



eBRAIN-Health

Public report

D4.1 - Open access full metadata and ontology models

Project number	101058516
Project title	eBRAIN-Health - Actionable Multilevel Health Data
Submission date	July 2024
Authors	Dr. Alpha Tom Kodamullil (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.	Building of Ontologies.....	4
4.1.1	Excerpt of ontology curation guideline for building ontologies.....	4
4.1.2	Building of ontologies	5
5.2.	Implementation of semantic framework	5
6.	Results.....	5
6.1.	Alzheimer's Disease Ontology.....	6
6.2.	Pathway Ontology (PO).....	9
6.3.	Neuroimaging Feature Ontology (NIFT).....	10
6.4.	Brain Region and Cell Type Ontology (BRCO)	11
6.5.	EEG/MEG and Feature Terminology	12
6.6.	Semantic Framework & Management system.....	14
7.	Conclusion, next steps	16

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 behavioural, clinical, life-style 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

Work Package 4 focuses on semantic data integration and multimodal knowledge integration. In this Work Package we ensure the fulfillment of the FAIR data principles in all our applications, continuously update existing resources via dedicated data steward and curation pipelines, the creation and provision of a semantic framework for neurodegenerative diseases in the form of central resources for controlled vocabularies and shared ontologies. Here, in this report we address task 4.1 ("Semantic framework for unified metadata annotation" with the deliverables D4.1 "Open access full metadata and ontology models"). The main objectives of this deliverable are to build and extend essential ontologies in the neurodegenerative field to extract knowledge hidden in unstructured text or scientific publications.

In the context of this task and deliverable, FRAUNHOFER worked on updating already published ontologies and is developing new ontologies for the eBRAIN-Health project. The already existing ontologies, specifically Alzheimer's Disease Ontology, Pathway Terminology System, Neuroimaging Feature Ontology, and Brain Region and Cell Type Ontology have been successfully updated. The updated version of the Alzheimer's Disease Ontology has been published and is now part of the OBO Foundry. FRAUNHOFER has built a specific instance of 'Ontology Lookup Service' (OLS) that is developed by the European Bioinformatics Institute (EBI) for exploring ontologies, acting as the semantic layer for the entire consortium. OLS acts as a repository for biomedical resources that aims to provide a single point of access to the latest ontology and terminology versions developed in this project. This semantic framework will also enable mapping and importing concepts from various ontologies as well as the meta-data framework for unified metadata annotation. This version of OLS is used as the semantic layer for annotating the relevant entities for retrieving the specific literature for extracting the cause-effect mechanisms as well as information about pathways co-associated with brain regions

With the deliverables here the demonstrators of the eBRAIN-Health Semantic Framework website are made available at the following address:

OLS, the ontology lookup service software to host the eBRAIN-Health ontologies, which is accessible at <https://ols.ebrain.bio.scai.fraunhofer.de/index>

Fraunhofer SCAI is responsible for the continuous maintenance of the OLS instance and its content.

4. Partners involved

FRAUNHOFER (lead), CHARITE, UNIRM1, UIO

5. Description of work performed

As a part of the ontology development and semantic data integration, various ontologies are being developed with the lead from Fraunhofer SCAI. There are disease-specific ontologies like Alzheimer's Disease Ontology and general ontologies like Neuroimaging Feature Ontology. Below are the detailed descriptions of work done to build and update the ontologies.

5.1. Building of Ontologies

4.1.1 Excerpt of ontology curation guideline for building ontologies

The scientific community defines several standards and best practices to create and develop ontologies. Open Biological and Biomedical Ontology (OBO) Foundry has defined in the past principles to “develop interoperable ontologies, that are both logically well-formed and scientifically accurate”. Hence, where possible, we have used the best practices. That includes the usage of terms that were already described in an existing OBO conform ontology. Furthermore, to make the curation reproducible and transparent, several annotations have been added to the ontologies. For example:

- `rdfs:label` annotation was added for each concept as exactly same as the concept name.
- `oboInOwl:hasDefinition` property contains the definitions for a concept, which is mandatory for all concepts.
- `rdfs:isDefinedBy` contains a reference to the source of the definition, if the definition couldn't be imported from an existing ontology.
- `importedFrom` annotation is used if a concept is reused from another ontology.
- `rdfs:seeAlso` is used to capture any additional relevant references.
- `oboInOwl:hasExactSynonym` includes exact synonyms. Source of the synonyms are the terms from articles or research papers.
- `oboInOwl:hasRelatedSynonym` includes related synonyms.
- `oboInOwl:hasDbXRef` is used to add additional link from PubMed/NCBI.

The curation guidelines contain several further rules that ontology experts follow during the curation process.

The curation guideline is also subject to continuous review. If regulations are adapted, this will be taken into account in our guidelines.

4.1.2 Building of ontologies

All ontologies were developed using protégé 5.5.0 (<https://protege.stanford.edu/products.php>) and whenever possible all ontologies followed Basic Formal Ontology as the defined hierarchy as ontology structure. All concepts that are defined in an existing OBO ontology are integrated by using OntoFox. All newly added entities (manual) have the IRI starting with <https://bio.scai.fraunhofer.de/ontology/BRCO>. For BFO higher order classes, the BFO IDs are used as IRI. For ontofox imports, IRI is the original OBO ID. The version control of the project is maintained using SCAI Gitlab.

5.2. Implementation of semantic framework

All the ontologies developed have been integrated into the semantic framework and ontology management system which has been developed leveraging the Ontology Lookup Service (OLS) instance. This serves as a dedicated repository for housing neurodegenerative diseases-related ontologies, providing a central platform for accessing, and querying semantic resources. Currently the semantic framework includes Alzheimer's Disease Ontology, Pathway ontology, Brain region and Cells ontology, Neuroimaging Feature ontology, and Human Physiology Simulation Ontology.

6. Results

T4.1 Semantic framework for unified metadata annotation (PM24)

Fraunhofer lead the development of semantic framework for the eBRAIN-Health project.

