Data Curation Glossary

On behalf of the Neuroinformatics and Biostatistics (NIBS) platform

This document contains 2 sections: the glossary and the authorship/contributions to the document. The glossary contains a specific set of terms (nomenclature, terminology) commonly used for the ONDRI and NIBS curation processes.

1. Glossary (pg. 2)
2. Authorship and contributions (pg. 5)

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

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

Version: 2.1.0.0

Curation glossary

This glossary contains shortened descriptions of commonly used terms, acronyms, and initialisms throughout the NIBS documentation. Terms are generally listed by importance (not alphabetically), and are grouped together when related.

NIBS: Acronym for ***N***euro***i***nformatics and ***B***io***s***tatistics platform.

Data package: Sets of files that include data and required companion files (e.g., dictionaries, readmes). **Note**: In most cases a data package is per cohort and per visit.

Cell (data): a single location (row by column) in a spreadsheet

Tabular (data): Data contained entirely within a single spreadsheet with a single piece of information per cell.

Non-tabular (data): Data that cannot be contained within a single spreadsheet, for examples a file or multiple files per participant such as neuroimaging, group level paired files such as genomics.

Wide (data): Tabular data with participants listed exactly zero (missing) or one time down the rows.

Long (data): Tabular data with participants listed zero, one, or many times down the rows because of a *repeated factor* (e.g., time, condition).

Platform: Interchangeably used to refer to the assessment team and/or the modality(ies) of data from the respective assessment team. For example “neuropsychology” or “neuroimaging” refer to platforms (NPSY and NIMG initialisms, respectively) but also to the modality(ies) of data those platforms produce.

Structural Standards: Standards for data and data packages specifically for formatting, structure, and content. For examples: having the correct files (DATA, DICT, README, etc…), uniform precision levels, correct formatting of names and labels (A-Za-z0-9 and underscores), correctly named files.

Project Standards: Standards for project-level and project-specific content. For examples: correct list of participants, valid date ranges, short codes.

Curation: The general process within a platform of preprocessing and preparation of data package(s) for the release process.

Curator: The person or persons within a platform or subplatform as the designated individual to curate data for release.

Consumer: The person or persons using the data for analysis, reporting, presentation, and/or publication purposes.

Standards: The set of requirements for data packages, including but not limited to required files and their formats.

Standards checks*:* A mostly programmatic check for adherence to the standards; some manual checks also performed.

Outlier analyses*:* Analysis pipeline designed to identify outliers (anomalies or potential errors) in a data set. Results provided by NIBS and to be checked and verified by platform.

Outlier*:* A highly deviated individual observation (participant) on a value (univariate) or multiple values (multivariate).

> ***Anomaly/anomalous outlier***: Outlying individual with verified correct values
>
> ***Error outlier***: Outlying individual with verified incorrect values (to be corrected)

Curation-to-release process: The NIBS team performs two steps once a data package has been submitted: standards checks and outlier analyses. On completion and agreement (between NIBS and platform) data can be ***released***.

Release: A data package exits the curation process and is made available to qualified researchers.

Release request: On completion of the curation-to-release process (agreed to by a platform and NIBS), the platform sends a request for release of the data package to the data release committee

Withdrawal (of data): A request from the platform to NIBS *during* the curation-to-release process but *prior* to release. Withdrawals occur because of errors or issues with data packages discovered by the platform before provided with reports and results from NIBS.

Retraction (of data): A request from the platform to the data release committee and NIBS *after* data have been released. Retractions occur because of errors or issues with data packages discovered by any qualified researcher and verified by the platform. In all cases a notification must be provided to all qualified researchers, and in serious cases data are immediately removed. In the event of retraction, corrected data must go through the ***curation-to-release process*** for ***re-release***

Re-release (of data): Subsequent ***releases*** of existing data packages, often because of additions, updates, or corrections.

Missing (data): Data or datum that *should* have been acquired for distribution but does not exist. Generally we refer to two global types of missing data (with specific definitions in the *Standards* document):

> ***Sporadic missing***: Single pieces of datum generally not recorded, e.g., participant did

not respond to specific questions or perform specific tasks.

Completely missing: Data missing on the whole for a participant, e.g., declined neuroimaging scans, pathology discovered in ocular image that prevents data distribution, miscalibration renders eye tracking task unusable

Codes: Prespecified short codes for ONDRI projects, visits, sites, etc… See the *Standards* document and **Appendix A** in the *Standards* document for a list of codes and code types.

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*
Derek Beaton	Derek Beaton Kelly Sunderland	Derek Beaton Kelly Sunderland	