CDS Specification

Overview

The Clinical Dataset Structure (CDS) is a simple and intuitive way to organize and describe clinical research datasets such that they are readily interoperable and easily reusable by humans and machines. The CDS is especially designed to optimize AI-readiness of datasets.

Why we need the CDS

During clinical research studies, multiple modalities of data are typically collected from study participants such as survey answers, blood test results, electrocardiogram data, magnetic resonance imaging (MRI), retinal images, etc. There exist some standards that specify how to structure the data and metadata for certain modalities such as the Brain Imaging Data Structure (BIDS) for MRI data, the SPARC Data Structure (SDS) for neuromodulation-related data, and the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) for observational clinical data. There is, however, no consensus on how to structure data and metadata from multiple modalities together into a consistently structured dataset. As a result, datasets from different clinical research studies are currently structured and documented differently which has two major impacts:

1. Datasets are not readily interoperable, meaning that it is difficult to combine datasets from different studies together. This limits, from instance, our ability to train AI/ML models on combined data from different studies.
2. Datasets are not easily reusable, meaning that it is difficult to understand and reuse data collected by someone else. This limits secondary analysis and reuse of data by external researchers not originally involved in a given study.

The Clinical Dataset Structure (CDS) is standard established to address these limitations by providing a simple and intuitive way for structuring clinical research data and associated metadata.

Development of the CDS

The CDS is developed as part of the AI-READI project funded by the NIH Bridge2AI Common Fund. We refer to the Governance section for information regarding the continuous development and maintenance of the CDS.

Acknowledgment

This project is funded by the NIH under award number 1OT2OD032644. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Structure

This documentation is conveniently organized so you can easily find what you are looking for:

- The "CDS Specification section" is for those of you who are interested in the fine details of the CDS. It contains the specification of the CDS, i.e., what is considered a datatype, how to create a CDS-compliant directory structure, what metadata files are required, what format should each metadata file follow, etc.
- The "Implementing the CDS section" contains a step-by-step guide to implement the CDS. We recognize that the specifications are quite extensive and can be difficult and time consuming to implement so we have developed support tools and resources designed to enable anyone to implement CDS without any specific knowledge or expertise. Those tools and resources are also introduced in this section.
- The "Design Rationale" section is for those of you who are interest on the rationale behind the design of the CDS. It contains detailed explanation on the different choices made in the CDS (Why one directory per datatype? Why the study_description.json file makes use of the ClinicalTrials.gov schema? etc.).
- The Resources section contains various information that didn't fit into the previous sections.

Maintenance

This documentation is maintained from its GitHub repository accessible here.

Suggestions/Feedback

Have a suggestion for improving the CDS? Need help implementing the CDS? Checkout instructions in our Contributing page.

Inspiration

The structure and some content of this documentation is inspired by the documentation of the BIDS specification.

If you use the CDS or any elements related to it, please reference the following resources:

```
B. Patel, S. Soundarajan, A. Gasimova, N. Gim, J. Shaffer, A. Lee (2024).
Clinical Data Structure (CDS). Zenodo. https://doi.org/10.5281/zenodo.14043020
```

lang: en-US title: General principles description: Description of the general principles of the CDS —

Definitions

- The keywords “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” in this document are to be interpreted as described in RFC2119.
- Clinical dataset: A set of data files and associated metadata files resulting from a clinical research study.
- Datatype: In a clinical research study, multiple modalities of data are collected. A datatype designate one modality or a group of modalities. We postulate that data from multiple modalities can be characterized as being of a single datatype if at least one of these is applicable:
  1. There exists an established standard structure for organizing data from these modalities together
  2. The modalities were collected through a shared method (instrument, device, etc.)
  3. The modalities cannot be interpreted separately
  4. The modalities are typically collected together

General specifications

The CDS specifies the following:

1. No data file must be present at the root level.
2. The data files must be organized into one directory per datatype at the root level as per the specification provided here.
3. Within each datatype-specific directory, data and metadata files must be organized according to an existing standard for that datatype, if available, or following the CDS-suggested structure, as explained here.
4. The following metadata files must be included at the root level:

