibly duplicate variables.

- Precision levels (i.e., the number of decimal places) are at the discretion of the platform but should be consistent across all data within a measure.

- All text and names are strictly limited to alphanumerics (A-Z, a-z, 0-9) and underscores (_).

- All DATES are required in YYYYMMMDD format where, e.g., January 15th, 2019 is 2019JAN15.

- If using time, format time as HH:MM:SS on a 24 hour clock: 00:00:00 is midnight (12:00 am) and 23:59:00 is 11:59 pm. The DICT must indicate which timezone this occurred in, or which timezone it uses as a reference (e.g., UTC). Be sure to check for daylight vs. standard time. The data type, for now, is coded as TEXT. ***NOTE:*** Please contact NIBS if you need to include time.

Methods document requirements.

The METHODS document provides summarized and detailed information on how data were curated. METHODS documents are a vital resource for data consumers because in most data, data consumers need to know how the data came about. The METHODS document has a standardized template. For access, contact NIBS. Below we provide an overview of the template.

Title Page/Page 1: ONDRI branded header with text to allow for: (1) the name of the document (which matches the package it is in), and (2) indication of authorship on behalf of a platform. This page also includes a list of the 4 standardized sections: Methods summary, Methods details, References, and Contributions & Contacts. There is blank space on this page that can be used for: (1) details on how to cite the document, (2) application of a license to the document, (3) versioning and/or dating of the document, or (4) other similar information to the aforementioned. No other information should appear on this page.

Page 2: *Methods Summary*. The *Methods Summary* is meant to be a short version of methods with references. Effectively, this should be written as in a way to be copy & pasted directly into publications and other materials. Generally this should be no more than a paragraph.

Page 3 (or more): *Methods Details*. The *Methods Details* is meant for a platform to expand, in as much detail as necessary or as much detail as they'd like, on how data for this particular package was curated. This can be broken down into multiple sections. The template has a style for section delimiters.

Page immediately following *Methods Details*: *References*. The *References* section is like any other reference section. We do not enforce any particular citation style (e.g., named vs. numbered). Rather, the platform should use the style *most commonly used* in their fields. The minimal requirement is that it is, at least, in an established citation format.

Final page: *Contributions and contacts*. The *Contributions and contacts* section serves two purposes: (1) contact information for the data package/platform, and (2) provide a listing of people who contributed to the METHODS document and how they contributed *to the document*. There is some leeway in contributions. Contribution types have been adapted from the CRediT system (https://casrai.org/credit/). We allow for four categories: (1) ***Conceptualization***: Ideas; formulation or evolution of overarching goals and aims. (2) ***Writing - original draft***: Preparation, creation and/or presentation of the published work, specifically writing the initial draft. (3) ***Writing - review & editing***: Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision, and (4) ***Additional contributions***: Contributions to this document not otherwise captured in Conceptualization, Writing (draft), or Writing (reviews/edits); additional contributions is akin to "acknowledgements".

Individuals are to be listed in boxes in the template. Individuals can be repeated across the contribution categories, though, if someone is listed in any of the first three categories, they should not be listed in the *Additional contributions* (and vice versa).

Tabular data packages

Tabular package example. Items in **blue bold** are folders and items in black are files.

Montreal Cognitive Assessment Example	Corresponding files & descriptions
OND01_ADMCI_01_CLIN_MOCA_2018AUG29_DATAPKG ├── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_README.csv ├── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_DICT.csv ├── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_DATA.csv ├── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_METHODS.pdf ├── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_MISSING.csv └── **OND01_ADMCI_01_CLIN_MOCA_2018AUG29**_SUP.txt	**OND01_*** ├── a README listing all files here ├── DICTIONARY for DATA ├── DATA ├── METHODS documentation ├── MISSING from this DATA.csv └── A SUP file with qualitative information

Tabular data packages are exactly one level: all necessary and required files exist within a single folder.

Content	***Requirements***	***Format, type (encoding)***	***Description***
*_README.csv	REQUIRED	Tabular, text (UTF-8)	Describes and lists the contents of the data package
*_METHODS.pdf	REQUIRED (Exemptions require approval).	Document, varies (PDF)	Describes how data were preprocessed and prepared for release; an extended version of a “Methods” section in publication.
*_DICT.csv	REQUIRED	Tabular, text (UTF-8)	Mapping and descriptions of columns in DATA
*_DATA.csv	REQUIRED	Tabular, text (UTF-8)	Data for distribution
*_MISSING.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Missingness descriptions and reasons when participants are ***entirely*** missing from DATA
_SUP.*	OPTIONAL	Any, optional (dependent)	Generally additional qualitative data such as notes
*_GLOSSARY.pdf	OPTIONAL	Document, varies (PDF)	Short descriptions of terms and background

NOTE: SUP (supplemental) is a placeholder name for additional information, notes, or generally qualitative information not otherwise distributed (e.g., clinical, administration, or instrumental notes; collection parameters). **Contact NIBS on how to prepare, format, and distribute SUP files.** Tabular data exists in two possible formats (also see the next page):

1. *Wide tabular*: Each participant exists exactly zero (missing) or exactly one time down the rows. Examples: Demographics, MoCA.

2. *Long tabular*: A participant can exist zero (missing), one, or many times down the rows and also requires at least one variable (i.e., a repeated factor) to denote occurence of participants in long formats. Example: Concomitant medications. ***NOTE:*** Use of long tabular is at the discretion of and subject to approval by NIBS.

	A	B	C	D	E	F	G	H	I	J
1	SUBJECT	VISIT	NPSY_SITE	NPSY_DATE	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6
2	OND01_BYC_0001	01	BYC	2017OCT26	YES	1	-23.6924	YES	10	0
3	OND01_BYC_0002	01	BYC	2017OCT26	NO	1	-84.5469	NO	11	1
4	OND01_BYC_0003	01	BYC	2017OCT26	NO	2	55.517	YES	20	2
5	OND01_SBH_0001	01	BYC	2017OCT26	NO	4	11.0302	YES	2	10
6	OND01_TWH_0002	01	TWH	2017APR09	YES	3	44.3537	NO	5	4

Illustration of wide tabular. Visualized with Comma Chameleon.

	A	B	C	D	E	F
1	SUBJECT	VISIT	CONMED_SITE	CONMED_DATE	VAR1	VAR2
2	OND01_BYC_0001	01	BYC	2017OCT26	MED1	Aricept
3	OND01_BYC_0001	01	BYC	2017OCT26	MED2	Vitamin B
4	OND01_BYC_0001	01	BYC	2017OCT26	MED3	Aspirin
5	OND01_BYC_0003	01	BYC	2017OCT26	MED1	Levodopa
6	OND01_SBH_0001	01	BYC	2017OCT26	MED1	Levodopa
7	OND01_SBH_0001	01	BYC	2017OCT26	MED2	Lactulose
8	OND01_TWH_0002	01	TWH	2017APR09	MED1	Aspirin
9	OND01_TWH_0002	01	TWH	2017APR09	MED2	Lyrica
10	OND01_TWH_0002	01	TWH	2017APR09	MED3	Citalopram
11	OND01_TWH_0002	01	TWH	2017APR09	MED4	Aricept
12	OND01_TWH_0002	01	TWH	2017APR09	MED5	Clonazepam

Illustration of long tabular. Note that each participant varies in their repetitions (see “VAR1”), however in some cases (e.g., repeated factor designs) each participant may have an equal number of repeats. Visualized with Comma Chameleon.

Common tabular package file requirements

Each file, their requirements, and examples follow in subsequent sections, and a concrete example can be found here: https://github.com/ondri-nibs/toy_data.

README requirements. Every data package must have a README file. README files list all subsequent files in a data package. The README file is meant to be the first thing read by data consumers in order to orient them as to which files contain what information. The README:

- must be released as a comma separated value (csv) file. README format details are below with an example in **Appendix B**.
- contains exactly two columns in the following order:
    - **FILE**: All files contained in the data package.
    - **DESCRIPTION**: Brief description (no more than 200 characters) of the files listed.
- contains at least three rows with the names (FILE) and descriptions (DESCRIPTION) of
    - METHODS
    - DICT
    - DATA

  Subsequent rows are included for MISSING, SUP*, and GLOSSARY as needed.

