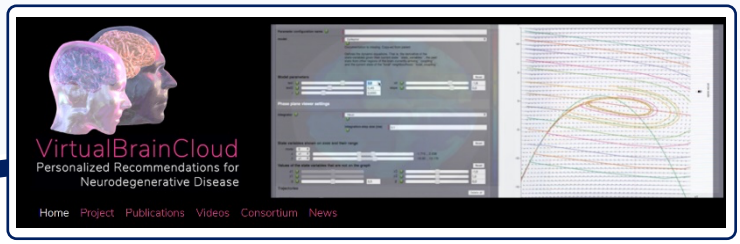
VirtualBrainCloud

Personalized Recommendations for Neurodegenerative Disease



https://twitter.com/tvb_cloud

Public deliverable report

D3.12: Evaluation and optimization of pipeline for practical use
Concluded

Date	November 2021
Authors	Fraunhofer SCAI: Vanessa Lage-Rupprecht, Alpha Tom Kodamullil, Philipp Wegner
Dissemination level:	**public**
Website	https://virtualbraincloud-2020.eu

This project has received funding from the **European Union´s Horizon 2020** research and innovation programme under **grant agreement**

Table of content

1. Introduction

Work Package 3 focuses, among other things, on the creation and provision of a semantic framework for neurodegenerative diseases in the form of central resources for controlled vocabularies and shared ontologies. Here, we address tasks 3.4 ("Workflows for clinical data curation and processing" and 3.6 ("FAIR data principle implementation on all modalities") with the deliverables 3.10 ("Framework for multimodal integrated annotations established and in use") and 3.12 ("Evaluation and optimization of pipeline for practical use concluded"). Main objectives of these deliverables were to realize multimodal knowledge integration and its usage, the fulfilment of the FAIR data priciples in all our applications, to continuously update existing resources via dedicated curation pipelines and software content updates, and to ensure reproducibility of applications through intelligent software deployment (see section 2.6 as an example).

Within TVB-Cloud two systems are implemented and made available for the consortium members. The content in these systems requires continuous updates.

- The Ontology Lookup Service (OLS) from Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), and
- the CTS2 system from University of Genova (UNIGE).

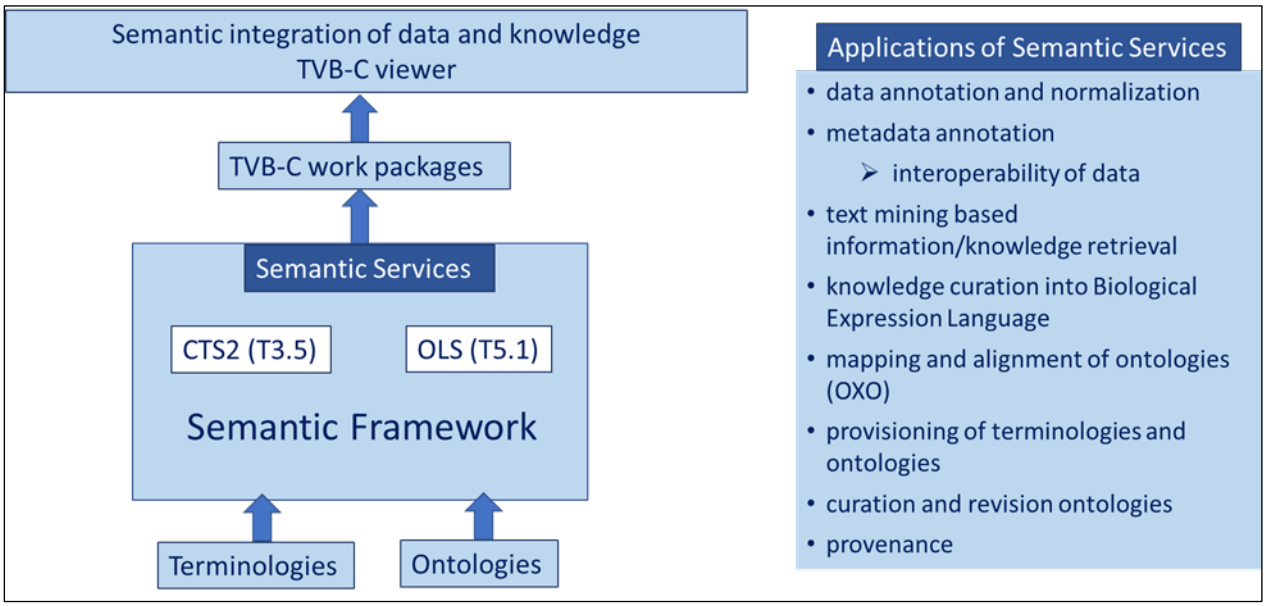


Figure 1: Schema of the Sematic Framework

The two deliverables share the same premise, but one was mainly developed by Fraunhofer SCAI and the other by the University of Genoa. In a context of shared semantics, CTS2 provides a modular, common and universally implementable set of operations, which can then be used to manage a set of terminologies chosen by service users in their own distribution environment. The Ontology Lookup Service (OLS) provides a repository for biomedical ontologies that aims to provide a single point of

access to the newest ontology versions. These services complement each other and generate a sematic framework. Technically, they are realized as independent services.

With these two deliverables two demonstrators of the websites are made available at the following addresses:

- OLS, the ontology lookup service software is being used to host the Referential Ontology Hub for Applications within Neurosciences (ROHAN), which is accessible at https://rohan.scai.fraunhofer.de/ols/index.