Here is the list of ontologies developed in the project:

6.1. Alzheimer's Disease Ontology

Fraunhofer SCAI have constructed and integrated Alzheimer's Disease Ontology (ADO), combining selected concepts from the former version of the ADO and the Alzheimer's Disease Mapping Ontology (ADMO). In addition to the existing entities available from these knowledge models, essential knowledge about AD from public sources, such as newly discovered risk factor genes and novel treatments, was also integrated. The updated version of ADO (<http://purl.obolibrary.org/obo/ado.owl>, version 2.0.1) has been constructed and is ready for application. In summary, with a total count of 39 854 axioms, 1963 classes (1910 classes imported from other OBO ontologies and 53 original class axioms) and 12 object properties, the updated ADO contains relevant knowledge ranging from research, preclinical, clinical and molecular interactions and mechanisms to diagnostics, study types and treatments.

The Alzheimer's Disease Ontology has been published in peer-reviewed journal Databases:

Zhang B, Lage-Rupprecht V, Wegner P, Sargsyan A, Gebel S, Jacobs M, Klein J, Hofmann-Apitius M, Tom Kodamullil A. Design of the formalized and integrated Alzheimer's Disease Ontology and its application in retrieving textual data via text mining. Database (Oxford). 2023 Dec 2;2023:baad085. doi: 10.1093/database/baad085. PMID: 38041858; PMCID: PMC10693436.

Alzheimer's Disease Ontology (ADO) is part of Open Biological and Biomedical Ontologies (OBO). <https://obofoundry.org/ontology/ado.html>

ADO ontology could be used in many application scenarios. Here we demonstrate how ADO assists researchers in extracting textual information via text mining. The ontology model was integrated into the semantic search engine Academia SCAIView (<https://academia.scaiview.com>) and could be selected as taggers to annotate texts. In total, the search engine accesses more than 36 million documents from the MEDLINE database, of which more than 18 million are annotated. Using the ADO ontology as a search filter results in a total of 30 021 056 documents retrieved [many documents were retrieved using ADO ontology as a tagger due to the wide range of concepts (and their synonyms) included in the upper and intermediate levels of ADO ontology]. Other sub-ontologies derived from ADO, such as 'signs and symptoms', 'genetic risk factors', and 'treatment', can also be used as taggers. We give several examples of how the taggers can be utilized as follows:

Publication filtering based on individual taggers

The integrated ADO-based models can be incorporated as bins tagging the abstracts of the documents that match the context. When we applied the whole ADO as the text tagger, a number of documents were retrieved, shown in Figure 1. While applying the other taggers, different numbers of documents were obtained (86 877 documents for the genetic risk factor tagger, 4 985 933 documents for the signs and symptoms tagger and 8 907 809 documents for the treatment tagger).

A

Table		Barchart		
Add to Query	Label	Relevance	Count of Documents with Concept in Query	Total documents
<input type="checkbox"/>	memory	1.25	7,374	773,664
<input type="checkbox"/>	inflammatory response	0.87	6,720	1,430,246
<input type="checkbox"/>	cognition	0.84	4,418	322,749
<input type="checkbox"/>	signaling	0.42	4,317	1,619,475
<input type="checkbox"/>	developmental process	0.40	6,793	5,125,866
<input type="checkbox"/>	activated T cell autonomous cell death	0.38	3,789	1,334,151
<input type="checkbox"/>	phosphorylation	0.37	2,967	682,161
<input type="checkbox"/>	metabolic process	0.35	3,394	1,179,529
<input type="checkbox"/>	biological regulation	0.34	4,209	2,161,656
<input type="checkbox"/>	biosynthetic process	0.32	5,121	3,555,230

Previous Page 1 of 379 10 rows Next

B

Table		Barchart		
Add to Query	Label	Relevance	Count of Documents with Concept in Query	Total documents
<input type="checkbox"/>	inflammatory response	0.55	1,371,290	1,430,246
<input type="checkbox"/>	developmental process	0.36	1,810,114	5,125,866
<input type="checkbox"/>	necrotic cell death	0.31	755,431	757,868
<input type="checkbox"/>	biological regulation	0.20	872,650	2,161,656
<input type="checkbox"/>	signaling	0.18	733,545	1,619,475
<input type="checkbox"/>	activated T cell autonomous cell death	0.18	664,926	1,334,151
<input type="checkbox"/>	behavior	0.17	868,711	2,436,243
<input type="checkbox"/>	biosynthetic process	0.17	1,038,584	3,555,230
<input type="checkbox"/>	metabolic process	0.12	510,623	1,179,529
<input type="checkbox"/>	gene expression	0.12	517,886	1,294,911

Previous Page 1 of 1626 10 rows Next

Figure 2: (A) Cross-analysis of AD-related symptoms and cellular processes. (B) Cross-analysis of AD-related symptoms and cellular processes.

Mentions of drugs in the context of AD-related signs and symptoms

We applied the 'signs and symptoms' filter and analyzed the resulting corpus concerning DrugBank annotations. The three most mentioned drug compounds in our search context were methamidophos, bifenthrin and flufenoxuron.

Biological processes/pathways that might play a role in AD

As the 'genetic risk factor' tagger was applied in SCAIView, the corpus was then analyzed for relevance in 'gene ontology biological processes' annotations. The analyzed corpus was filtered using 'pathway' and its synonyms. We added each pathway term to the query and noted the number of literature studies extracted. The Wnt signaling pathway and Notch signaling pathway are the most intensively researched pathways concerning the genetic risk factors found in AD patients.

6.2. Pathway Ontology (PO)

The Pathway Ontology (PO) consists of 4483 classes, organized in a hierarchical structure. Each class has been annotated, whenever available, with additional information such as definition, source identifiers, and synonyms. PO was created by integrating terms from the Integrating Network Objects with Hierarchies (INOH) and Pathway Ontology (PW) and was populated with pathways from four popular public biomedical pathway databases, namely KEGG, Reactome, BioCarta, and Pathway Interaction Database. Further enrichment of the PO was achieved by analysis of relevant scientific text. For this purpose, the phrases (with frequency of occurrence) from MEDLINE abstracts containing the word “pathway” and three words preceding it were extracted using a sub-corpus. If the pathway name inside the four-word phrase was either absent from the pathway reference name list or absent from the pathway synonym list, it was added to the pathway reference name dictionary.

