



eBRAIN-Health

Public report

D4.2 - Open access data catalogue and variable mapping

Project number	101058516
Project title	eBRAIN-Health - Actionable Multilevel Health Data
Submission date	July 2024
Authors	Dr. Alpha Tom Kodamullil, Sathvik Guru Rao (FRAUNHOFER)
Dissemination level	Public (PU)
Public project website	https://ebrain-health.eu/



Funded by
the European Union

Table of content

1. eBRAIN-Health	3
2. eBRAIN-Health consortium	3
3. Introduction	4
4. Partners involved	4
5. Description of work performed	4
5.1. Variable Mapping with Common Data Model & Data Stewardship Tool	5
5.1.1. Common Data Model (CDM)	5
5.1.2. Data Stewardship tool (DST)	5
5.2. Data Catalogue and Data Viewer	7
6. Results.....	8
6.1. Variable Mapping with Common Data Model	8
6.2. Data Steward Tool for data standardization	9
6.3. eBRAIN-Health Data Catalogue.....	10

1. eBRAIN-Health

The Project eBRAIN-Health will deliver a distributed research platform for modeling and simulating complex neurobiological phenomena of human brain function and dysfunction in a data protection compliant environment. It will provide thousands of multilevel virtual brains from patients and healthy human controls for research and innovation. Brain data from multiple sources will be pre-processed. Solving the societal grand challenge of dementia is a big task. Yet it appears feasible in a collective approach. Therefore, we will build an interdisciplinary digital twin for dementia for modeling and simulating complex phenomena at the service of research infrastructure communities. eBRAIN-Health-Cloud will offer end-to-end services for personalized complex brain modeling and simulations in distributed e-infrastructures with data protection by design and by default and simulation-ready human multiscale brain data that range from molecular (genomics, proteomics, metabolomics) and cellular to electrophysiology and imaging to behavioral, clinical, lifestyle and environmental data as well as data from wearables. Brain data are pre-processed and annotated such that they all relate to a common reference 3D brain space.

eBRAIN-Health-Cloud constitutes a blend of three large-scale research programs: the FET Flagship Human Brain Project with its EBRAINS Research Infrastructure, the EOSC project Virtual Brain Cloud with its Virtual Research Environment for sensitive data and the H2020 project AI-MIND with intelligent tools for dementia risk estimation. The project will have synergies to topics of the Digital Europe Program, such as artificial intelligence, cybersecurity and supercomputing and the Health Data Space. eBRAIN-Health-Cloud offers a next generation clinical research infrastructure and creates an open yet protected space for groundbreaking digital health innovation by the research infrastructure communities comprising academia and the private sector.

2. eBRAIN-Health consortium

- CHARITE – Universitaetsmedizin Berlin, Germany
- EBRAINS, Belgium
- Forschungszentrum Juelich GmbH, Germany
- Stichting Radboud Universiteit, Netherlands
- Universidad Pompeu Fabra, Spain
- OSLO Universitetssykehus, Norway
- tp21 GMBH, Germany
- Fraunhofer Gesellschaft zur Foerderung der Angewandten Forschung eV, Germany
- INDOC RESEARCH EUROPE gGmbH, Germany
- Universitaet Wien, Austria
- Universidad Complutense de Madrid, Spain
- EODYNE Systems SL, Spain
- ATHENA – Research and Innovation Center, Greece
- University of Oslo, Norway
- Universita degli Studi di Roma la Sapienza, Italy
- Alzheimer Europe, Luxembourg
- Institute National de Recherche en Informatique et Automatique, France
- Centre Hospitalier Universitaire Vaudois, Switzerland
- The University of Edinburgh, United Kingdom

[Find the partners on our website](#)