Data Dictionary Requirements. Every DATA file must have a corresponding dictionary (DICT) file. All dictionary files must briefly and adequately explain all columns in their respective data file.

- The data dictionary file must be released as a comma separated value (csv) file. Dictionary format details are found in **Appendix B**.

- Each dictionary file must contain exactly four columns in the following order:
    a. **COLUMN_LABEL:** All columns contained in the data file, in the exact order as they appear in the DATA file.

    b. **DESCRIPTION:** A brief description (no more than 200 characters) of the items listed in COLUMN_LABEL.

    c. **TYPE:** The type of data for each item listed in COLUMN_LABEL (e.g. CATEGORICAL, TEXT). Options for this field are provided below and in **Appendix B**.

    d. **VALUES:** A brief description (no more than 200 characters) of the levels, possible ranges, and/or codes that correspond to the data described in COLUMN_LABEL. For examples “non-negative integers”, “0 - 15 in integers”, “0 - 200 cm”, “Y = Yes or N = No”.

Missing File Requirements. All data packages must account for every enrolled participant *at the time of the respective visit*. Participants without at least one data point should not appear in the DATA file but in a corresponding MISSING file. MISSING files must briefly explain why all data are missing for necessary participants.

- The MISSING file must be released as a comma separated value (csv) file. MISSING format details are below with examples in **Appendix B**.

- Each MISSING file must contain six columns in the following order:

    a. **SUBJECT** and **VISIT**, each with the same requirements as the DATA file.

    b. **Columns describing the site(s) and date(s) of acquisition**. The SITE and DATE columns must have the same column names as are used in the DATA file. If data were collected but are not usable, the site and date of acquisition should be reported. If data were never collected, the cells should contain valid missing codes.

    c. **MISSING_CODE**: a valid missing code that accurately describes the reason all data are missing. See **Appendix A**.

    d. **DESCRIPTION**: a brief description (in words; no more than 200 characters) of the reason all data are missing.

Supplemental Material Requirements. *Optional.* In some cases, a platform may have additional or supplemental information that would be useful or essential for analyses. For examples: (1) additional information beyond the missing data code (e.g. more details regarding the artifacts on an image), or (2) guides and instructions on handling data. When SUP files are text, it is recommended that supplementary files are released as a csv file, with column names consistent with those in the data file (i.e. SUBJECT and VISIT). If new column names are necessary, they should follow the column name requirements of the data file. If additional information is more of a narrative, then it is recommended to distribute a .pdf using the METHODS template. Other SUP files are possible if they are necessary for the package. **Contact NIBS if SUP files are to be included in release.**

Glossary Requirements. It is recommended that all platforms include data *glossaries*. Glossaries are generally expanded versions of dictionaries that include more detailed descriptions of individual variables, or entire instruments that also include references. **Contact NIBS if GLOSSARY files are to be included in release.**

DATA and DICT file structures

Below describes the requirements for the DATA file format. Many of these requirements also apply to the corresponding DICT file. Certain items, such as SUBJECT, VISIT, SITE(s), and DATE(s), are required to be in a specific order.

- Data files should include only enrolled participants with at least one data point. Any enrolled participants without data should not be included in the data file, but in a corresponding missing file (see **Missing File Requirements** below). Screen failed participants should never appear in data files. See VALID IDS COHORT.csv file in the NIBS Sandbox folder > Cross-Platform Information > IDS for Distribution for information on which participants to include at which time points.

- **The first column of every DATA file must be SUBJECT. Likewise, the first row of every DICT file must be SUBJECT.** SUBJECT denotes participants with their original (i.e., enrollment) IDs. The use of enrollment IDs ensures that the ID remains static throughout the study; which enables users to link participants' longitudinal data. See also **Curation**.

- **The second column of every DATA file must be VISIT. Likewise, the second row of every DICT file must be VISIT.** VISIT denotes the visit number with the standardized visit code. Visit codes can be found in **Appendix A**. The use of standardized visit codes ensures a common element to identify participants for cross-sectional analyses within a time point.

- **All DATA files require *at least* one column to indicate the SITE of data acquisition. Likewise, SITE rows of every DICT file correspond to the SITE columns of DATA.** Site columns must appear after VISIT and before DATE columns. Site ***must*** be the actual site of data acquisition (not the intended site of collection and not the recruitment site). For example, if a participant was scheduled for an assessment at Sunnybrook (SBH) but for some reason data had to be collected at Baycrest (BYC), the site of acquisition is Baycrest (BYC). Site codes can be found in **Appendix A**. The SITE variable must be formatted as follows:

  [REQUIRED PLATFORM CODE]_[OPTIONAL SUBPLATFORM CODE]_SITE_[OPTIONAL SUPPLEMENTAL CODE].

- **All DATA files require *at least* one column to indicate date of data acquisition. Likewise, DATE rows of every DICT file correspond to the DATE columns of DATA.** Date columns must appear after the SITE columns. Date ***must*** be the actual date of data acquisition (not the date that the data were entered into the database or processed). DATE required in YYYYMMDD format. The DATE variable must be formatted as follows:

  [REQUIRED PLATFORM CODE]_[OPTIONAL SUBPLATFORM CODE]_DATE_[OPTIONAL SUPPLEMENTAL CODE].

- **All subsequent DATA column labels are strictly limited to containing alphanumerics (ABC and 123) and underscores (_). Likewise, the order of DATA column labels correspond to the rows of the DICT file.** It is recommended that all subsequent column labels are fully capitalized. Column names/labels (e.g., variable names) should be short, meaningful, and human + machine readable.

- **No duplication of column names** (even with mixed case). For example: VARIABLE_1, Variable_1, and variable_1 are considered equivalent and are therefore duplicate column names.

NOTE: It is at the discretion of each platform to determine which of the following site and date formats are most useful to include in their data files. For example, if a platform collects data across multiple dates and these data exist in a single distributed data file, but the platform deems only one date to be necessary and sufficient for accurate data analyses, then only one date is required in that file.

There are five possible scenarios for indicating the site(s) and date(s) of acquisition. We describe formatting requirements for only the two most common scenarios below. Please contact NIBS if your data fall into the uncommon scenarios. See **Appendix A** for site codes.

SCENARIO 1: All data per participant were acquired at only one site and on only one date.

SOLUTION 1: The third and fourth columns of the data file must adhere to site and date format (see criteria 5 and 6 on pg 3-4). For examples: NPSY_SITE, NIMG_SITE, EYTK_SITE, and NPSY_DATE , NIMG_SABR_DATE, EYTK_DATE for site and date respectively.

SCENARIO 2: All data per participant were acquired at only one site but across multiple dates.

SOLUTION 2: The third column of the data file must adhere to site format (see criterion 5 on pg 3-4). For examples: NPSY_SITE, NIMG_SITE, EYTK_SITE. Site must be followed by the date columns. Because multiple date columns will be required the date columns must include information differentiating between the dates reported. Date columns must adhere to date format (see criterion 6 on pg 4). For examples:

- CLIN_DEMOG_DATE_PT and CLIN_DEMOG_DATE_SP would mean “Clinical platform, demographics, date of acquisition, participant data” and “Clinical platform, demographics, date of acquisition, study partner data”. respectively.
- EYTK_DATE_PRO and EYTK_DATE_ANTI would indicate the pro and anti-saccade tasks were performed on separate dates and would mean “Eyetracking, date of acquisition, pro-saccade task”, and “Eyetracking, date of acquisition, anti-saccade task”, respectively.

SCENARIO 3: Participant data was acquired at multiple sites but only on one date.
SOLUTION 3: Contact NIBS.

SCENARIO 4: Participant data was acquired at multiple sites across multiple dates.
SOLUTION 4: Contact NIBS.

SCENARIO 5: Participant data was acquired in a non-consistent order of dates, for example a subset of participants (say, group A) received some assessments on one date, with follow up dates for remaining data. But another subset (say, Group B) received assessments in the opposite order as Group A.
SOLUTION 5: Contact NIBS.