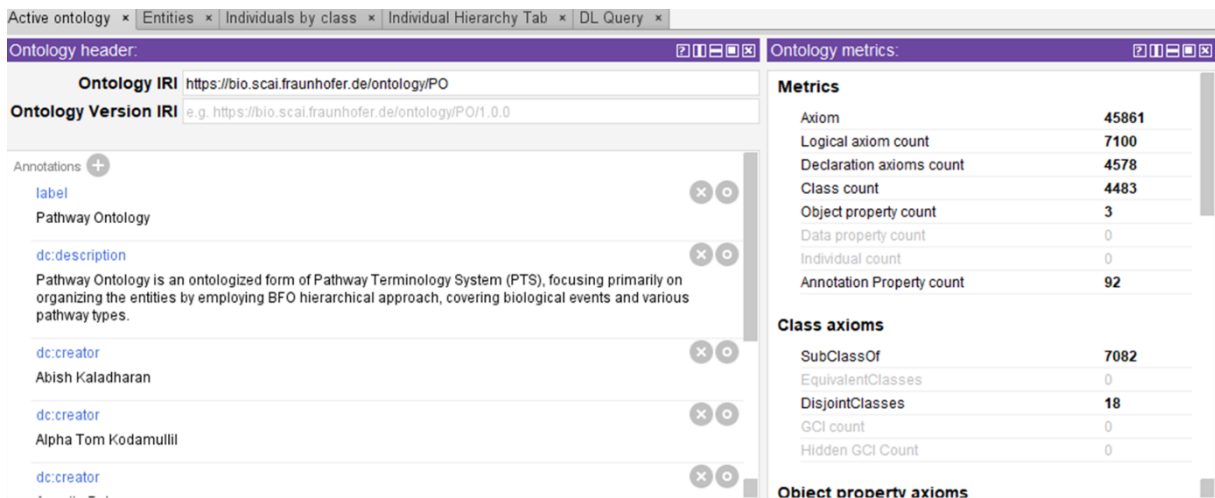


Figure 3: The overall statistics of Pathway Ontology

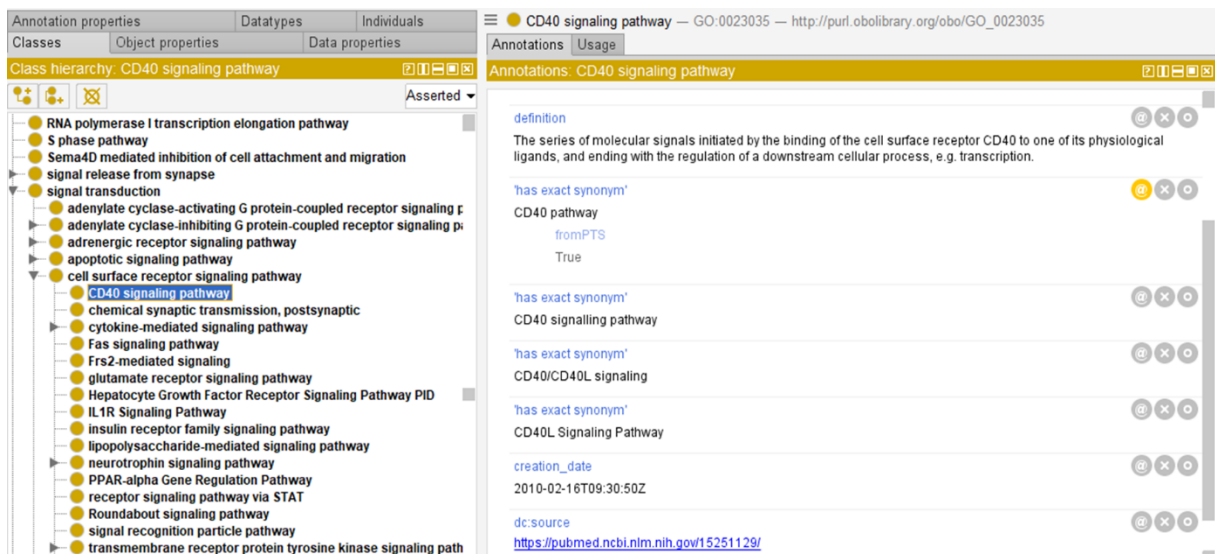


Figure 4: The screenshot of the pathway ontology with various concepts along with annotations

Integrating the PO into the literature-mining environment such as SCAIView (the environment that enabled us to perform very context-specific literature searches based on combined semantics from

multiple ontologies and terminologies), allows us to link pathways with other interested biological concepts such as imaging features. For example, we would be able to query the literature for cellular pathways relevant to the brain anatomy in both healthy and AD conditions.

6.3. Neuroimaging Feature Ontology (NIFT)

Within the Neuroimaging Feature Ontology (NIFO), published in 2017, there are in total of 1,221 terms. In the context of eBRAIN-Health NIFT was updated and currently it has 1126 concepts. NIFT can be used to correlate clinical diagnosis with imaging features, retrieval and mining figure captions and full text from publications, and annotation of image scans.