3. Introduction

Data from Alzheimer's Disease (AD) cohorts are fragmented, inherently unstructured, noisy, and often incomplete, with large amounts of non-standardized terms/concepts. We address this issue by harmonizing and optimizing clinical data processing and alignment using sophisticated tools for natural language processing and annotation. As a first step, we have created and organized the necessary ontologies and terminologies needed for the eBRAIN-Health within Task 4.1. These ontologies and terminologies are then stacked into a semantic framework (eBRAIN-Health Semantic Framework – referred to in Deliverable 4.1), which acts as the basis for the semantic interoperability of the datasets. Metadata annotation is standardized by mapping and harmonizing various data points from various datasets to the developed controlled vocabularies and ontologies via eBRAIN-Health Semantic Framework by using it as a centralised ontology store. Based on the eBRAIN-Health semantic framework, we have built the Alzheimer's Disease core data model for dementia and extended the variable mappings using the Data Steward tool developed at SCAI. The Data Steward Tool (DST) allows semi-automatic semantic integration of clinical data into ontologies and global data models and data standards. We have aligned and mapped data dictionaries from datasets available in eBRAIN-Health with the available mapping in ADataViewer. ADataViewer lets you explore the AD data landscape and identify cohort datasets that suit your research needs.

4. Partners involved

Partners involved: FRAUNHOFER (lead), CHARITE, UNIRM1 (for EEG/MEG case), UIO

5. Description of work performed

Overview

In this deliverable, the first task was to collect various AD-related datasets from project partners and publicly available AD cohort datasets for building an openly accessible data catalogue. The second task was to build a common data model for AD, which acts as a semantic data model that provides a standardized framework for variable mapping, data harmonization, and interoperability among datasets.



Figure 1: The overall workflow of the tools developed and connected for the variable mapping, cataloguing and cohort mapping and finally the visualization.

5.1. Variable Mapping with Common Data Model & Data Stewardship Tool

5.1.1. Common Data Model (CDM)

In the context of eBRAIN-Health project, we have built and extended the existing common data model on Alzheimer's Disease. Fraunhofer SCAI has built a common data model using the variables coming from 38 different cohort sources from various neurodegenerative diseases (AD, Ataxia etc.). Here is the list of various data resources used for building the data model: 'DESCRIBE', 'DELCODE', 'SCAregistry', 'Clinical Research Consortium for the Study of ' 'Prospective Study of Individuals at Risk for Spinocerebellar Ataxia', 'Textmining', 'AddNeuroMed - Merge Cerebellar AtaxiaDataset', 'Alzheimer's Disease Neuroimaging Initiative - Phase 1', 'Alzheimer's Disease Neuroimaging Initiative - Phase 2', 'Alzheimer's Disease Neuroimaging Initiative - Phase 3', 'Alzheimer's Disease Neuroimaging Initiative - MERGE', 'Alzheimer's Disease Neuroimaging Initiative - GO Phase', 'Memory Clinic UKB', 'German Center for Neurodegeneration Research', 'European Spinocerebellar Ataxia Registry', 'European Spinocerebellar Ataxia Type 3/Machado-Joseph Disease Initiative', 'JADNI', 'EPAD', 'ABVIB', 'IGNORE', 'ANM', 'PharmaCog', 'ADNI', 'ARWIBO', 'I-ADNI', 'DOD-ADNI', 'WMH-AD', 'ROSMAP', 'AIBL', 'NACC', 'PREVENT-AD', 'IGNORE.1', 'A4', 'OASIS', 'VASCULAR', 'EMIF', 'EDSD', 'VITA'.

This common data model consists of 1539 unique data variables and 2995 Variable mappings which are good enough to map and standardize any new incoming data sets in NDD field.