- README.md
- LICENSE.txt
- CHANGELOG.md
- healthsheet.md
- study_description.json
- dataset_description.json
- participants.tsv and participants.json
- dataset_structure_description.json

5. No empty directories must be included anywhere in the dataset.

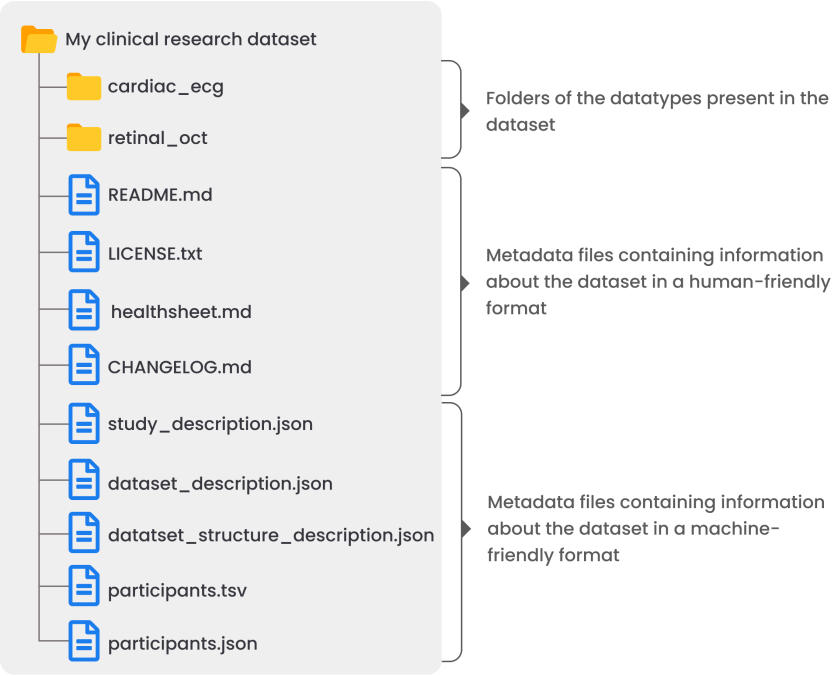

Figure 1: Illustration of the root level of a dataset structured following the CDS.

About

The CDS specifies to maintain one directory per datatype at the root level of a dataset. This page provides specification for deciding what constitutes a datatype and how to name each directory.

Specifications

Datatype directories

The data must be organized into one directory per datatype at the root-level. What constitutes a datatype is left to dataset authors' best interpretation of the datatype definition provided here.

Directory naming

Since there is no standard for datatype names, we recommend naming each datatype directory such that it best reflects its content. The following naming convention must be followed: Only a-z characters (lowercase English alphabet) and 0-9 (numerical characters) are allowed with the exception of underscores that must be used to separate words (no white space allowed).

Documenting directory structure

The name of each datatype directory, what datatype it contains, etc. must be documented in the dataset_structure_description.json metadata file.

About

The CDS specifies to maintain one directory per datatype at the root level of a dataset. This page provides specification about how data must be organized within each datatype directory.

Specifications

Datatype directories

Inside each datatype directory, the data and metadata must be structured according to applicable standards, if available. Resources such as fairsharing.org are recommended for identifying relevant datatype-specific standards.

If there is no existing community-agreed standard for structure data from a datatype, the CDS suggests following this structure within that datatype directory:

- Include one directory per modality
- Inside each modality directory include one directory per device/tool used to collect data from that modality.
- Inside each device directory, include one directory per participant
- Inside each participant directory, include the participant's data and participant specific metadata if applicable.
- Directly under the datatype directory and alongside the modality directories, include a manifest.tsv metadata file according to the indications provided in the dedicated section below.

manifest.tsv metadata

The manifest.tsv metadata file is intended to document all the data files within a datatype directory. It should follow this specification:

- A column called "filename" must be included where the paths of the data files, relative to the datatype directory, need to be listed.
- Additional column should be included as necessary to properly document relevant information about each datatype.
- The following naming convention must be followed for all column names: Only a-z characters (lowercase English alphabet) and 0-9 (numerical characters) are allowed with the exception of underscores that must be used to separate words (no white space allowed).