General requirements for text-based files and tabular data

- Commas in text-based data should be avoided. Short text does not require commas and lists can be separated with semicolons (;).

- Consistent encapsulation of cells, i.e., all cells must be encapsulated with double quotes (“”) or use no encapsulation. If quotes are necessary within a cell (e.g., as part of a description or response), then use single quotes (‘’) though it is best to avoid quotations within data.

- No leading or trailing whitespace (e.g., single spaces, tabs, returns) within cells.

- For DATA files: derived variables (e.g., means, SDs, proportions) from other variables within the DATA file should *not* be included in the same spreadsheet as the original or source data (Broman & Woo, 2018). In general derived values should not be distributed. If you have questions contact NIBS.

- For DATA files: blank cells should not exist with one exception: a column exists conditionally on another column and data do not exist (e.g., follow-up questions such as the number of cigarettes smoked per day, following the question “Do you smoke?”).

- For DATA files: contents of one cell should map to one value.

- For DATA files: only include enrolled participants *at the time of the respective visit*. A list of ONDRI included participants (by IDs x visit) is distributed by NIBS.

Non-tabular data packages

Non-tabular data packages contain coherent/consistent data from the same process. Non-tabular packages are exactly three levels: all necessary and required files exist within a hierarchy of three folders. First, an overview with a visualization of the levels are provided, followed by additional details, requirements, and limitations for each level.

- **Level 1 (top level)**: Contents of this level are contained within a folder and similar to the contents of *Tabular* data packages and also include uniquely named folders ("Level 2") that contain additional data. There exists exactly one Level 1 item.

- **Level 2 (within Level 1)**: Contents of this level are contained within a folder and similar to the contents of *Tabular* data packages and also include exactly one folder called *DATAFILES*. There exists at least one Level 2 item.

- **Level 3 (within Level 2)**: Contents strictly limited to: (i) a file or multiple files per participant that are the data of interest for this data package, (ii) a FILELIST.csv file, (iii) possibly a MISSING.csv file, or (iv) possibly a common DICT.csv file. There exists exactly one Level 3 ***per*** Level 2.

Non-tabular package "wide" example. Items in **`blue bold`** are folders and items in `black` are files.

Resting state fMRI example	Corresponding levels & descriptions
OND01_VCI_01_NIMG_RSFMRI_2019JAN04_DATAPKG ├── [...]_README.csv ├── [...]_METHODS.pdf ├── **VCIEPI** │ ├── [...]_VCIEPI_[...]_README.csv │ ├── [...]_VCIEPI_[...]_MISSING.csv │ ├── [...]_VCIEPI_[...]_DICT.csv │ ├── [...]_VCIEPI_[...]_DATA.csv │ ├── [...]_VCI_EPI_template_mask.nii.gz │ ├── [...]_VCI_EPI_template.nii.gz │ └── **DATAFILES** │ ├── [...]_FILELIST.csv │ ├── [...]_MISSING.csv │ ├── OND01_BYC_1006_RSFMRI_VCIEPI.nii.gz │ ├── OND01_CAM_1020_RSFMRI_VCIEPI.nii.gz │ ├── [...] │ ├── OND01_TWH_5025_RSFMRI_VCIEPI.nii.gz │ └── OND01_TWH_5026_RSFMRI_VCIEPI.nii.gz └── **MNI** ├── [...]_MNI_[...]_README.csv ├── [...]_MNI_[...]_MISSING.csv ├── [...]_MNI_[...]_DICT.csv ├── [...]_MNI_[...]_DATA.csv ├── [...]_VCI_MNI_template_mask.nii.gz ├── [...]_VCI_MNI_template.nii.gz └── **DATAFILES** ├── [...]_FILELIST.csv ├── [...]_MISSING.csv ├── OND01_BYC_1006_RSFMRI_VCIMNI.nii.gz ├── OND01_CAM_1020_RSFMRI_VCIMNI.nii.gz ├── [...] ├── OND01_TWH_5025_RSFMRI_VCIMNI.nii.gz └── OND01_TWH_5026_RSFMRI_VCIMNI.nii.gz	**LEVEL 1 (TOP LEVEL PACKAGE NAME)** ├── README ├── A METHODS document ├── **LEVEL 2.1 (FIRST LEVEL 2)** │ ├── README │ ├── MISSING from this DATA.csv │ ├── DICTIONARY │ ├── DATA (applicable at level 2.1) │ ├── A required ADDITIONAL file │ ├── A required ADDITIONAL file │ └── **LEVEL 3 (DATAFILES)** │ ├── DATAFILES listing │ ├── MISSING files at this level │ ├── Participant file │ ├── Participant file │ ├── [...] │ ├── Participant file │ └── Participant file └── **LEVEL 2.2 (SECOND LEVEL 2)** ├── README ├── MISSING from this DATA.csv ├── DICTIONARY ├── DATA (applicable at level 2.2) ├── A required ADDITIONAL file ├── A required ADDITIONAL file └── **LEVEL 3 (DATAFILES)** ├── DATAFILES listing ├── MISSING files at this level ├── Participant file ├── Participant file ├── [...] ├── Participant file └── Participant file

Non-tabular packages can range in complexity, much more so than tabular packages. The above example shows one of moderate complexity: there are two Level 2s and there are additional (necessary) files within each Level 2, and there different MISSING files (reflected in Level 3), with MISSING participants in each Level 2. The illustration on the follow page shows a minimum and maximum example of non-tabular packages. ***This package features one file per participant within the DATAFILES folder and within the rows of the DATA.csv files, which is of the "wide" type because a participant has exactly zero (missing) or one file/row.***

Non-tabular package "long" example. Items in **`blue bold`** are folders and items in `black` are files.

GENEAActiv (sensor) example	Corresponding levels & descriptions
OND06_VCI_01_SNSR_GNAC_2020APR14_DATAPKG ├── [...]_README.csv ├── [...]_METHODS.pdf ├── **ACCELEROMETER** │ ├── [...]_ACCELEROMETER_[...]_README.csv │ ├── [...]_ACCELEROMETER_[...]_DICT.csv │ ├── [...]_ACCELEROMETER_[...]_DATA.csv │ ├── [...]_ACCELEROMETER_[...]_MISSING.csv │ └── **DATAFILES** │ ├── [...]_FILELIST.csv │ ├── [...]_MISSING.csv │ ├── OND06_SBH_0867_[...].edf │ ├── OND06_SBH_0867_[...].edf │ ├── [...] │ ├── OND06_SBH_5309_[...].edf │ └── OND06_SBH_5309_[...].edf └── **TEMPERATURE** ├── [...]_TEMPERATURE_[...]_README.csv ├── [...]_TEMPERATURE_[...]_DICT.csv ├── [...]_TEMPERATURE_[...]_DATA.csv ├── [...]_TEMPERATURE_[...]_MISSING.csv └── **DATAFILES** ├── [...]_FILELIST.csv ├── [...]_MISSING.csv ├── OND06_SBH_0867_[...].edf ├── OND06_SBH_0867_[...].edf ├── [...] ├── OND06_SBH_5309_[...].edf └── OND06_SBH_5309_[...].edf	**LEVEL 1 (TOP LEVEL PACKAGE NAME)** ├── README with METHODS.PDF ├── A METHODS document ├── **LEVEL 2.1 (FIRST LEVEL 2)** │ ├── README with MISSING, DICT, DATA │ ├── DICTIONARY │ ├── DATA (applicable within this level 2.1) │ ├── MISSING participants from this DATA.csv │ └── **LEVEL 3 (DATAFILES)** │ ├── DATAFILES listing │ ├── MISSING participant files, this level │ ├── Participant file │ ├── Participant file │ ├── [...] │ ├── Participant file │ └── Participant file └── **LEVEL 2.2 (SECOND LEVEL 2)** ├── README with MISSING, DICT, DATA ├── DICTIONARY ├── DATA (applicable within this level 2.2) ├── MISSING participants from this DATA.csv └── **LEVEL 3 (DATAFILES)** ├── DATAFILES listing ├── MISSING participant files, this level ├── Participant file ├── Participant file ├── [...] ├── Participant file └── Participant file

Non-tabular packages can range in complexity, much more so than tabular packages. The above example shows one of moderate complexity: there are two Level 2s and there are additional (necessary) files within each Level 2, and there different MISSING files (reflected in Level 3), with MISSING participants in each Level 2. The illustration on the follow page shows a minimum and maximum example of non-tabular packages. ***This package also features multiple files per participant within the DATAFILES folder and within the rows of the DATA.csv files, which is of the "long" type because a participant can have zero (missing), one, or many files/rows.***

Non-tabular minimal and maximal examples

The following example is a schematic with placeholder names, meant to illustrate the minimal and maximal examples of non-tabular packages. Key: [...] required filename format, SITE is a placeholder for OND01 sites, #### is a placeholder for the 4 digits of a OND01 participant ID, *.* are meant to capture the tail of the file name and the extension (e.g., RSFMRI.nii.gz).