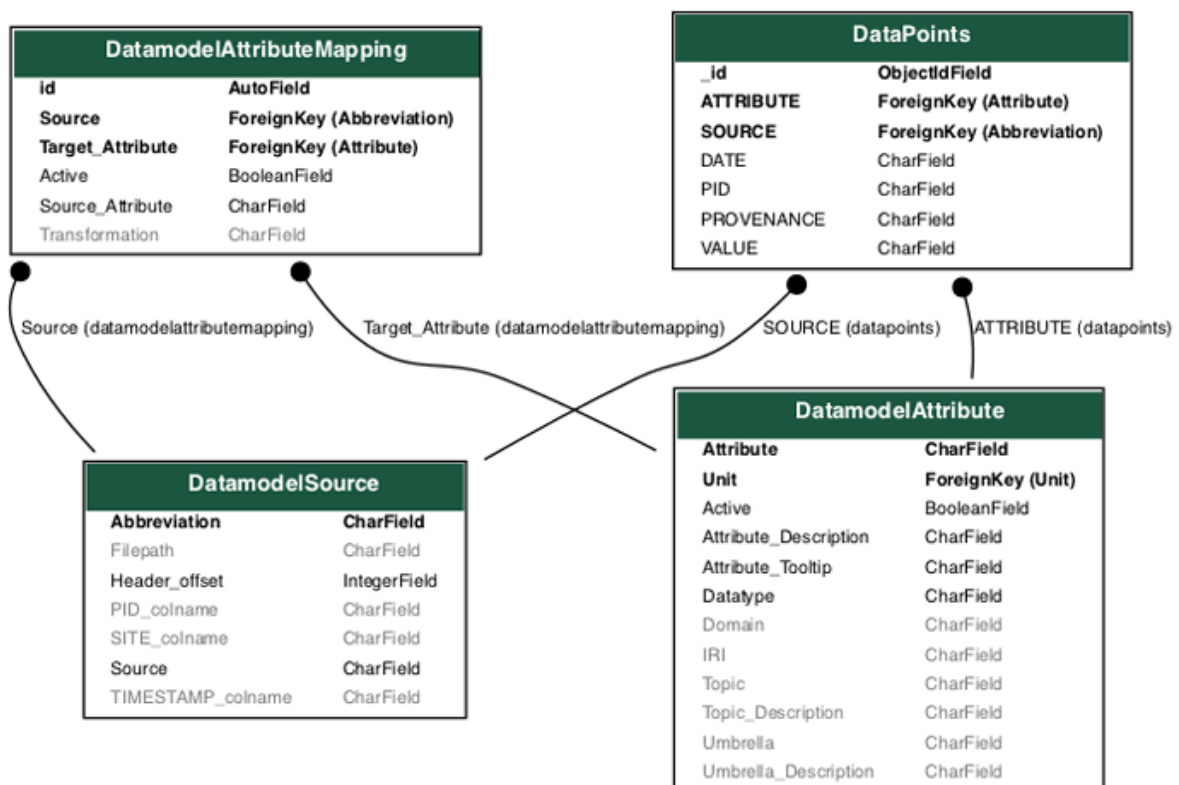


Figure 2: UML Diagram of CDM on Database level

5.1.2. Data Stewardship tool (DST)

The Data Steward Tool (DST) which can be used to automatically standardize clinical datasets, map them to established ontologies, align them with OMOP standards (a widely used data model for health data), and export them to a FHIR-based format. To standardize specific variables from dementia which could not be mapped to existing ontologies or data standards, we developed a CDM in the context of dementia, built from various datasets such as —Alzheimer's disease datasets [ADNI and AddNeuroMed], routine clinical datasets (from University Hospitals Bonn and Aachen, Germany) and

other datasets covering neurodegenerative diseases from DZNE (<https://www.dzne.de>) and EUROSCA (<http://www.euroscas.org>).

We implemented a web application called the DST that allows users to read, edit and use the CDM to standardize real-world research data. The tool consists of RESTful APIs served by a Django application on a MongoDB database; thus, providing a user-friendly environment that makes the underlying data accessible to the scientific community. To demonstrate the DST, we accessed and investigated major dementia studies and identified the variables they shared. We found that most of these variables comprised patient information, such as age, sex or initial diagnosis, biomarker measurements, as well as measurements from neuropsychological tests, such as the Boston Naming Test. This set of variables was then used as the groundwork to extend the data model as we gained access to new studies by connecting the variables of additional studies to equivalent variables present in the CDM. Finally, we extracted data types and value ranges for each of the variables and assigned every variable to a specific data modality.

data-steward.bio.scai.fraunhofer.de/data-steward

Download Screens



Data Steward

Welcome to the SCAI Data Steward

This service provides several features such as uploading and editing the current clinical datamodel, which is used right now to normalize data in the used in the [Clinical Viewer](#). Moreover you can download the datamodel in .xlsx format after you added attributes, mappings or new sources.