Directory naming

Since there is no standard for modality and device names, we recommend naming each modality and device directory such that it best reflects its content. The participant directories should be named exactly as the ID used for that participant in the study.

The following naming convention must be followed for all directories: Only a-z characters (lowercase English alphabet) and 0-9 (numerical characters) are allowed with the exception of underscores that must be used to separate words (no white space allowed).

Documenting directory structure

The structure under each datatype folder must be documented in the dataset_structure_description.json metadata file.

About

The README.md file is a metadata file that contains a detailed description of the dataset in a human-friendly format. You can think about it as a detailed abstract for your dataset, i.e. the first thing that a human user of the data will read.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `README` and must be in the markdown format. The full name with extension must thus be `README.md`.

Content

There is no mandatory requirements for the content of this metadata file. It is left up to the authors of the dataset to include information about their data they think would be critical to potential reusers of their dataset. The following section titles and content are highly recommended when available/applicable:

- **Include the name of the dataset at the top of the file**
- **Identifier**: In this section, indicate the DOI or any other globally unique and persistent identifier of the current version of the dataset.
- **Version number**: In this section, indicate the current version number of the published dataset (suggested format is major_release.minor_release, c.f.).
- **Publication date**: In this section, indicate the date when the current version of the dataset was published (i.e., made available outside of the project members openly or through a restricted access). It is suggested to follow the ISO 8601 date format (YYYY-MM-DD).
- **License**: In this section, mention the name of the data reuse license (refer to the LICENSE.txt file in your dataset for additional details).
- **Dataset access/restrictions**: In this section, explain how the dataset can be accessed and any conditions/restrictions for accessing it.
- **Overview of the study**: In this section, provide a high-level description of the study associated with the dataset. Include for instance identifiers of the study, a brief overview of the study protocol, external links (website, manuscripts, protocols, etc.) to find out more about the study, etc.
- **Description of the dataset**: In this section, provide a detailed description of the dataset. Include the number of study participants (refer to the participants.tsv file in your dataset for additional information), the datatypes collected, data deidentification approaches if any, the overall number of files and total size of the dataset, etc.
- **Data standards followed**: In this section, indicate the standards followed to structure the dataset, format the data files, etc. Make sure to include identifiers of the standards when available and/or link to the associated documentation.
- **Resources**: In this section, mention any specific resources (software, documentation, manuscripts, etc.) that are required/useful for using the data. Make sure to include identifiers and/or links to the resources.
- **How to cite**: In this section, provide instructions on how to cite the dataset if reused. Using the American Psychological Association (APA) style is suggested.

- **Contact**: In this section, provide contact information of someone who can be reached out with questions regarding the dataset. You can also refer to the study_description.json and dataset_description.json metadata files for information about contact person/entity, authors, and contributors of the dataset.
- **Acknowledgement**: In this section, provide acknowledgement to the funding source and other with identifiers and/or links as applicable.

About

The LICENSE.txt is a metadata file that contains the terms under which the dataset is shared. This file is intended to provide a human readable overview of requirements and conditions for reusing the data.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `LICENSE` and must be in the text format. The full name with extension must thus be `LICENSE.txt`.

Content

This file must contain the license terms.

About

The healthsheet.md metadata file is based on the healthsheet document established in 2021 to document motivation, composition, collection process, recommended uses, and so on of a healthcare dataset (see here for more information about healthseet). It is intended to improve communication between dataset creators and dataset consumers, and encourage dataset generators to prioritize transparency and accountability.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `healthsheet` and must be in the markdown format. The full name with extension must thus be `healthsheet.md`.

Content

The content of this metadata file must be exactly as specified in the paper "Healthsheet: Development of a Transparency Artifact for Health Datasets" that introduced the concept for this healthsheet.