MINIMAL

```
OND01_*
├── [...]_README.csv
├── [...]_METHODS.pdf
└── FOLDER_1
    ├── [...]_README.csv
    └── DATAFILES
        ├── [...]_FILELIST.csv
        ├── OND01_SITE_####_*.*
        ├── OND01_SITE_####_*.*
        ├── [...]
        ├── OND01_SITE_####_*.*
        └── OND01_SITE_####_*.*
```

MAXIMAL

```
OND01_*
├── [...]_README.csv
├── [...]_METHODS.pdf
├── [...]_DICT.csv
├── [...]_DATA.csv
├── [...]_MISSING.csv
├── [...]_SUP.txt
├── [...]_ADDITIONAL_1*.*
├── [...]_ADDITIONAL_[...]*.*
├── [...]_ADDITIONAL_N*.*
├── FOLDER_1
│   ├── [...]_README.csv
│   ├── [...]_METHODS.pdf
│   ├── [...]_DICT.csv
│   ├── [...]_DATA.csv
│   ├── [...]_MISSING.csv
│   ├── [...]_SUP.txt
│   ├── [...]_ADDITIONAL_1*.*
│   ├── [...]_ADDITIONAL_[...]*.*
│   ├── [...]_ADDITIONAL_N*.*
│   └── DATAFILES
│       ├── [...]_FILELIST.csv
│       ├── [...]_MISSING.csv
│       ├── [...]_DICT.csv
│       ├── OND01_SITE_####_*.*
│       ├── OND01_SITE_####_*.*
│       ├── [...]
│       ├── OND01_SITE_####_*.*
│       └── OND01_SITE_####_*.*
├── FOLDER_[...]
└── FOLDER_N
    ├── [...]_README.csv
    ├── [...]_METHODS.pdf
    ├── [...]_DICT.csv
    ├── [...]_DATA.csv
    ├── [...]_MISSING.csv
    ├── [...]_SUP.txt
    ├── [...]_ADDITIONAL_1*.*
    ├── [...]_ADDITIONAL_[...]*.*
    ├── [...]_ADDITIONAL_N*.*
    └── DATAFILES
        ├── [...]_FILELIST.csv
        ├── [...]_MISSING.csv
        ├── [...]_DICT.csv
        ├── OND01_SITE_####_*.*
        ├── OND01_SITE_####_*.*
        ├── [...]
        ├── OND01_SITE_####_*.*
        └── OND01_SITE_####_*.*
```

Non-tabular Level 1 required, allowable, and optional content

Content	*Requirements*	*Format, type (encoding)*	*Description*
LEVEL 2 folders	REQUIRED	Folder(s)	Folder or folders with unique names.
README.csv	REQUIRED	Tabular, text (UTF-8)	Describes and lists the contents of the data package
METHODS.pdf	REQUIRED (exemptions require approval).	Document, varies (PDF)	Describes how data were preprocessed and prepared for release; an extended version of a "Methods" section in publication. ***Must pertain to all contents in Levels 2 and 3.***
DICT.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Mapping and descriptions of columns in DATA.csv
DATA.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Data for distribution. ***Requirement:*** These data must be dependent on DATAFILES in all subsequent levels (or vice versa)
MISSING.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Missingness descriptions and reasons when participants are ***entirely*** missing from the package
Additional files	CONDITIONAL REQUIRED / OPTIONAL	Any, optional (dependent)	Any additional files that pertain to all DATAFILES in all subsequent levels
SUP*.*	OPTIONAL	Any, optional (dependent)	Generally qualitative data such as notes

Non-tabular Level 2 required, allowable, and optional content

Content	*Requirements*	*Format, type (encoding)*	*Description*
DATAFILES	REQUIRED	Folder	Folder that contains a data file or multiple data files per participant
README.csv	REQUIRED	Tabular, text (UTF-8)	Describes and lists the contents of the data package
METHODS.pdf	CONDITIONAL REQUIRED	Document, varies (PDF)	Describes how data were preprocessed and prepared for release; an extended version of a "Methods" section in publication.
DICT.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Mapping and descriptions of columns in DATA
DATA.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Data for distribution
MISSING.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Missingness descriptions and reasons when participants are ***entirely*** missing from its respective level 2
Additional files	CONDITIONAL REQUIRED / OPTIONAL	Any, optional (dependent)	Any additional files that pertain to all DATAFILES *at this particular level*
SUP*.*	OPTIONAL	Any, optional (dependent)	Generally qualitative data such as notes

Non-tabular Level 3 required, allowable, and optional content

Content	***Requirements***	***Format, type (encoding)***	***Description***
FILELIST.csv	REQUIRED	Tabular, text (UTF-8)	List of participant files that also includes SUBJECT, VISIT, SITE, and DATE (akin to DATA).
Participant files	REQUIRED	Any, required (dependent)	A data file or multiple data files per participant. ***Required:*** Filenames must at least be the ONDRI participant ID; additional portions of the file names may be necessary but should be as short as possible.
MISSING.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Missingness descriptions and reasons when participant files are ***entirely*** missing from **DATAFILES**
DICT.csv	CONDITIONAL REQUIRED	Tabular, text (UTF-8)	Only included when all the participant files meet the following two conditions: (1) are text-based and (2) have identical columns. Mapping and descriptions of columns in all *Participant files*.

Non-tabular package requirements

NOTE: Non-tabular packages contain many files and are likely of modest (100s of megabytes), large (1-10 gigabytes), and very large (10+ gigabytes) sizes and can be difficult to manage, check, and transfer. ***Before any preparation of these packages contact NIBS.***

The non-tabular package can range from compact with only one datafile per participant as the sources of data in Level 3 (**MINIMAL**) or could contain multiple datafiles per participant in Level 3 and DATA (with corresponding DICT and MISSING files) in Levels 1 and 2 (**MAXIMAL**). Non-tabular packages can have exactly zero or one file per participant (analog to wide tabular) or zero, one, or many files per participant (analog to long tabular). Any DATA/DICT part of non-tabular packages could be wide tabular or long tabular. Non-tabular packages

- are comprised of coherent datafiles that measure the same process (i.e., various resting state fMRI files, sensors during the same measurement period). Non-tabular packages ***cannot*** contain data across different measurements (e.g., mixture of structural and functional imaging, multiple eye-tracking tasks, different genetics chips)

- require consistency with respect to wide (exactly zero or one instance) or long (zero, one, or multiple instances) data structures. For example, if there are multiple data files per participant within the DATAFILES folder, then an equal number of listings should occur in the FILELIST.csv file ***as well as*** any DATA files within the package (at Level 1 or the respective Level 2). **Generally the "long" format should be avoided if possible. Please contact NIBS to discuss long vs. wide non-tabular formats**

- contain at least one data file per participant (or explanations of missing files)

- require README files within Level 1 and all Level 2 folders

- can contain tabular data packages as part of Level 1 and Level 2 under the condition that any DATA.csv files within the non-tabular package are dependent on the individual data files or vice versa: <u>any analyses should require both DATA and corresponding DATAFILES</u>. If a dependency does not exist, then DATA must be distributed as a separate tabular data package

    - Any DATA & DICT (and MISSING when required) files at Level 1 must pertain to <u>all</u> datafiles across Level 2s

    - Any DATA & DICT (and MISSING when required) files at a Level 2 must pertain to <u>only</u> datafiles within its respective Level 3

    - DATA & DICT (and MISSING when required) are formatted exactly as they would be for tabular packages

README requirements. Every Level 1 and Level 2 folder within non-tabular data packages must contain a README file. README files list all subsequent files and folders in a data package at that level. The README files

- must be comma separated value (csv) file. README format details are below with an example in **Appendix B**.