Another key feature of the Data-Steward is the graph explorer of the datamodel. That graph provides deeper insight into the model currently used by SCAI. Edges in that graph represent the connections between different sources and its attributes as well as the mappings between those attributes.

Figure 3: Landing page of the Data Stewardship Tool, where you can upload your datasets and can do a semi-automatic mapping of datasets.

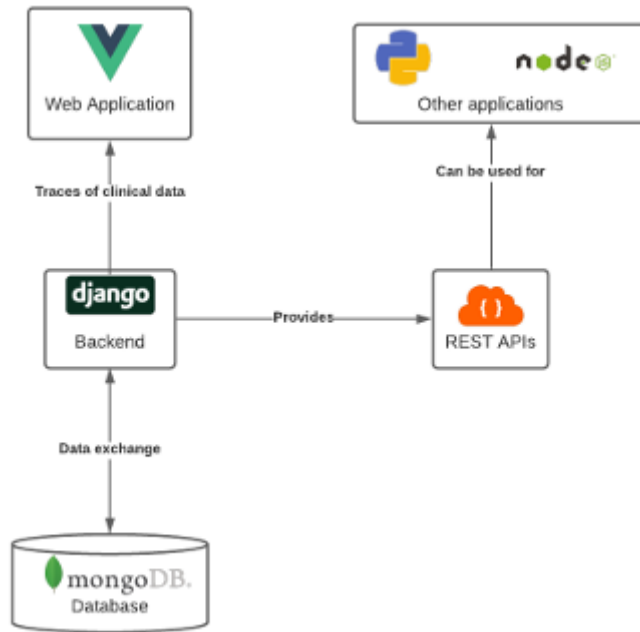
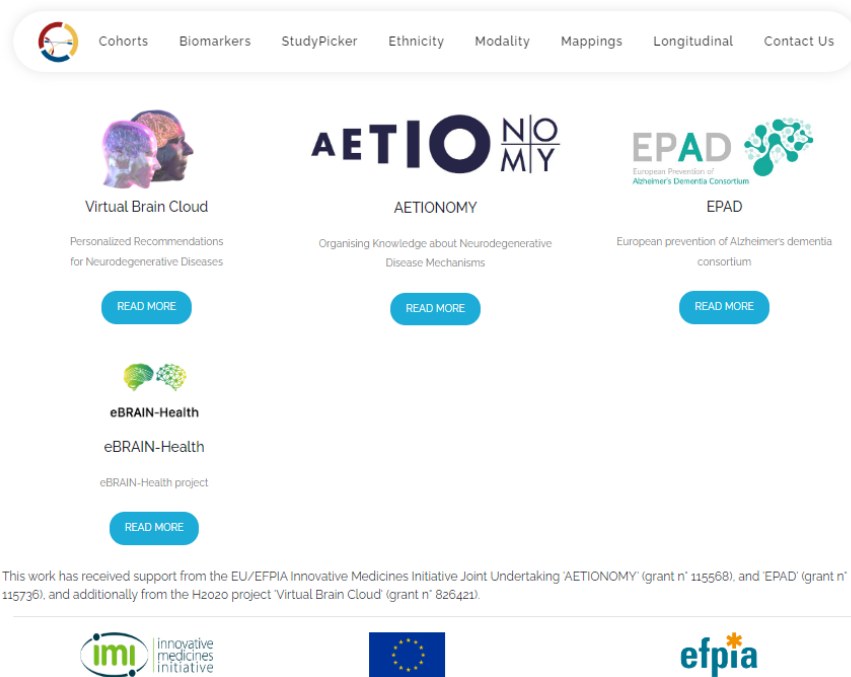


Figure 4: Interactions of the different Components of the DST.

5.2. Data Catalogue and Data Viewer

Fraunhofer extended the already built ADataViewer with the new datasets coming from the eBRAIN-Health for building the data catalogues. An interactive platform that facilitates the exploration of 20 cohort datasets with respect to longitudinal follow-up, demographics, ethnoracial diversity, measured modalities, and statistical properties of individual variables.