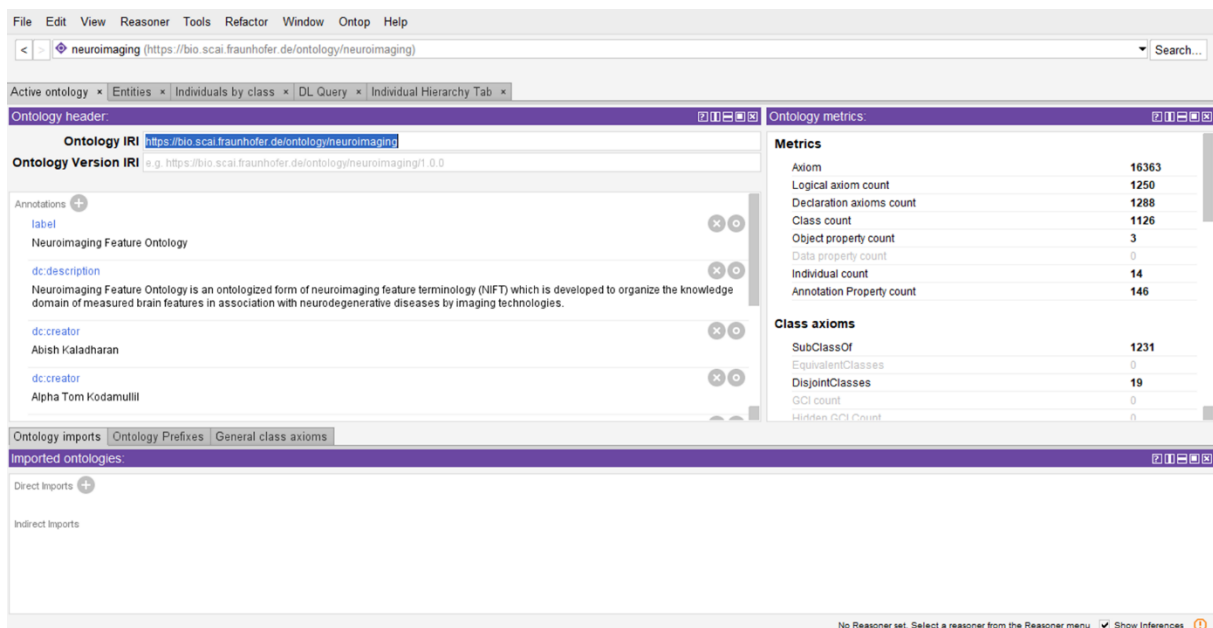
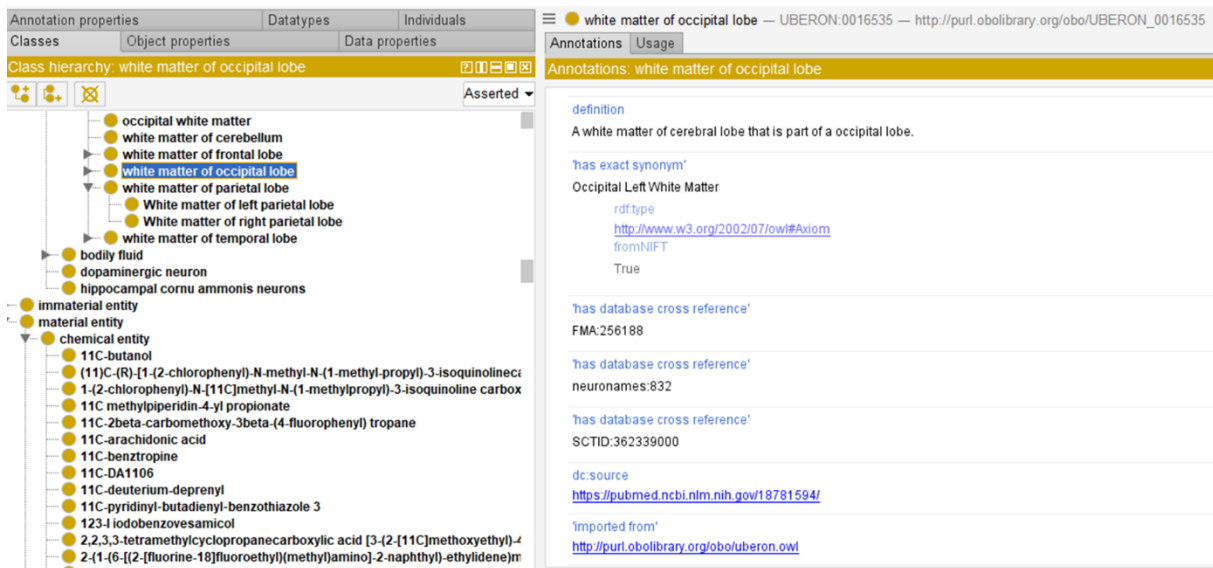


Figure 5: The overall statistics of NIFO

Below is a figure showing the hierarchical structure of NIFO as visualized in the Protégé OWL Editor. This figure depicts the higher level concepts the terminology namely Algorithms, Brain Region, Clinical Indices, Clinical trial information, Imaging Technique, Measured Feature, and Radiopharmaceutical compound:

Apart from the 7 major classes of terms from NIFO, other major classes of entities were added:

- Imaging parameter (pulse sequence, echo time, repetition time etc.)
- Imaging data processing (normalization, smoothing etc.)
- Behavioral data (reaction time, motor performance etc.)
- Statistical analysis (Pearson correlation, post-hoc etc.)
- Units of measure (time unit, magnetic strength etc.)
- Image artifacts (metal artifact, noise artifact etc.)
- Types of imaging (fMRI, gradient echo MRI etc.)
- Image file format (DICOM, JPG, PNG etc.)
- Computer language (python, R etc.)



The screenshot displays the NISO NIF Ontology (NIFO) interface. On the left, a class hierarchy is shown for 'white matter of occipital lobe', including subclasses like 'white matter of cerebellum', 'white matter of frontal lobe', and 'white matter of parietal lobe'. On the right, the 'Annotations' tab for 'white matter of occipital lobe' (UBERON:0016535) is active, showing a definition: 'A white matter of cerebral lobe that is part of a occipital lobe.' It also lists various database cross-references such as FMA:256188, neuronames:832, and SCTID:362339000, along with the source URL: <https://pubmed.ncbi.nlm.nih.gov/18781594/>.

Figure 6: The screenshot of the NIFO with various concepts along with annotations

6.4. Brain Region and Cell Type Ontology (BRCO)

BRCO stands for Brain Region and Cell Type Ontology, which was built with the purpose of organizing the knowledge domain of brain anatomy with a top-down granularity, from gross regions to cell types. BRCT describes the types of both neural and non-neural cells in the human brain and the cell types were integrated into their corresponding anatomical hierarchies. BRCO is dedicated ontology derived from Brain Region and Cell type Terminology (BRCT). BRCT contains more than 1,300 classes with average number of 5 children which showed a satisfactory F-score (0.80) in a named entity recognition task on an independent testing corpus composed of 100 manually annotated MEDLINE abstracts. Annotating unstructured data (e.g., from large cohort data sets like ADNI or the UK biobank or literature) using BRCT allows to apply automated workflows to link imaging features with other interested biological entities such as genes, pathways. Below is the figure that shows the overall statistics and structure of BRCO.