- contains exactly two columns in the following order:
    - **FILE**: All files and folders contained at that level in the data package.
    - **DESCRIPTION**: Brief description (no more than 200 characters) of the files and folders listed at that level.

- Level 1:
    - contains at least two rows with the names (FILE) and descriptions (DESCRIPTION) of
        - METHODS
        - Any Level 2 folders

- Level 2:
    - contains at least one row with the names (FILE) and descriptions (DESCRIPTION) of
        - DATAFILES

Subsequent rows for READMEs are included for DATA, DICT, additional files, MISSING, and SUP* as needed. READMEs required within each Level 2 folder.

Folder name requirements. Non-tabular data packages have three levels, each contained within a folder. Folder names at Level 1 and Level 3 are strictly defined; folder names at Level 2 can vary as necessary for the platform. Requirements, limitations, and options are listed below for each level.

Level 1: A single folder that contains the entirety of the data package. This folder requires the naming convention of

[ONDRI CODE]_[COHORT CODE]_[VISIT CODE]_[PLATFORM CODE]_[SUBPLATFORM CODE]_[DATA SET CODE]_[DATE OF RELEASE FOR CURATION]_DATAPKG.

Note: [SUBPLATFORM CODE]_[DATA SET CODE] can be of any length, subject to NIBS approval. The "core" items of names are the first four and the last two elements.

Each released file name must be unique. Each partition of the above format is described below and separated by underscores ("_"). SUBPLATFORM CODE and DATA SET CODE are described last and in more detail.

- **ONDRI CODE:** ONDRI study to which these data belong. See **Appendix A**.
- **COHORT CODE:** Cohort (disease) to which these data belong. See **Appendix A**.
- **VISIT CODE:** Visit at which data were collected. See and **Appendix A**.
- **PLATFORM CODE:** Assessment platform to which the data belong. See **Appendix A**.
- **SUBPLATFORM CODE:** Some platforms have multiple data files. For examples:
    - Genomics (GNMC) has two sources of genetic data: NeuroX and ONDRISeq.
    - Neuroimaging (NIMG) has multiple data modalities and types, e.g., resting state fMRI, DTI, SABRE-LE estimates. See **Appendix A**.
- **DATA SET CODE**: Some platforms have multiple data files. For examples:
    - Clinical (CLIN) will have multiple partitions that generally reflect certain types of data (e.g., demographics, MOCA, questionnaires, disease specific). These will be denoted by brief additional descriptors e.g., Neuropsychology (NPSY) and Neuroimaging's SABRE-LE (NIMG_SABR) each have two data sets: a "full" with comprehensive and "item level" data and a "minimum" which generally reflect summaries derived from the full data. See **Appendix A**.
- **DATE OF RELEASE FOR CURATION:** The date a platform puts data into LabKey's curation folder with the intent that it goes through adherence check and outlier analyses.
- **DATAPKG**: Truncation of "Data package".

Level 2: These folders must be uniquely named strictly using alphanumeric [A-Za-z0-9] and underscores ("_"). Folders cannot start with digits [0-9] or underscores. If digits are required they should immediately precede an underscore or appear at the end and only used if necessary (e.g., per naming convention, not to iterate). **NOTE**: A platform should define these folders in advance because they are required to be common across cohorts and visits.

Level 3: The folder name must be "DATAFILES".

File name conventions & requirements.

- Level 1:
    - For nearly all files, follow the conventions for the tabular data packages
    - For “Additional files” and folder names (Level 2s) make clear names that are human and machine readable, limited to alphanumerics (A-Za-z0-9) and underscores (_)

- Level 2:
    - For nearly all files, follow the conventions for the tabular data packages
    - For “Additional files” make clear names that are human and machine readable, limited to alphanumerics (A-Za-z0-9) and underscores (_)
    - A folder named DATAFILES (required)

- Level 3:
    - [...]_FILELIST.csv
    - [...]_MISSING.csv
    - For individual files, all files must be named with at least their ***release ID***, followed by the file extension. When necessary or convenient, the released ID can be followed by additional but brief subnames (e.g., subplatform, data type) separated by underscores, then the file extension.
    - [...]_DICT.csv is a “sometimes required” file: when all of the individual files are generally text or tabular, and have the same column structure, curators can and should provide a DICT file to describe the (common) column names (across all files). ***NOTE***: This only applies when ***all*** data files within Level 3 have exactly the same column structure.

Additional file(s) requirements. In many cases, non-tabular data have additional files that do not qualify as SUP nor DATA, but are required in order to perform analyses. These include, for examples, templates and masks from imaging data, stimulus onset timings from tasks, or meta-data and acquisition parameters.

- Any additional files are permissible so long as they are deemed necessary for all analyses. **NOTE:** The number of these files should be limited to whatever is minimal.

- Additional files can be of any type or format

- Additional files can be included in Level 1 or Level 2. The primary requirement is that the additional files apply to all data files in subsequent levels: any additional files at Level 1 apply globally across the entire data package, where as any additional files at Level 2 apply only to the DATA & DATAFILES at that respective level.

- Any additional files must be listed in the respective README files

NOTE: Before any preparation of non-tabular packages and these files, contact NIBS.

DATAFILES and FILELIST.csv requirements. The FILELIST.csv file(s) and DATAFILES folder(s) comprise the bulk of Level 3. Together, DATAFILES is the core data of non-tabular data packages. For examples, see **Appendix B**.

- DATAFILES could exist in “wide” and “long” analogs (see below for MISSING files):
    - “wide” non-tabular data contain exactly one file per participant within the DATAFILES folder
    - “long” non-tabular data contain more than one file per participant within the same DATAFILES folder

  NOTE: In most cases long non-tabular data are unnecessary, as the files can be split across multiple “Level 2” folders and thus the wide format. **Long non-tabular data packages are at the discretion, direction, and approval of NIBS.**

- The DATAFILES folder must contain: (1) one or multiple data files per participant and (2) the FILELIST.csv file, and could contain a MISSING.csv file. If the data package is:

    - “wide” (zero or one instance per participant) then the MISSING.csv file should include only one entry per entirely missing participant file

    - “long” (zero, one, or multiple instances) then the MISSING.csv file should include as many entries per participant as missing files, e.g., if participants should have 5 files each, but a participant is missing 3, they will have 3 entries in the MISSING.csv file.

- FILELIST.csv must contain all appropriate identifying columns: SUBJECT, VISIT, DATE(s), SITE(s), and FILENAME. In almost all cases only a single DATE and single SITE correspond to a data file, thus FILELIST.csv will have exactly five columns. The FILENAME column must list the exact and entire filename (including extensions) of the corresponding subjects data files.

Missing File Requirements. All data packages must account for every enrolled participant *at the time of the respective visit*. Participants without at least one data point should not appear in the DATA file but in a corresponding MISSING file. MISSING files must briefly explain why all data and data files are missing for necessary participants.

- The MISSING file must be released as a comma separated value (csv) file. MISSING format details are below with examples in **Appendix B**.

- Each MISSING file must contain six columns in the following order:
    a. **SUBJECT** and **VISIT**, each with the same requirements as the DATA file.

    b. **Columns describing the site(s) and date(s) of acquisition**. The SITE and DATE columns must have the same column names as are used in the DATA file. If data were collected but are not usable, the site and date of acquisition should be reported. If data were never collected, the cells should contain valid missing codes.

    c. **MISSING_CODE**: a valid missing code that accurately describes the reason all data are missing. See **Appendix A**.

    d. **DESCRIPTION**: a brief description (in words; no more than 200 characters) of the reason all data are missing.