About

The CHANGELOG.md is a metadata file that contains information about the changes between different versions of the dataset that are released. This file is intended to provide a human and machine-readable overview of different dataset versions that are released, their release date, and changes included between the different versions.

Specifications

Requirement

This metadata is mandatory for all datasets, including the first release.

Name and format

This metadata file must be named `CHANGELOG` and must be in the markdown format. The full name with extension must thus be `CHANGELOG.md`.

Content

The content must be structured following the Dataset changelog v1.0.0 conventions.

About

The study_description.json is a metadata file that contains provenance metadata, contextual metadata, as well as additional metadata about the study associated with the dataset. This metadata file is intended to prioritize machine readability.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `study_description` and must be in the JSON format. The full name with extension must thus be `study_description.json`.

Content

This metadata file must be structured as per the JSON schema provided here.

About

The dataset_description.json is a metadata file that contains provenance metadata, contextual metadata, as well as additional metadata necessary for reuse of the dataset. This metadata file is intended to prioritize machine readability.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `dataset_description` and must be in the JSON format. The full name with extension must thus be `dataset_description.json`.

Content

This metadata file must be structured as per the JSON schema provided here.

About

The dataset_structure_description.json is intended to document the structure of the dataset by specifying what each directory contains and what is. This metadata file is intended to prioritize machine readability.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

This metadata file must be named `dataset_structure_description` and must be in the JSON format. The full name with extension must thus be `dataset_structure_description.json`.

Content

This metadata file must be structured as per the JSON schema provided here.

About

The participants.json and participants.tsv files are metadata files that contain information about the participants in the study. The participant.tsv file contains the information while the participant.json file acts as a sidecar that describes the columns in the participant.tsv file.

Specifications

Requirement

This metadata is mandatory for all datasets.

Name and format

The `participants.json` file must be in JSON format and the `participants.tsv` must be in TSV format.

Content

participants.tsv The content of the `participants.tsv` file must be as follows:

- All column labels must follow the following naming convention: Only a-z characters (lowercase English alphabet) and 0-9 (numerical characters) are allowed with the exception of underscores that must be used to separate words (no white space allowed).
- The first column must be labeled `participant_id` and must list the IDs used in the study for all participants in the dataset.
- There must be one label for each datatype directory named exactly as as the corresponding directory. A boolean value (true/false) must be assigned for each participant based on if data from that datytype is included in the dataset for the participant or not.
- If participants are organized into multiple group/cohort, a column labeled `group` must be included with the group name specified for each participant
- Optionally include other columns as deemed adequate to facilitate data reuse such as `age`, etc.

participants.json The `participants.json` is intended to document the meaning and allowable values of each column in the the `participants.tsv` file. Its content must be as follows:

- There must be one key for each column label in the `participants.tsv` file except for the labels corresponding to the root-level datatype directories
- All sub-keys must follow the same naming convention as the column labels in the `participants.tsv` file
- A `description` sub-key must be included for each key to describe what the corresponding column represents

- A `data_type` sub-key must be included to indicate what type of data is associated with that column (e.g., `string`, `integer`, etc.)
- If there are a limited number of possible values for a given column, a sub-key called `levels` must be included to list and define the different possible values
- If a key value is expressed in a given unit, a sub-key named `unit` must be included to specify the unit
- Other sub-keys can be included as deemed necessary for understanding the content of the corresponding column.

Step-by-guide for implementing the CDS

Coming soon...

Templates

Coming soon...

Examples

Coming soon...

Overview

Coming soon...

All notable changes to the CDS will be documented in this file.

The format is based on Keep a Changelog and this project adheres to Semantic Versioning.

v.0.1.0 - 2024-xx-xx

Added

- First beta version released

Coming soon...

Providing feedback/suggestions

Have feedback, suggestions, or questions related to the CDS? Submit them by opening a GitHub issue. If you want to suggest changes to this docs, you can also submit a PR by following the instructions in the developer instructions below.

Developer instructions

Instructions for making changes to the documentation, submitting a PR, and publishing a new version of the documentation and guidelines are available here.