ONTARIO BRAIN INSTITUTE | INSTITUT ONTARIEN DU CERVEAU

OBI Data Quality Framework

Stage 2 – Data Collection, Curation and Study Completion

Version: 1.2 - Date: August 8, 2024

1. Introduction

Capturing high quality data is critical to science and to supporting translation to clinical care. To build trust in the data, OBI aims to achieve world-class data quality practices. Data quality measures will be captured and shared as part of the Brain-CODE data release metadata in a standardized manner, so it is essential to maximise this quality up-front. The following guidelines aim to support the OBI research networks in establishing best-in-class practices for data quality across modalities. Please note that these guidelines may evolve as methods, techniques, and new modalities arise. If your modality is not covered by this section, please contact the OBI informatics team at help@braincode.ca to discuss your modality and agree on a suitable data quality guideline.

This document covers Stage 2 of OBI's Data Quality Framework and follows Stage 1 requirements and recommendations. This document should be used by any Data Producer (including Integrated Discovery Programs) that will be contributing data to OBI's Brain-CODE platform. Items stated with "***must***" indicate a process or output that is mandatory to complete by Data Producers. Items stated with "***should***" or "***recommends***" indicate highly recommended processes or outputs, but variations can be implemented in consultation with OBI.

2. General Data Quality Guidelines

The following requirements and recommendations apply to all study data modalities.

1. Data Producers ***must*** ensure timelines and frequencies of collection align with protocols and milestones as stated in the contract agreements. This should be stated clearly in the Data Management Plan.

2. Brain-CODE data capture tool users ***must*** complete electronic data capture (EDC) tools training (see the EDC Training Materials subfolder in Stage 1). Please contact help@braincode.ca for EDC training.

3. Data Producers ***must*** provide OBI with metadata about their study dataset.
    - This metadata should include study protocol information, cohort types, modalities, subject information, instruments and experimental conditions for each subject and assessment made.
    - OBI will track metadata information received from Data Producers and follow up in case of missing or updated information.

4. Brain-CODE REB ethics and consent tracking:
    - Data Producers ***must*** send their protocol and consent forms to OBI ***prior to*** REB submission for approval by OBI. Any changes ***must*** also be sent to and reviewed by OBI ***prior to*** submitting to REB.
    - Data Producers ***must*** send approved consent forms, REB approval letters, amendments and the REB submission package to OBI for approval ***prior to*** participant recruitment.
    - Please submit all documents to governance@braincode.ca.

5. Confidentiality and privacy practices ***must*** adhere to OBI's governance and security policies available on the Brain-CODE website: https://braininstitute.ca/docs/Brain-CODE-Governance-Policy-version-FINAL.pdf

6. Data Producers ***must*** complete the standard Brain-CODE REDCap Subject Enrollment Form and enter encrypted Ontario health card numbers to the Brain-CODE Subject Registry. This applies to both studies that capture data on Brain-CODE and studies that transfer data to Brain-CODE.

7. OBI ***recommends*** several data curation software tools, but Data Producers are not limited to these. Please refer to the OBI Software Tools Catalogue to explore currently used tools by the OBI IDP network.

8. OBI ***strongly recommends*** using the REDCap ***Participant Status Form*** designed by the Brain-CODE team that allows for a standardized method to track participant status during the study as participants complete scheduled study timepoints, and to track participant withdrawal (see Stage 1 document "OBI DQF REDCap Tools & Procedures"). It is ***recommended*** that data collection personnel at each site complete this form after each participant visit/data collection timepoint.

9. Data collection and sharing workflows
    - Data Producers ***must*** monitor their data on a regular basis. Data Producers and OBI will curate the data to reach a "**minimally curated dataset**". A minimally curated dataset is a dataset that is processed to minimize errors and missing information. OBI will perform data quality checks and reporting on a regular basis (e.g., quarterly) and share this with the Data Producers to address any quality issues. This will include evaluation of missing data and standards compliance, as applicable. If additional errors are detected, Data Producers will have the opportunity to make any necessary corrections.
    - The following diagram illustrates the general data collection, monitoring, curation and sharing cycle.