The screenshot shows a navigation bar with links: Cohorts, Biomarkers, StudyPicker, Ethnicity, Modality, Mappings, Longitudinal, and Contact Us. Below are four project cards:

- Virtual Brain Cloud**: Personalized Recommendations for Neurodegenerative Diseases. [READ MORE](#)
- AETIONOMY**: Organising Knowledge about Neurodegenerative Disease Mechanisms. [READ MORE](#)
- EPAD**: European prevention of Alzheimer's dementia consortium. [READ MORE](#)
- eBRAIN-Health**: eBRAIN-Health project. [READ MORE](#)

At the bottom, there is a funding notice: "This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking 'AETIONOMY' (grant n° 115568), and 'EPAD' (grant n° 115736), and additionally from the H2020 project 'Virtual Brain Cloud' (grant n° 826421)." Logos for imi, the European Union, and efpia are displayed.

Figure 5: The lists of projects in Alzheimer's disease that use the ADataViewer, variable mapping, and data catalogue.

Feature Mappings Across Cohorts

The prerequisite for working across multiple datasets is interoperability. This aspects includes the availability of variables, same naming conventions, and comparable representations. While value representations can be explore in the Biomarker section, here we focus on naming conventions and semantic mappings of variables across cohorts.

Mappings to standardized ontologies can be found in the table below. Variable names printed in **red** describe variables which were listed in the dataset and its metadata but held only missing values across all participants of the respective dataset (i.e. no data was available). **Disclaimer:** While comprehensive, this work is not a complete mapping of every single variable available in these cohorts.

Clinical (I)		Clinical (II)		Demographics		Family		Comorbidities	
APOE		CSF		Plasma		Lifestyle		PET	
MRI General Brain	MRI Diencephalus	MRI Basal Ganglia	MRI Cerebellum	MRI Ventricles	MRI Brain Poles Volume	MRI Hippocampus	MRI Others		

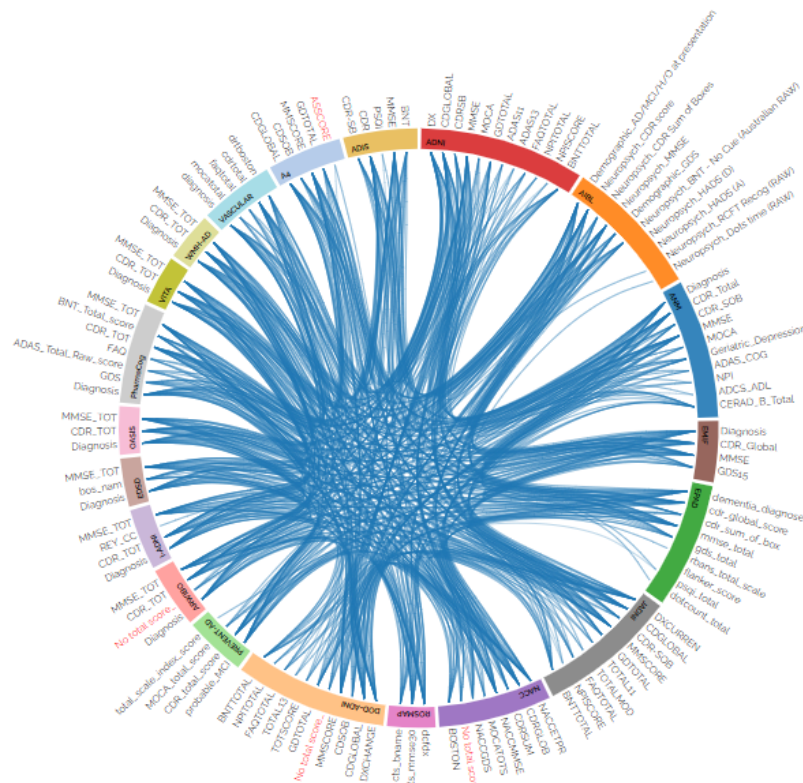


Figure 6: Feature mappings of various datasets including the ones from eBRAIN-Health visualized in the ADataViewer.