- MISSINGness is hierarchical (and has dependencies) for non-tabular packages
    a. MISSING files for DATA/DICT within Level 1 or any Level 2 describe the missingness for its corresponding DATA entry and it is expected that there are no DATAFILES (Level 3) for these participants. MISSING for Level 3 describes the missingness for its corresponding file(s). These could stem from the same reason, or have different reasons. ***Contact NIBS on how to best address MISSINGNESS across DATA entries and DATAFILES.***

    b. Level 1: When a participant is missing *entirely* from the package; no instances of data should exist in any subsequent Levels.

    c. Level 2: When a participant is missing from this level and its subsequent Level 3 (DATAFILES)

    d. Level 3: When a participant is missing ***only*** a specific DATAFILE.

Other data packages

In some cases, data may require an alternate format from the specified types because either (1) the data conform to neither data package type or (2) it is best practice to maintain data in a well-established convention or standard, and adapt the ONDRI standards for those data. One example of data that do not conform to either package are the NeuroX platform (i.e., genome-wide data).

The decision to define and require alternate or additional standards for data release is at the sole discretion of NIBS. In these cases, NIBS will work with the platform to define a custom data package type that has features of the existing conventions (e.g., .ped/map files of genome-wide data) and the ONDRI standards (folder structures, dictionaries, README, missing, etc…).

Data Examples and Software

The ONDRI-NIBS Github page houses much of our public-facing materials. With respect to data and data standards, there are some important repositories on the Github page. These are briefly described below.

Data Examples

Examples of tabular and non-tabular data can be found in the "toy_data" repository: https://github.com/ondri-nibs/toy_data. The toy_data repository includes full examples of both types of data packages, and a README to help guide users through the repository.

Software

All public-facing software can be found here: https://github.com/ondri-nibs. However, there are several packages tailored for standards or to accommodate ONDRI data.

The Standards ShinyApp. We provide a Shiny/R app to perform standards checks: https://github.com/ondri-nibs/standards_app. At this time, the "Standards ShinyApp" generally handles structural standards, some project standards (e.g., OND01, OND05), and allows for some customization for project standards checks.

The Standards package. We provide an R package to perform *structural standards* checks: https://github.com/ondri-nibs/standards_package. This package does not offer *project standards* checks (e.g., correct participants, short codes, date ranges). The "Standards Package" emphasizes correct structuring and formatting of contents within a data package.

The ondricolors package. ONDRI has a set of standardized colors designated for each of the recruitment cohorts (from OND01, but applies to other OND-based projects): https://github.com/ondri-nibs/ondricolors. The "ondricolors package" ensures the correct use of colors assigned to cohorts by their short codes.

The ONDRIdf package. We provide an R package designed to read in DATA.csv and DICT.csv pairs, which stores DICT information as part of a data.frame (the most common data object in R), and also accommodates ONDRI's missing codes. *Currently this package is a prototype, with some known issues. Please use with caution.* Please see the code in the ONDRIdf package for some suggested strategies on how to maintain DATA & DICT information as a single object in R: https://github.com/ondri-nibs/ONDRIdf.

References

Arregoitia, L. D. V., Cooper, N., & D'Elía, G. (2018). Good practices for sharing analysis-ready data in mammalogy and biodiversity research. *Hystrix, the Italian Journal of Mammalogy*, *29*(2), 155–161. https://doi.org/10.4404/hystrix-00133-2018

Baumer, B. S. (2018). Lessons From Between the White Lines for Isolated Data Scientists. *The American Statistician*, *72*(1), 66–71. https://doi.org/10.1080/00031305.2017.1375985

Broman, K. W., & Woo, K. H. (2018). Data Organization in Spreadsheets. *The American Statistician*, *72*(1), 2–10. https://doi.org/10.1080/00031305.2017.1375989

Ellis, S. E., & Leek, J. T. (2018). How to Share Data for Collaboration. *The American Statistician*, *72*(1), 53–57. https://doi.org/10.1080/00031305.2017.1375987

Ercole, A., Brinck, V., George, P., Hicks, R., Huijben, J., Jarrett, M., Vassar, M., Wilson, L., & Collaborators, the D. (2020). Guidelines for Data Acquisition, Quality and Curation for Observational Research Designs (DAQCORD). *Journal of Clinical and Translational Science*, *4*(4), 354–359. https://doi.org/10.1017/cts.2020.24

Appendix A: Codes

Appendix A provides a list of short codes for global usage across OND01, OND05, OND06 and OND08.

Global missing codes. Missing data are defined missing or undefined values for individual cells of data for rows (participants) and columns (variables) that are otherwise intact. These codes are meant for usage as descriptive and static missingness at present (i.e., a specific data file), and not meant to infer any potential future missingness. Reasons for missing data generally fall into several categories. We have outlined those categories and the missing codes below. All missing indications must be preceded by "M_" to denote that this is missing and not some other valid data (e.g., free text).

ABBREVIATION	BRIEF_DESCRIPTION	LONG_DESCRIPTION
M_CB	Cognitive/behavioural impairment	The participant could not participate in the assessment or task because of cognitive impairment (it is assumed as a result of the disease).
M_PI	Physical impairment	The participant could not participate in the assessment or task because of an impairment (likely not temporary) that is related to the assessment and physically prevents the participant from completing that task. For example, glaucoma would be a disability (and is permanent) that prevents participants from doing the SDOCT assessment but a broken hip would not fall under M_PI (rather M_OTHER). physical impairment . If additional information is available to indicate that the physical impairment was a result of the disease (e.g., severe tremor), or not because of the disease (e.g., lifelong colour-blindness) please provide additional information as needed in SUP* materials.
M_VR	Verbal refusal	The participant chose not to participate in the assessment or task.
M_AE	Administrative/administration error	The administrator of the assessment or task made a mistake that caused the participant's score to be invalid or otherwise unobtainable. E.g., miscalibration of instruments by administrators, incorrect administration, scheduling conflicts.
M_DNA	Did not apply	The task or assessment was not performed, or a value could not be derived because it did not apply to the participant. For examples, a questionnaire asks if a participant can organize his/her own medications, but the participant does not take any medications so the question does not apply, percentage of error types when no errors occurred.
M_TE	Technical/equipment error	The instruments or tools (hardware or software) used for

		data acquisition, collection, or processing failed. Examples: Computer crashed, power failure, response button box or keyboard was not properly hooked up or defective, processing failures.
M_NP	Not part of protocol	Data did not exist for a participant because the assessment or task was not part of protocol. Examples: the ALS cohort was not asked to complete a particular assessment; a new instrument was introduced after baseline and screening.
M_ART	Artifacts	Data are not usable due to artifacts. Example: unusable FLAIR because of foreign bodies or motion.
M_TBC	To be completed	*To be used sparingly.* Data are currently unavailable but will eventually be added. The only circumstance under which this should occur is when a release is required but data have not yet been processed, curated, or subjected to outlier analyses. ***Requires approval from NIBS.***
M_OTHER	Other	*To be used sparingly.* When missing data are not covered by previous categories *and* when none of the existing missing codes apply. If a common reason for missing data exists it should not be qualified as “OTHER”, then we may need additional missing codes. Examples could include ‘No show’ without additional information or difficulty with reaching participant. If additional codes may be required, please contact Neuroinformatics. ***Requires approval from NIBS.***

Recruitment/primary site codes.

Site code	**Site Name**
BYC	Baycrest Health Sciences
CAM	Centre for Addiction and Mental Health
EBH	Elizabeth Bruyère Hospital
HDH	Hotel Dieu Hospital
HGH	Hamilton General Hospital
LHS	London Health Sciences Centre
MCM	McMaster Hospital
PCH	Providence Care Hospital
PKH	Parkwood Institute
SBH	Sunnybrook Health Sciences Centre
SMH	St. Michael’s Hospital
TBR	Thunder Bay Regional Health Sciences Centre

TOH	The Ottawa Hospital
TWH	Toronto Western Hospital

Platform-specific data-acquisition and site codes.