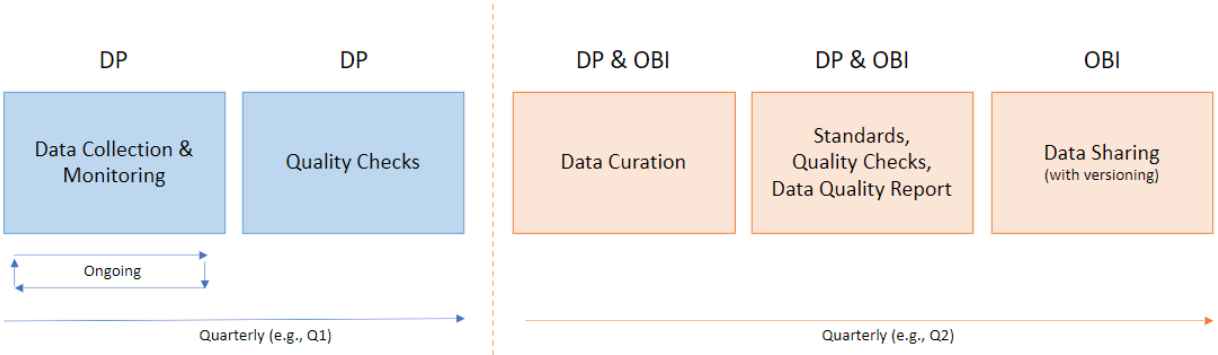

Figure 1. General data collection, monitoring, curation, and sharing workflow

10. Data producers ***must*** report (flag) all participant-identifying information within their Data Management Plan, including both direct and indirect identifiers, from all data collected as part of their data collection and curation procedures and remove identifiers as required by ethics approvals. This will help OBI make the data reusable under appropriate privacy controls. OBI will perform a secondary check before data are made more widely available.

11. Data Producers and OBI ***must*** complete the **Study Quality Checklist** (see section "Stage 2" in Stage 1 document #5) and prepare all applicable data quality files (as outlined in State 2 document #3 "OBI Data Quality Reporting Guide") to support OBI Data Quality Reporting. This ***must*** be conducted each quarter (or as agreed upon with OBI).
    - See Stage 2 document #3: "OBI DQF Data Quality Reporting Guide"

12. Study closeout and completion ***must*** be planned for in consultation with OBI.

3. Clinical and Tabular Data Quality Guidelines

Please review Stage 1 document "OBI DQF REDCap Tools & Procedures" for further training in support of the following guidelines.

1. Brain-CODE REDCap electronic case report forms (eCRFs) ***must*** be used as the primary data capture tool, when possible. Despite prior cases where double entry from paper and transcription to eCRFs was used, it is now ***recommended*** to capture data directly into eCRFs to reduce chances of errors and reduce participant burden.
    - Paper forms ***should*** only be used when electronic data capture is not available or poses a significant barrier for data collection.
    - Data Producers may agree with OBI to transfer data on a scheduled basis if this is the only mechanism for clinical data collection.

2. eCRFs ***must*** be tested by Data Producers and validated by the Brain-CODE team prior to use in production.

3. It is ***strongly recommended*** that OBI's missing data codes (see Stage 2 document #3 "OBI Data Quality Reporting Guide" and Stage 1 document "OBI DQF REDCap Tools & Procedures") be implemented in the missing data codes feature within the study's REDCap project so they can be inserted directly into fields in REDCap forms to identify missing items and analyzed using REDCap's Data Quality tools.

4. All data input fields ***must*** conform to input validation rules by the data collection software and protocol including numerical fields (e.g., age), timestamps or dates, and categorical fields (e.g., gender). File uploads ***must*** conform with the approved file types. The use of free-text input fields ***must*** only be used for required cases (e.g., qualitative

feedback). All free-text input fields ***must*** be documented for the data curation teams in a data dictionary.

5. Regular (e.g., weekly) data monitoring of 100% of captured fields is ***recommended*** for new research coordinators and can reduce to a lesser amount (e.g., 50%, 20%) via spot checks as data entry reliability improves.

6. REDCap's ***Data Resolution Workflow*** is ***strongly recommended*** as the formal process to correct errors and document changes. Furthermore, this will support compliance with good clinical practices.
    - ***If*** REDCap is not used for collection, Data Producers ***must*** implement robust data monitoring and correction measures to meet "OBI DQF Study Quality Checklist" (Stage 1 document #5) reporting requirements before transferring the data to Brain-CODE.
    - Please refer to the Stage 1 document "OBI DQF REDCap Tools & Procedures" for more details on the Data Resolution Workflow and how to use it.

7. Clinical instrument scoring calculations ***should*** be conducted in REDCap and must be verified before finalizing the datasets for sharing. Scoring/values for individual items ***should*** be included in the data export in addition to calculated scores.

8. Re-training ***must*** be provided by Data Producers to respective staff when major changes are made to (electronic) case report forms or to data collection or curation procedures.

4. Imaging Data Quality Guidelines

1. Images and related data files ***must*** be uploaded/transferred to SPReD/XNAT. Any data created during further offsite image processing and related documentation detailing the processing methods, ***must*** be transferred to Brain-CODE in accordance with the regular upload schedule.

2. Brain-CODE standard Subject ID and session ID conventions ***must*** be used when capturing data on SPReD/XNAT.

3. Standard file formats will be adopted by Brain-CODE when preparing data for reuse. Data Producers ***must*** ensure that curated data can be transformed into Brain-CODE data

structures and formats and provide support with this. The following standards will be used for imaging and time-series data:

- MRI, PET, CT, and X-ray: DICOM format upload to SPReD
- Task-based fMRI event data: shared as exported .CSV files (e.g., from E-Prime) at a minimum EEG and other time-series data: EDF+ format, unless it explicitly contravenes modality-specific BIDS recommendations (e.g., BIDS-MEG).
- The Brain-CODE team will implement the standard BIDS naming conventions for both imaging and time-series data. As part of the BIDS conversion:
  i) DICOM images will be converted to NIFTI format.
  ii) Time-series data will be converted to EDFs.