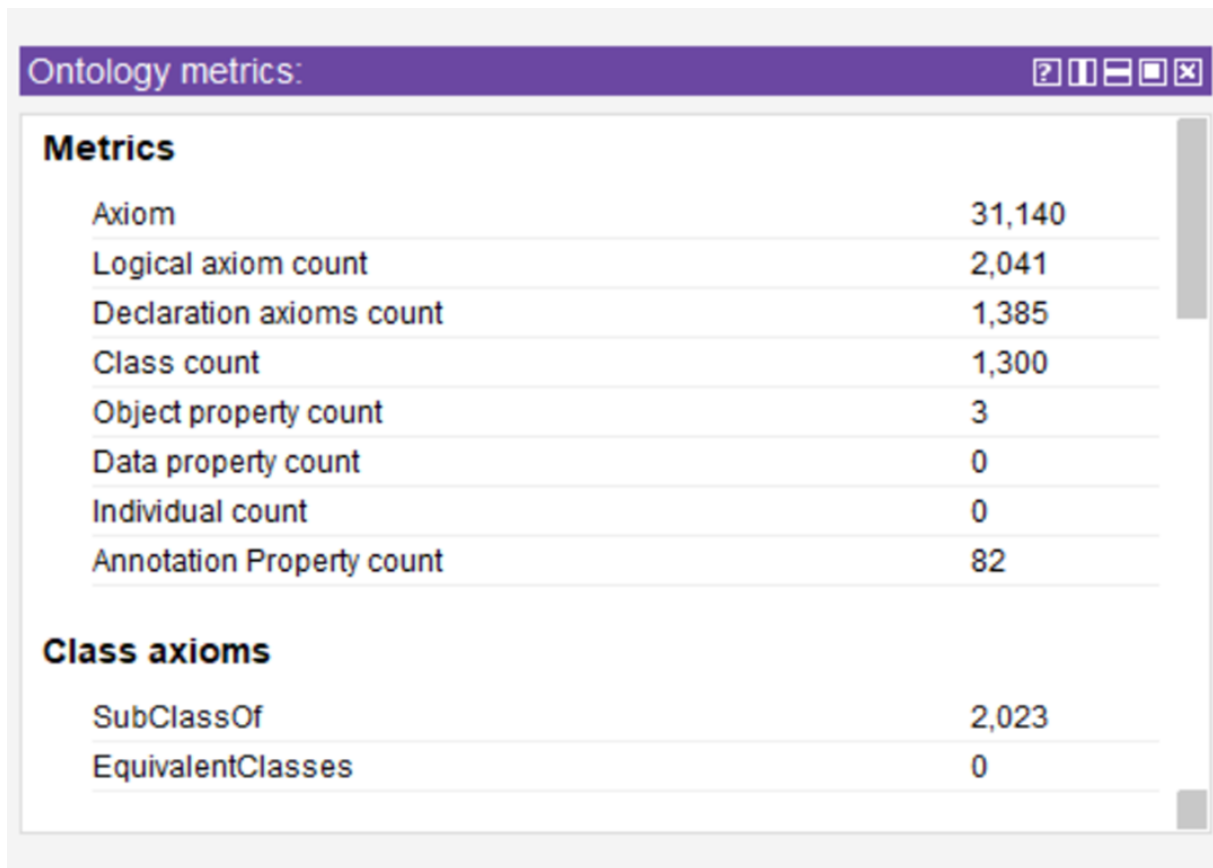


Figure 7: The overall statistics of BRCO

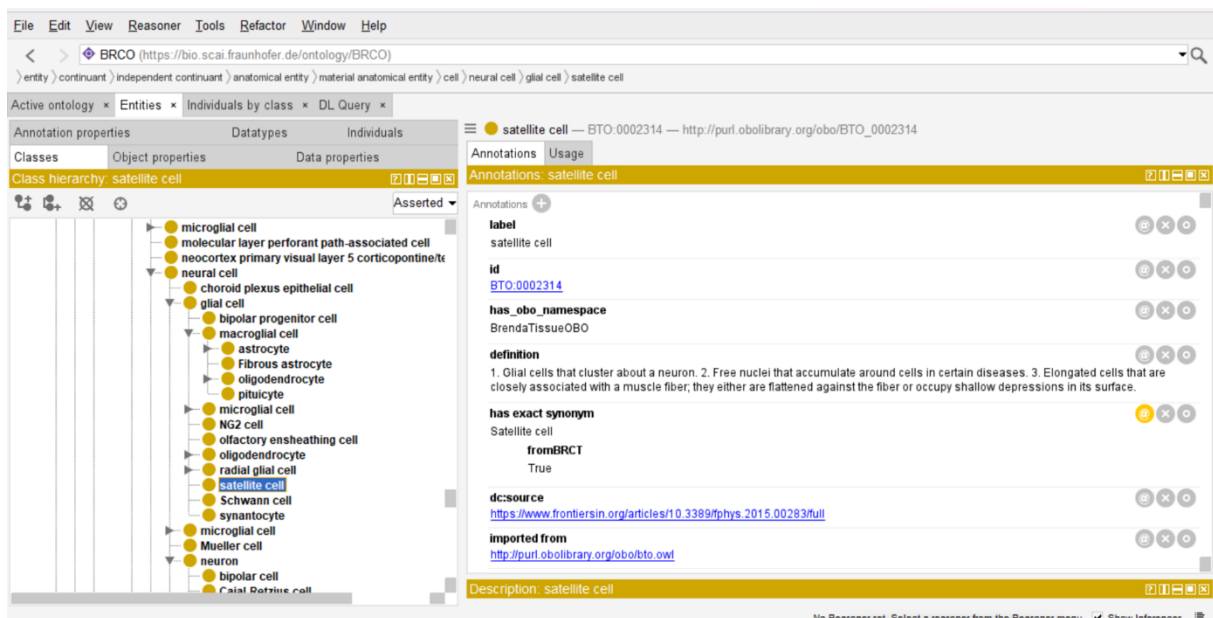


Figure 8: The screenshot of the NISO with various concepts along with annotations

6.5. EEG/MEG and Feature Terminology

The EEG / MEG Feature terminology is being developed as a Clinical decision support system (CDSS) to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information. The objective is to make use of web-applications or integration with electronic health records (EHR) and computerized provider order entry (CPOE)

systems. We have collected inputs on EEG data and the whole procedure from Prof. Dr. Claudio Babiloni, UNIROMA1. From UNIROMA1, we got inputs in the form of mindmap.