6. Results

6.1. Variable Mapping with Common Data Model

The CDM (Common Data Model) consists of common variables representing cohort demographic information, various clinical assessments, and further dementia-related biomarkers. The CDM was designed to store metadata about the variables themselves, including definitions, data type, and value ranges as well as information about the variable names already mapped onto it. The CDM can constantly and rapidly expand as more studies are mapped to it. By enriching the variables present in the data model with mappings to other resources, the system is capable of standardizing data from different origins. Finally, while the data model is stored in a database, it can be exported to OWL (Web

Ontology Language), RDF (Resource Description Framework) as well as tabular formats such as excel and csv.

data-steward.bio.scai.fraunhofer.de/data-steward/table

Model Upload Wizard Model Explorer Clinical Viewer Download Contact

Datamodel Mappings

[Download as .csv](#)

Source ↑	External Variable	Internal Variable	Description (Internal Variable)	Transformation
AddNeuroMed - Merge Dataset	Age	AGE_FV	Age at first visit	
AddNeuroMed - Merge Dataset	Sex	SEX	Sex	cm001
AddNeuroMed - Merge Dataset	MMSE	MMSE_SUM_R	Mini Mental State Examination (raw v.	
AddNeuroMed - Merge Dataset	CDR_SOB	CDR_SOB	Clinical Dementia Rating Scale Sum	
AddNeuroMed - Merge Dataset	CDR_Total	CDR_SUM	Global Clinical Dementia Rating Scz	
AddNeuroMed - Merge Dataset	ADAS_COG	ADAS_COC_11	Alzheimer's Disease Assessment Sc	
AddNeuroMed - Merge Dataset	MOCA	MOCA	Montreal Cognitive Assessment	
AddNeuroMed - Merge Dataset	Geriatric_Depression	GDS_15	Geriatric Depression Scale (15 item	
AddNeuroMed - Merge Dataset	NPI	NPI_TOTAL	Neuropsychiatric Inventory - Total (
AddNeuroMed - Merge Dataset	CERAD_A_Total	VFC_ANIM_R	Verbal fluency cat. (animals) (raw v.	

Figure 7: Variable Mapping Catalogue behind the Data Steward Tool

6.2. Data Steward Tool for data standardization

We developed a web application called the DST that provides an interface to visualize the model, extend it with new variables, add mappings, and read clinical data through a user-friendly interface. Furthermore, the DST is capable of automatically mapping external variables onto the CDM through fuzzy string matching. Performing full-text searches throughout the entire data model grants the possibility of using the system as a searchable variable catalog for dementia research. Additionally, uploading clinical data onto the system allows for data standardization and the storage of harmonized data in a central data repository. The uploaded standardized data can be queried via a RESTful API. Through these APIs, users can access the content of the DST using their own tools (e.g. Jupyter notebook). Uploading and standardizing data is a guided multi-step process, where the user begins by loading the data, often a 2D data table, to an interface via drag and drop and is then given the opportunity to manually map variables that were not found in the current data model. During this process, the user is supported with autosuggestions to automatically add terms from eBRAIN-Health Semantic Framework (Deliverable 4.1) if no suitable variable can be found. Finally, the DST provides a graph-based view of the model where the user can interactively explore the entirety of the model (Please refer: <https://data-steward.bio.scai.fraunhofer.de/data-steward/graph>).