4. Data Producers, with support from OBI, ***must*** assign quality scores ("usable", "questionable", "unusable") on XNAT, validate parameters, and label images for classification. Please contact the Brain-CODE team for training.

5. Primary (raw) imaging files ***must*** be minimally curated (assigning a quality score to each scan and documenting preprocessing procedures in addition to minimizing errors and identifying missing data) in priority to meet Brain-CODE upload schedules. OBI will support this process once data are on SPReD/XNAT. Specifically, OBI will perform the following steps:
   - Perform naming checks, parameter checks, and redact headers, followed by conversion to NIFTI, then conversion to BIDS naming and structure.
   - Generate QC metrics.

6. Derived image analyses and related process documentation can be added to release packages at later stages if necessary.
   - Imaging curation pipelines and derived processing tools are catalogued in the OBI Software Tools Catalogue.

7. Imaging devices ***should*** be calibrated monthly using phantoms or as appropriate per site.

5. Molecular and Genomic Data Quality Guidelines

1. Lab protocols (e.g., next generation sequencing [NGS] protocols) ***should*** be collected with a standard set of field descriptors.
    - Standard schemas are ***recommended*** and available on GitHub (note: Case, Administrative, and Data schemas are not necessary and are redundant with existing Brain-CODE fields).

2. Genetic sequences ***should*** be mapped to reference genomes and call variants. The latest version of the references should be employed.

3. Standard gene annotations ***should*** capture the following information:
    - What does the gene do?
    - Is the gene a variant?
    - Which cell is the gene active in?
    - What proteins is the gene associated with?
    - What phenotypic information is the gene associated with?

    For example, some Data Producers have adopted the American College of Medical Genetics and Genomics Pathogenicity Guidelines: https://www.jove.com/t/57266/targeted-next-generation-sequencing-bioinformatics-pipeline-to

4. Standard file formats ***must*** be adopted:
    - Genetic alignment data in BAM or CRAM file formats (GA4GH) (these can include non-aligned data with the "unmapped" flag set in BAM/CRAM).
    - Genetic variant call format (VCFs) files in UTF-8 encoded text format.
    - Biosamples and other molecular data in CSV tabular format or UTF-8 encoded text format.
    - Molecular and genomics curation pipelines and derived processing tools are catalogued in the OBI Software Tools Catalogue.

5. Data Producers ***must*** ensure their processed genomic and molecular data is transferred to Brain-CODE at a frequency agreed upon with OBI.
    - If it is more appropriate to store raw data at the molecular/genomics sites, OBI may enter into an agreement to ensure ongoing access. However, processed data ***must*** be transferred to Brain-CODE in any case (e.g., BAM files, VCF files).
    - File md5 hash signatures ***must*** be created and shared with the Brain-CODE team prior to upload.

For data modalities not described in sections 3-5, please consult with the Brain-CODE team to adopt suitable quality guidelines by emailing help@braincode.ca.

6. Data Transfers

1. Data Producers ***must*** transfer all available data from **all participants** on a quarterly basis (or as agreed upon with OBI) for the given modality (even if those data were transmitted in prior transfers). Prior to data transfer, Data Producers ***must*** ensure that all participants are entered into the Brain-CODE Subject Registry (for encrypted OHIP numbers, if collected) and have completed Subject Enrollment Forms (***recommended*** upon participant's initial study visit).

2. Prior to the initial data transfer, Data Producers ***must*** establish a detailed Data Transfer Plan with OBI **for each study**. Please complete the "OBI Data Transfer Plan Template" (Stage 2 Templates folder) and submit to OBI. If any changes are made to the plan, please send an updated version of the Data Transfer Plan.

3. Data Producers ***must*** designate personnel who will work with the Brain-CODE team on post-transfer curation of data.

4. Data Producers and OBI ***must*** conduct quality control and curation on the data in accordance with the "OBI DQF Data Quality Reporting Guide" (Stage 2 document #3).

7. Next Steps

OBI is currently developing a strategy for the use of OBI-approved vocabularies and ontologies. In the future, Data Producers, with OBI support, ***should*** perform semantic coding by annotating data fields, measures, and methods using the presented vocabularies and ontologies.

- See Stage 2 document #2: "OBI DQF Recommended Vocabularies"

8. Support

If you have specific informatics questions regarding your research project, please contact your program's Neuroinformatics Lead and/or Program Manager. For any general questions or support requests regarding this Data Quality Framework, please contact help@braincode.ca.