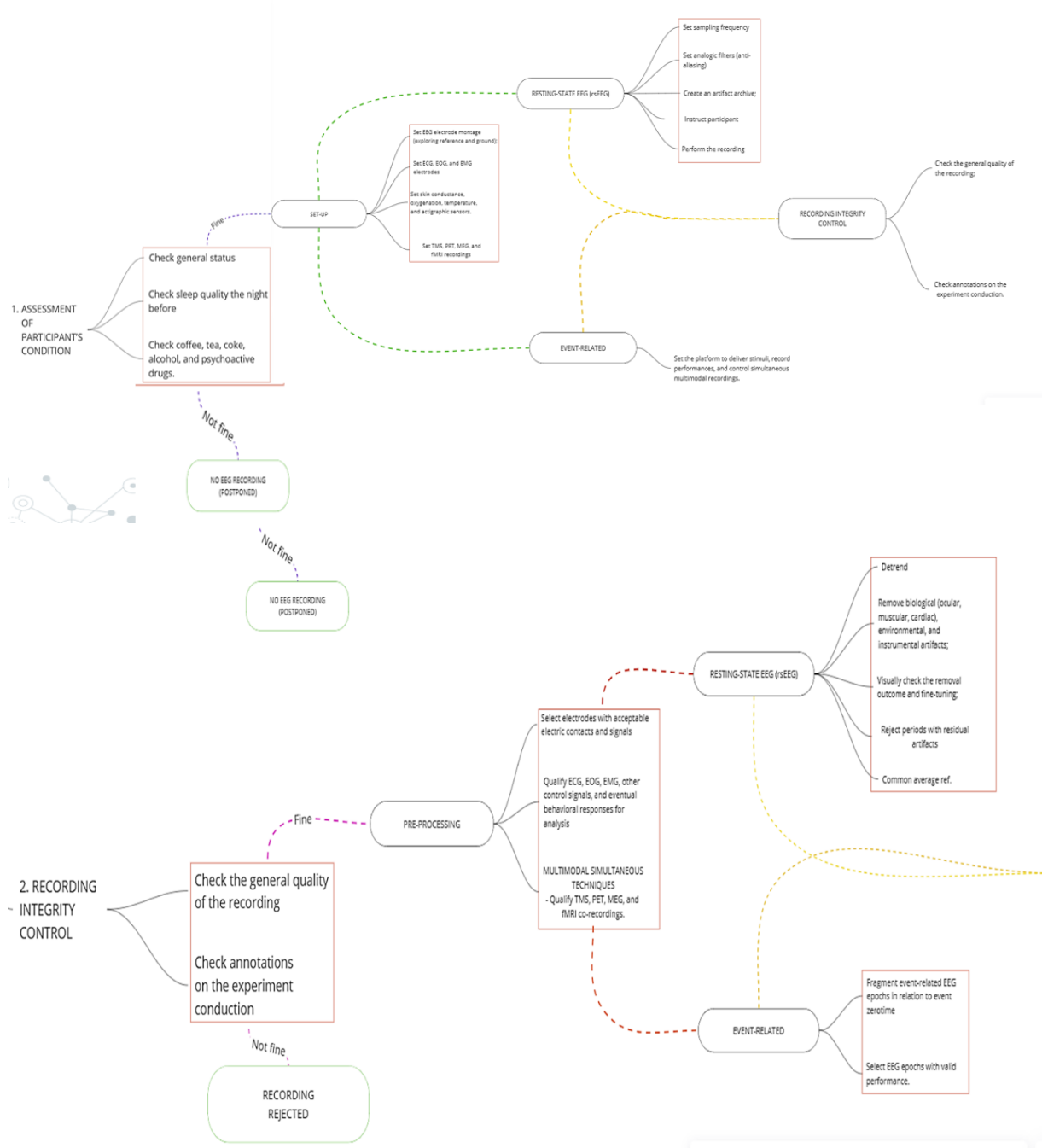


Figure 9: The screenshot of the mindmap from UNIROMA

We use Protégé Ontology Editor to try building a knowledge-based system. Terms within the workflow are categorized as Classes, Individuals and Object Properties and we use Semantic Web Rule Language (SWRL) to create the rules, which form the workflow.

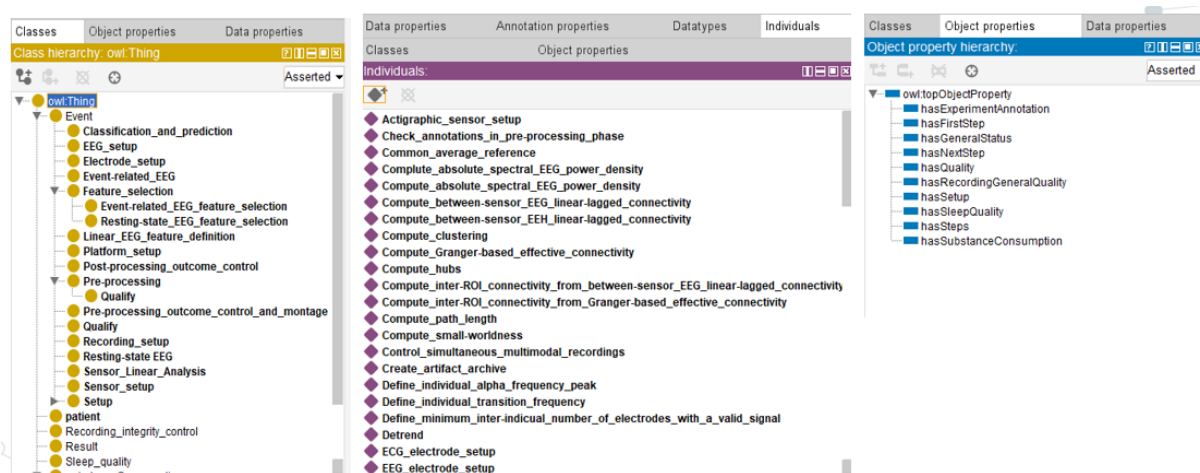


Figure 10: The screenshot of the current EEG Ontology based on the inputs from UNIROMA