Site code	Site Name	Platform
IEI	Ivey Eye Institute	SDOC
KEI	Kensington Eye Institute	SDOC
OEI	Ottawa Eye Institute	SDOC
ROB	Robarts Research Institute	GNMC/NIMG
TAN	Tanz Centre for Research in Neurodegenerative Disease	GNMC
SJH	St Joseph's Healthcare Hamilton	NPSY/NIMG
QNS	Queen's University	NIMG
SBH_G	Sunnybrook Health Sciences Centre: GE scanner	NIMG
SBH_S	Sunnybrook Health Sciences Centre: Siemens scanner	NIMG
OTH	Other (e.g., participant's home, assisted living facility)	CLIN; NPSY; **Use of OTH requires approval from Neuroinformatics.**

NOTE: Contact NIBS if any site or platform-specific site/acquisition codes are missing.

Primary platform codes.

Primary platform code	Primary platform
CLIN	Clinical
NIMG	Neuroimaging
EYTK	Eye tracking
SDOC	Spectral domain optical coherence tomography
GABL	Gait & Balance
NPSY	Neuropsychology
GNMC	Genomics
NPTH	Neuropathology
NIBS	Neuroinformatics/Biostatistics
SNSR	Sensor technology

Subplatform codes. NOTE: Not all are listed as these can update over time.

Primary platform code	Primary platform	Subplatform code	Subplatform description
NIMG	Neuroimaging	SABR	SABRE-LE data
NIMG	Neuroimaging	DTI	White matter imaging
NIMG	Neuroimaging	FMRI	resting state fMRI
GABL	Gait & Balance	GAIT	Gait
GABL	Gait & Balance	BLNC	Balance
NPSY	Neuropsychology	LANG	Speech & Language
EYTK	Eye tracking	IPAST	Interleaved Pro & Anti Saccade Task
EYTK	Eye tracking	FVIEW	Behavior in the Free Viewing task
GNMC	Genomics	OSEQ	ONDRISeq Next-Generation Sequencing
GNMC	Genomics	NRX	NeuroX Genotyping
SDOC	Spectral domain optical coherence tomography	RT	Retinal thickness
SDOC	Spectral domain optical coherence tomography	RNFL	Retinal nerve fiber layer
SDOC	Spectral domain optical coherence tomography	OCLA	Ocular assessment
SNSR	Sensor technology	GNAC	GeneActiv device
SNSR	Sensor technology	BITF	Bittium Faros 180 device
SNSR	Sensor technology	ADAM	Device for speech data
SNSR	Sensor technology	NONW	NONIN Pulse Oximeter device for sleep data

NOTE: If a platform has multiple data sources or subplatforms please contact NIBS, as we need to establish unique codes for each data sources or subplatforms.

Cohort codes

Cohort (disease) code	Cohort description
ADMCI	Alzheimer's Disease/Mild Cognitive Impairment
VCI	Vascular Cognitive Impairment
FTD	Frontotemporal Dementia
PD	Parkinson's Disease
ALS	Amyotrophic Lateral Sclerosis
ALL	All cohorts included in a single file. ***Not to be used at this time; reserved for the possible use by NIBS and/or for eventual merging.***

Visit codes

Visit name	Visit time (window)	Visit code
Visit 1	Screening (within 8 weeks of consent)	01
Visit 2	Baseline (within 8 weeks of consent)	02
Visit 3	6 month (+/- 2 weeks)	03
Visit 4	12 month (+/- 4 weeks)	04
Visit 5	18 month (+/- 2 weeks)	05
Visit 6	24 month (+/- 4 weeks)	06
Visit 7	30 month (+/- 2 weeks)	07
Visit 8	36 month (+/- 4 weeks)	08
CV	Cross visit	Collection rolls across visits and is not constrained to any particular visit. **Requires NIBS approval.**
ALL	All visits together	All visits included in a data package. ***Reserved strictly for the use in package names and descriptors, never for a VISIT code in the data.*** **Requires NIBS approval.**

Data type codes. For a better idea of categorical, ordinal, and numeric (interval and ratio scale) data types please see https://en.wikipedia.org/wiki/Level_of_measurement

CODE	DESCRIPTION	Occurrence
TEXT	Free text	To be used only for free text
DATE	Date	To be used only for dates (in YYYYMMMDD format)
CATEGORICAL	Nominal data	When data have non-ordered categories, e.g., Male & Female, disease group, genotype, Yes & No.
ORDINAL	Ordered levels with unknown or unequal increments between levels.	When data levels are rank ordered with unequal or unknown increments between levels. Examples: confidence judgements, sliding scales, income levels, (arguably) Likert scales.
NUMERIC	Generally interval or ratio scale. Also includes counts.	When data exist across large ranges that generally span (1) negative and positive values or (2) values with meaningful and defined zeros (e.g., counts, temperature). Platforms should provide some additional information (in the dictionary) to indicate increments or how values are obtained.
MIXED	A single variable (column) that contains a mixture of data types. Examples include: a mixture of ordinal and numeric, a mixture of continuous and counts, a mixture of categorical and ordinal.	To be used only in very specific (and rare) cases. Please contact the Neuroinformatics team if: (a) you have a mixed variable, (b) you might have a mixed variable, or (c) you are unsure if you have a mixed variable. **Requires approval from NIBS.**
TIME	A timestamp or timestamp/range format	To be used only for data that requires time of day (or a time range) with hours, minutes, and/or seconds. Reference timezone must be included. For now, when using TIME, format this as specified, but list the variable type as TEXT until further notice from NIBS. **Requires approval from NIBS.**

NOTE: Occurrence should also include units whenever there are units.

NOTE: Contact NIBS if you have a variable that is of mixed types or does not fit into any of the above categories.

Appendix B: Examples

Appendix B provides illustrative examples of formatted files, but shown here in tables. For a more thorough set of examples, please refer to the **toy_data** repository, which has two complete (illustrative) data packages (a tabular, and a non-tabular): https://github.com/ondri-nibs/toy_data

Example README.csv file

FILE	DESCRIPTION
OND01_VCI_01_NPSY_MINIMUM_2017DEC01_DATA.csv	Neuropsychology platform's minimum data set for VCI participants at Visit 01.
OND01_VCI_01_NPSY_MINIMUM_2017DEC01_DICT.csv	Data dictionary for neuropsychology platform's minimum data set for VCI participants at Visit 01.
OND01_VCI_01_NPSY_MINIMUM_2017DEC01_METHODS.pdf	Methods document with detailed information for neuropsychology platform's minimum data set for VCI participants at Visit 01.

Example DATA.csv and corresponding DICT.csv files

SCENARIO 1.

DATA

SUBJECT	VISIT	NPSY_SITE	NPSY_DATE	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7
OND01_BYC_0001	01	BYC	2017OCT26	Y	1	-23.69	M	10	0	Lamb Gun Church River Fish Desk Desk
OND01_BYC_0002	01	BYC	2017OCT02	N	1	-84.54	F	11	1	Pencil Church Fish
OND01_BYC_0003	01	BYC	2017OCT26	Y	2	55.51	M	20	2	Church Fish Desk Ranger Ranger R River E Fish R Water SC E
OND01_SBH_0001	01	BYC	2017OCT02	Y	4	11.03	M	2	10	Lamb Council E Fish Bird Desk Pencil
OND01_TWH_0002	01	TWH	2017APR09	N	3	44.35	F	5	4	Church Fish Pencil Cloud Fish R

DICT

COLUMN_LABEL	DESCRIPTION	TYPE	VALUES
SUBJECT	Participant ID	TEXT	
VISIT	Visit code	CATEGORICAL	See Neuroinformatics dictionary(ies)
NPSY_SITE	Site of data acquisition	CATEGORICAL	See Neuroinformatics dictionary(ies)
NPSY_DATE	Date of data acquisition	DATE	
VAR1	Short description of var 1	CATEGORICAL	Y = Yes; N = No
VAR2	Short description of var 2	ORDINAL	1 = Like; 2 = Somewhat like; 3 = Neutral; 4 = Somewhat dislike; 5 = Dislike

VAR3	Short description of var 3	NUMERIC	Possible range: -100 to 100
VAR4	Short description of var 4	CATEGORICAL	M = male; F = female
VAR5	Short description of var 5	ORDINAL	Possible range: 1-25 (1 = dislike; 25 = like)
VAR6	Short description of var 6	NUMERIC	Possible range: 0 to 10 by integers
VAR7	Short description of var 7	TEXT	Responses to RAVLT

SCENARIO 2. Version 1: Specific measures happened on specific dates.