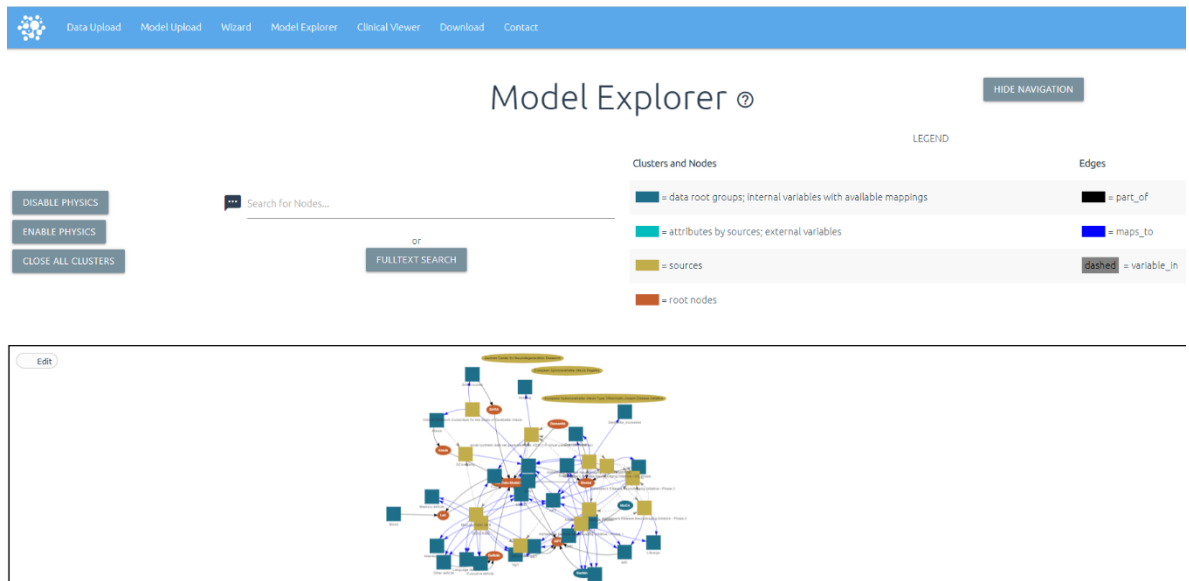


Figure 8: A graphical view of the common data model which act as basis for the Data Stewardship Tool.

6.3. eBRAIN-Health Data Catalogue

From the eBRAIN-Health context, we have the following datasets: OUH (AI-MIND), UEDI (Brain Health Scotland, EPAD, PREVENT), VUmc (EMIF, EPND), UCM (HCP-COBRA, Madrid), UNIRM1 (Rome), CHUV (MIP). External collaborators - provide access to additional cohorts: OBI (CBN01, ONDRI, ONDRI@HOME, ReMINDD), Oxford University (Dementias Platform UK), where we have already done variable mapping for EPAD, PREVENT, EMIF and Dementias Platform UK. We still need to get the data points from missing datasets and need to map the variables to the existing data model and if needed, extend the model.

The complete catalogue of AD datasets can be accessed at: <https://adata.scai.fraunhofer.de/cohorts>

● ABVIB	280	not reported	not reported	not reported	227	12	USA	not reported	DOI
● ADIS	75 (prospective)	25 (prospective)	25 (prospective)	25 (prospective)	0	no follow-up	Spain	Amyloid positive	DOI
● ADNI	2249	813	1016	389	1978	6	USA, Canada	NINCDS-ADRDA	DOI
● AIBL	1378	803	134	181	1019	18	Australia	NINCDS-ADRDA	DOI
● ANM	1702	793	397	512	1254	12	Europe	NINCDS-ADRDA	DOI
● ARWIBO	2617	1476	208	281	80	12	Italy	NINCDS-ADRDA	DOI
● DOD-ADNI	458	181	27	0	458	12	USA, Canada	SFVAMC	DOI
● EDSO	474	183	140	151	0	no follow-up	Europe	NINCDS-ADRCA	DOI
● EMIF	1221	386	526	201	0	no follow-up	Europe	NINCDS-ADRDA	DOI
● EPAD	2096	2071	not reported	14	1596	12	Europe	NINCDS-ADRDA	DOI
● I-ADNI	262	2	54	187	0	no follow-up	Italy	NINCDS-ADRDA	DOI
● JADNI	537	151	233	149	518	6	Japan	NINCDS-ADRDA	DOI
● NACC	40858	15894	3649	11761	27657	12	USA	UDS Form D1	DOI
● OASIS	564	400	122	42	168	6	USA	CDR	DOI

Figure 9: The data catalogue for eBRAINS-Health

Disclaimer

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058516. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or other granting authorities. Neither the European Union nor other granting authorities can be held responsible for them.