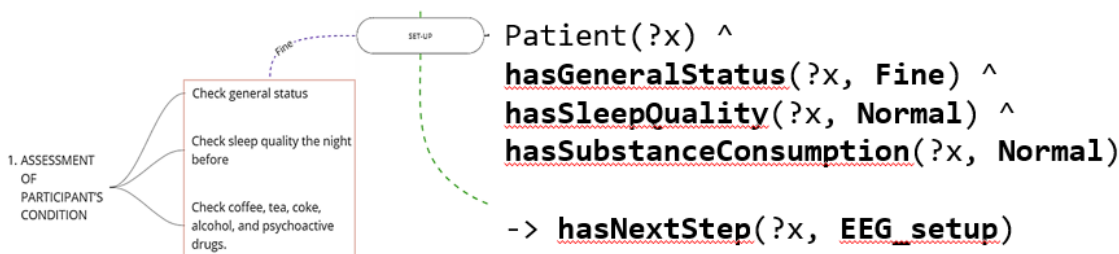
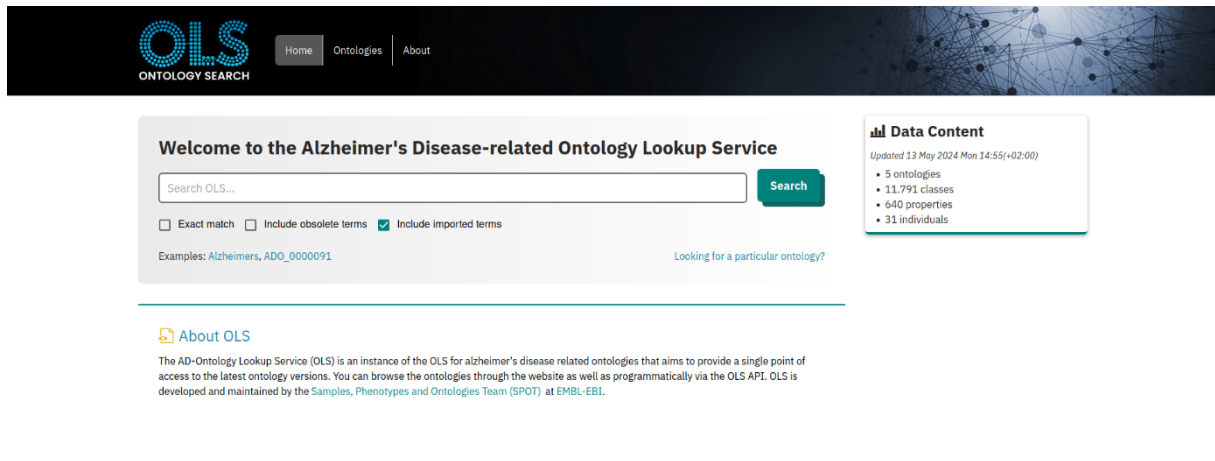


Figure 11: The screenshot of how the inputs from MindMap are transferred into SWRL rules

Apart from the terms from the workflow that are coming from Mindmap, we are enriching EEG ontology with additional terms related to EEG.

6.6. Semantic Framework & Management system

As of May 2024, the eBRAIN-Health semantic framework contains 11,791 classes and 640 properties/relations from 5 ontologies developed in the context of the project. Figure 12 and 13 shows a screenshot of the homepage and resources page of the eBRAIN-Health semantic framework service that lists the ontologies respectively. These ontologies comprise the majority of relevant entities and their relationships that are required to describe and harmonize data and knowledge about neurodegenerative diseases.



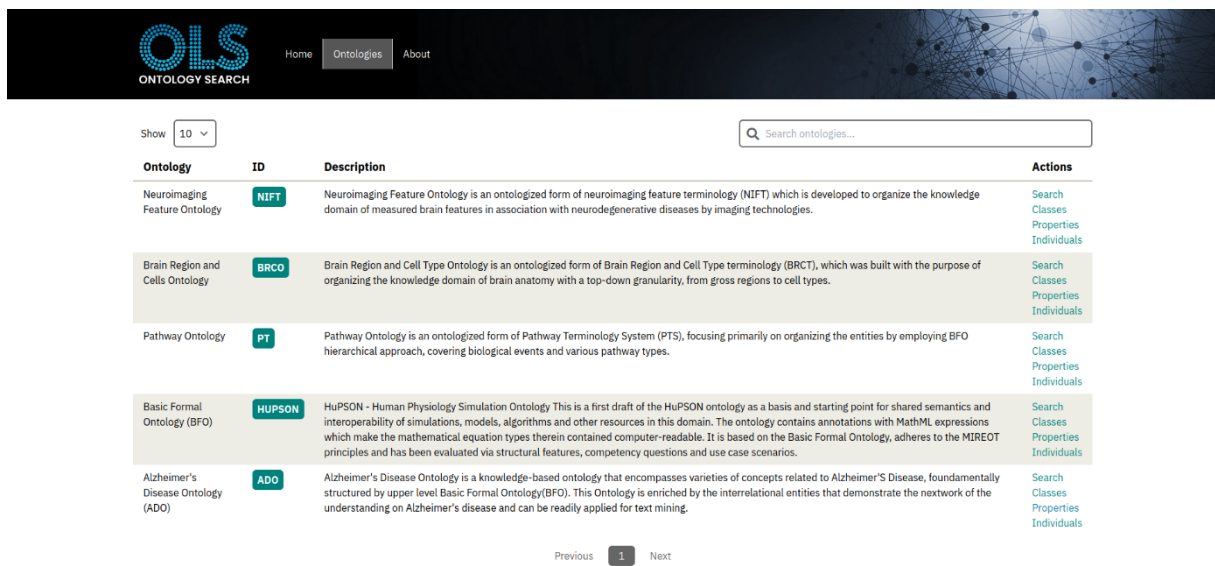
The screenshot shows the OLS homepage with a search bar, navigation tabs (Home, Ontologies, About), and a 'Data Content' box indicating 5 ontologies, 11,791 classes, 640 properties, and 31 individuals. A 'Welcome to the Alzheimer's Disease-related Ontology Lookup Service' message is displayed above the search interface.

Semantic Framework for AD by SCAI.BIO



Powered by OLS from [EMBL-EBI](#)