DATA

SUBJECT	VISIT	NPSY_SITE	NPSY_DATE_PRIM	VAR1	VAR2	VAR3	NPSY_DATE_SECD	VAR4	VAR5	VAR6	VAR7
OND01_BYC_0001	01	BYC	2017OCT26	Y	1	-23.69	2017NOV09	M	10	0	Lamb Gun Church River Fish Desk Desk
OND01_BYC_0002	01	BYC	2017OCT02	N	1	-84.54	16OCT2017	F	11	1	Pencil Church Fish
OND01_BYC_0003	01	BYC	2017OCT26	Y	2	55.51	2017NOV09	M	20	2	Church Fish Desk Ranger Ranger R River E Fish R Water SC E
OND01_SBH_0001	01	BYC	2017OCT02	Y	4	11.03	16OCT2017	M	2	10	Lamb Council E Fish Bird Desk Pencil
OND01_TWH_0002	01	TWH	2017APR09	N	3	44.35	2017APR23	F	5	4	Church Fish Pencil Cloud Fish R

DICT

COLUMN_LABEL	DESCRIPTION	TYPE	VALUES
SUBJECT	Participant ID	TEXT	
VISIT	Visit code	CATEGORICAL	See Neuroinformatics dictionary(ies)
NPSY_SITE	Site of data acquisition	CATEGORICAL	See Neuroinformatics dictionary(ies)
NPSY_DATE_PRIM	Date of data acquisition for initial tests	DATE	This date applies to VAR1, VAR2, and VAR3
VAR1	Short description of var 1	CATEGORICAL	Y = Yes; N = No
VAR2	Short description of var 2	ORDINAL	1 = Like; 2 = Somewhat like; 3 = Neutral; 4 = Somewhat dislike; 5 = Dislike
VAR3	Short description of var 3	NUMERIC	Possible range: -100 to 100
NPSY_DATE_SECD	Date of data acquisition for the data for follow-up tests	DATE	This date applies to VAR4, VAR5, VAR6, and VAR7
VAR4	Short description of var 4	CATEGORICAL	M = male; F = female

VAR5	Short description of var 5	ORDINAL	Possible range: 1-25 (1 = dislike; 25 = like)
VAR6	Short description of var 6	NUMERIC	Possible range: 0 to 10 by integers
VAR7	Short description of var 7	TEXT	Responses to RAVLT

NOTE: The mapping of each date to their respective data must also be described in the platform's data dictionary. Data files should be organized in a fashion where data columns immediately follow their respective date columns. In the above example, variables VAR1, VAR2, and VAR3 belong to NPSY_DATE_PRIM, and variables VAR4, VAR5, VAR6, and VAR7 belong to NPSY_DATE_SECOND.

SCENARIO 2. Version 2: Only some participants required a follow up visit to complete tests.

DATA

SUBJECT	VISIT	NPSY_SITE	NPSY_DATE_PRIM	NPSY_DATE_SECD	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7
OND01_BYC_0001	01	BYC	2017OCT26	2017NOV09	Y	1	-23.69	M	10	0	Lamb Gun Church River Fish Desk Desk
OND01_BYC_0002	01	BYC	2017OCT02		N	1	-84.54	F	11	1	Pencil Church Fish
OND01_BYC_0003	01	BYC	2017OCT26		Y	2	55.51	M	20	2	Church Fish Desk Ranger Ranger R River E Fish R Water SC E
OND01_SBH_0001	01	BYC	2017OCT02		Y	4	11.03	M	2	10	Lamb Council E Fish Bird Desk Pencil
OND01_TWH_0002	01	TWH	2017APR09	2017APR23	N	3	44.35	F	5	4	Church Fish Pencil Cloud Fish R

DICT

COLUMN_LABEL	DESCRIPTION	TYPE	VALUES
SUBJECT	Participant ID	TEXT	
VISIT	Visit code	CATEGORI	See Neuroinformatics dictionary(ies)

		CAL	
NPSY_SITE	Site of data acquisition	CATEGORICAL	See Neuroinformatics dictionary(ies)
NPSY_DATE_PRIM	Date of data acquisition for initial tests	DATE	Date of testing
NPSY_DATE_SECD	Date of data acquisition for the data for follow-up tests (if required)	DATE	Date of follow up testing for participants that did not complete testing on the first visit
VAR1	Short description of var 1	CATEGORICAL	Y = Yes; N = No
VAR2	Short description of var 2	ORDINAL	1 = Like; 2 = Somewhat like; 3 = Neutral; 4 = Somewhat dislike; 5 = Dislike
VAR3	Short description of var 3	NUMERIC	Possible range: -100 to 100
VAR4	Short description of var 4	CATEGORICAL	M = male; F = female
VAR5	Short description of var 5	ORDINAL	Possible range: 1-25 (1 = dislike; 25 = like)
VAR6	Short description of var 6	NUMERIC	Possible range: 0 to 10 by integers
VAR7	Short description of var 7	TEXT	Responses to RAVLT

NOTE: The mapping of each date to their respective data must also be described in the platform's data dictionary. In the above example participants should have completed the tests on one date, but some had to return to complete them. There is no indication of which variables belong to which dates in this example.

MISSING

SUBJECT	VISIT	NPSY_SITE	NPSY_DATE	MISSING_CODE	DESCRIPTION
OND01_CAM_0001	03	CAM	2015APR29	M_AE	Administration Error (AE) - Site Coordinator (certified to administer the scale) thought that scale would only apply for stroke related events and has rated all Participants as 0 = No symptoms at all
OND01_CAM_0002	03	BYC	M_PI	M_PI	Participant could not attend visit as she became immobilized after serious fall.
OND01_CAM_0003	03	PKH	2017FEB21	M_TE	Technical error- Device malfunction which erased participant's data entirely
OND01_CAM_0004	03	TWH	M_OTHER	M_OTHER	Participant cancelled the current visit and did not complete the NPSY assessment

For non-tabular examples, please see the above on how to format the file common to tablar & non-tabular packages, and also see https://github.com/ondri-nibs/toy_data

Contributions and contacts

Creation and authorship this document is outlined below. There are four roles, three of which are defined by the CRediT system (https://casrai.org/credit/) with adaptations of the definitions as necessary, see also (CITE: https://www.pnas.org/content/115/11/2557):

- ***Conceptualization***: Ideas; formulation or evolution of overarching goals and aims.
- ***Writing - original draft***: Preparation, creation and/or presentation of the published work, specifically writing the initial draft.
- ***Writing - review & editing***: Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision
- ***Additional contributions***: Contributions to this document not otherwise captured in Conceptualization, Writing (draft), or Writing (reviews/edits).

Conceptualization, Writing (draft), and Writing (review/edit) are not exclusive from one another but are exclusive from Additional Contributions. If one appears in Contributions they do not appear in Conceptualization, Writing (draft), and Writing (review/edit) and vice versa. *Names are listed alphabetically*.

Conceptualization	*Writing (initial draft)*	*Writing (review/edit)*	*Additional Contributions*
Stephen Arnott Derek Beaton Donna Kwan Kelly Sunderland	Stephen Arnott Derek Beaton Donna Kwan Kelly Sunderland	Stephen Arnott Derek Beaton Malcolm Binns Donna Kwan Paula McLaughlin Stephen Strother Kelly Sunderland	Jedid Ahn Brian Coe Allison Dilliot Kristen Lutz Mojib Javadi Mojdeh Zamyadi