Figure 12: Semantic framework for AD homepage

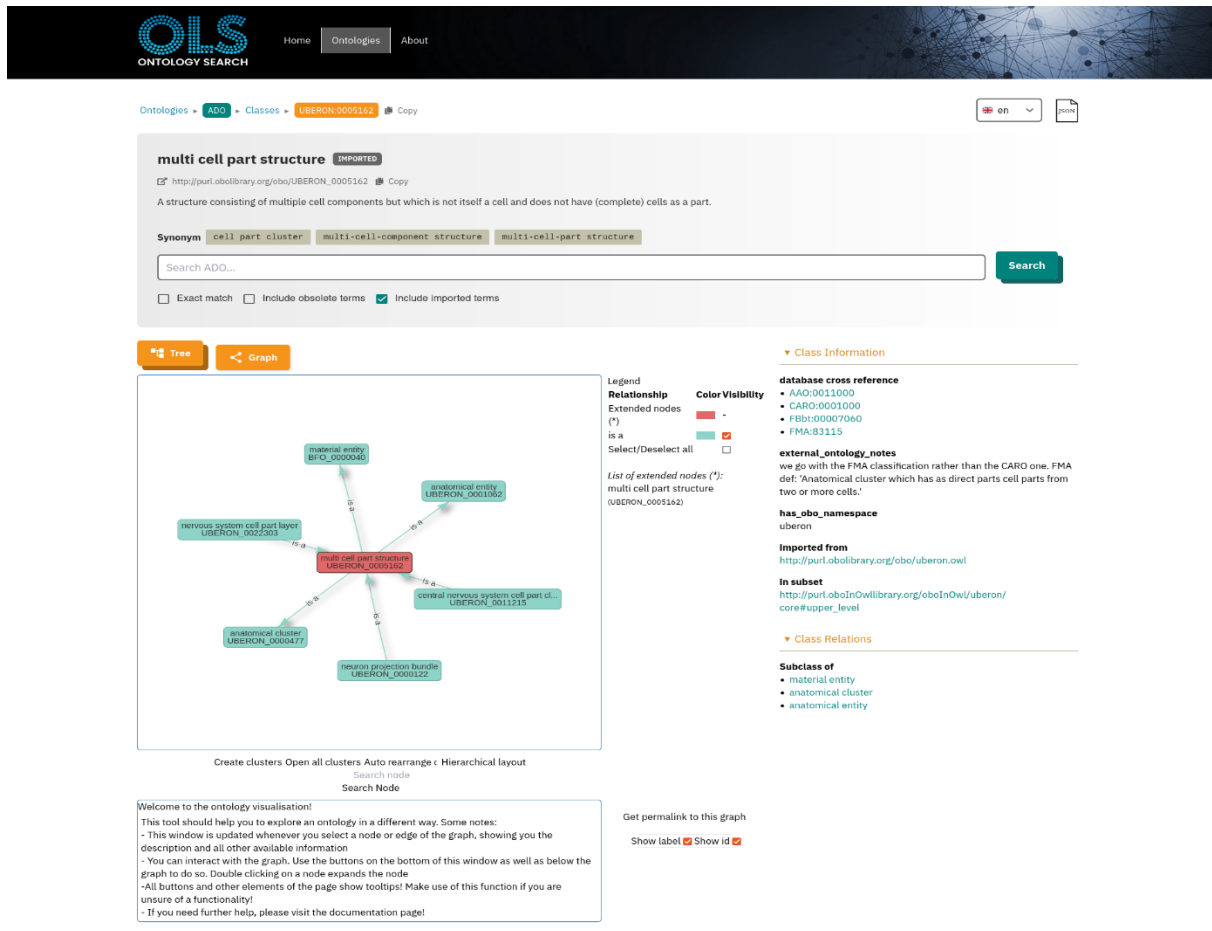


The screenshot displays a table of ontologies with columns for Ontology, ID, Description, and Actions. The table lists five ontologies: Neuroimaging Feature Ontology (NIFT), Brain Region and Cells Ontology (BRCO), Pathway Ontology (PT), Basic Formal Ontology (BFO) (HUPSON), and Alzheimer's Disease Ontology (ADO).

Ontology	ID	Description	Actions
Neuroimaging Feature Ontology	NIFT	Neuroimaging Feature Ontology is an ontologized form of neuroimaging feature terminology (NIFT) which is developed to organize the knowledge domain of measured brain features in association with neurodegenerative diseases by imaging technologies.	Search Classes Properties Individuals
Brain Region and Cells Ontology	BRCO	Brain Region and Cell Type Ontology is an ontologized form of Brain Region and Cell Type terminology (BRCT), which was built with the purpose of organizing the knowledge domain of brain anatomy with a top-down granularity, from gross regions to cell types.	Search Classes Properties Individuals
Pathway Ontology	PT	Pathway Ontology is an ontologized form of Pathway Terminology System (PTS), focusing primarily on organizing the entities by employing BFO hierarchical approach, covering biological events and various pathway types.	Search Classes Properties Individuals
Basic Formal Ontology (BFO)	HUPSON	HUPSON - Human Physiology Simulation Ontology This is a first draft of the HUPSON ontology as a basis and starting point for shared semantics and interoperability of simulations, models, algorithms and other resources in this domain. The ontology contains annotations with MathML expressions which make the mathematical equation types therein contained computer-readable. It is based on the Basic Formal Ontology, adheres to the MIREOT principles and has been evaluated via structural features, competency questions and use case scenarios.	Search Classes Properties Individuals
Alzheimer's Disease Ontology (ADO)	ADO	Alzheimer's Disease Ontology is a knowledge-based ontology that encompasses varieties of concepts related to Alzheimer's Disease, fundamentally structured by upper level Basic Formal Ontology(BFO). This Ontology is enriched by the interrelational entities that demonstrate the network of the understanding on Alzheimer's disease and can be readily applied for text mining.	Search Classes Properties Individuals

Figure 13: List of ontologies in the semantic framework

Updating and revision of the ontologies is an ongoing process within WP4. We are continuously updating ontologies, especially the Neuroimaging Feature Ontology, HUPSON, Brain Region and Cell Type Ontology, and Alzheimer's Disease Ontology (ADO) which act as the basis for the semantic framework. EEG ontology is currently being developed. The platform also allows for graph visualization of the concept of interest and its related terms as shown in figure 14.



Semantic Framework for AD by SCAI.BIO



Powered by OLS from EMBL-EBI

Figure 14: graph view of concepts

7. Conclusion, next steps

Work Package 4 has successfully developed a semantic framework that standardizes data integration for neurodegenerative diseases through the use of controlled vocabularies and shared ontologies. This framework includes publicly available resources such as the updated Alzheimer's Disease Ontology, now part of the OBO Foundry, and other key ontologies like the Pathway Terminology System, Brain Region and Cell type Ontology, EEG/MEG ontology, and Neuroimaging Feature Ontology.

These updated ontologies support various applications, including text mining, standardization, and clinical decision support. The Alzheimer's Disease Ontology, for example, has been effectively used for text mining to identify correlations between Alzheimer's symptoms and cellular processes, as well as biological pathways involved in the disease. The Neuroimaging Feature Ontology (NIFT) and Brain Region and Cell Type Ontology (BRCO) standardize imaging features and brain anatomy, respectively, while the EEG ontology is specifically designed to enhance clinical decision support systems.

Importantly, all these ontologies are publicly available on a dedicated website, ensuring broad accessibility for researchers and practitioners.

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.