- CTS2 compliant interface: The read and query part are open to everyone, while the access to the maintenance part is limited by a login. This access can be given to any member that wants to make a trial. The address is: http://www.medinfo.dibris.unige.it/VBC_CTS2 (for more information see delivery D3.14: "Integrated, open-source terminology versioning and management tool CTS2 compliant application shared with partners").

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

2. Results

2.1. ROHAN service - content update

By November 2021 the repository contains 14 terminologies and ontologies. Altoghether they consist of 57,361 terms, 2,378 properties and 1,204 individuals. Figure 2 shows a screenshot of the resources page of ROHAN service that lists the ontologies. These ontologies comprise the majority of relevant entities and their relationships that are required to describe and harmonize data and knowledge about neurodegenerative diseases. Updating and revision of the ontologies is an ongoing process within WP3 and WP5. We are continuously updating ontologies, especially the Clinical Trial Ontology (CTO), Neuroimaging Feature Ontology, HUPSON, Brain Region and Cell Type Ontology, Alzheimer Disease Ontology (ADO) which act as the basis for the semantic framework.

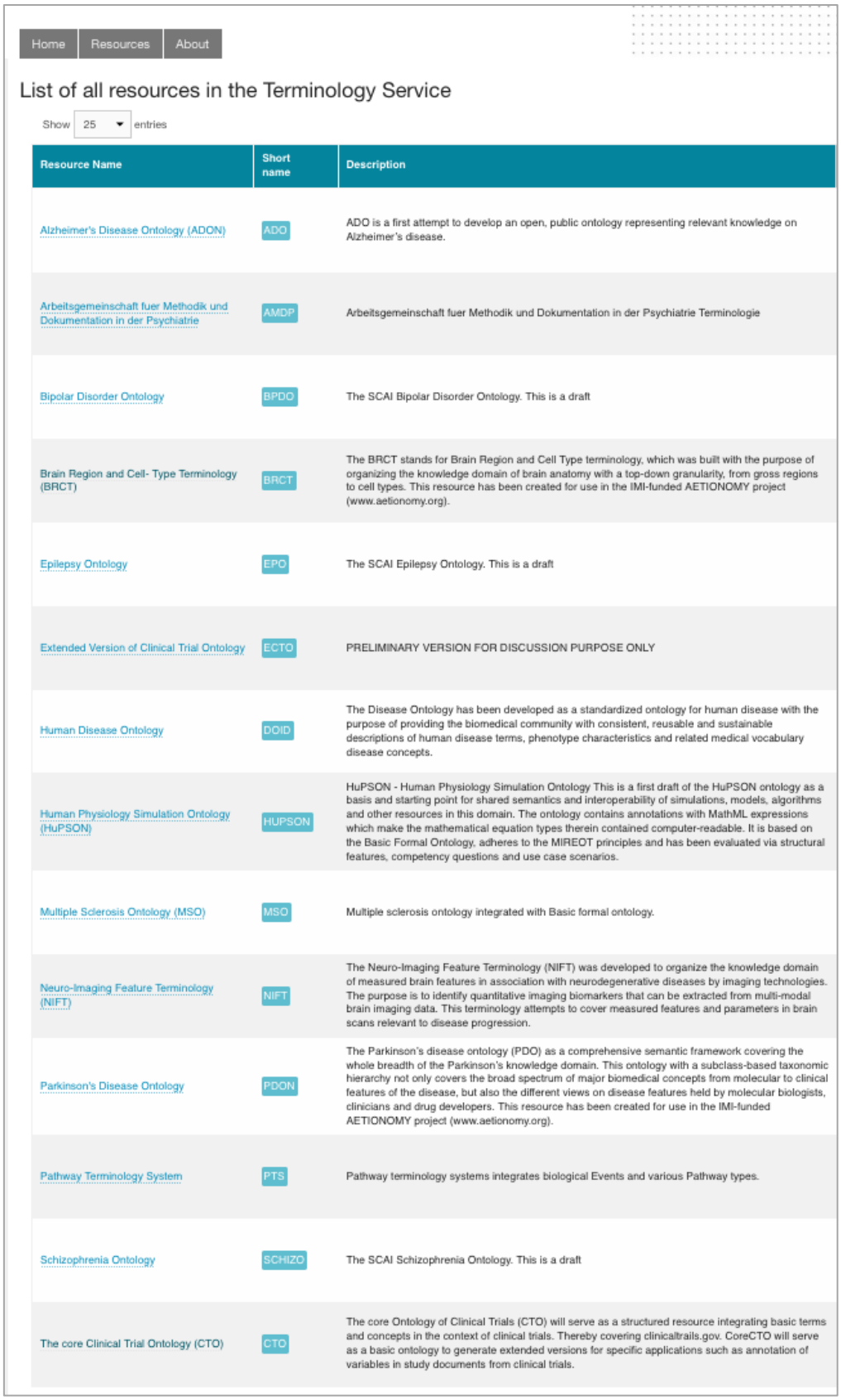
In the next section, we describe the updating of the Clinical Trial Ontology and the Alzheimer Disease Ontology. Other ontologies will be updated accordingly, if necessary.

2.2. Clinical Trial Ontology

To model the field of clinical trials, the Clinical Trial Ontology for Neurodegeneration Diseases (CTO-NDD) was developed for the IMI-funded AETIONOMY (https://www.aetionomy.eu/) project. During the TVB-Cloud project, this ontology has been updated. The overall structure of the updated ontology is based on community-agreed standards, like basic formal ontology (BFO) and, in addition, the best practices as well as the principles of Open Biological and Biomedical Ontology (OBO) Foundry are applied. The development of the ontology is undertaken by using Protégé OWL editor (https://protege.stanford.edu/) in ontology web language (OWL) format.

We have continued to work on the CTO. There are now two ontologies, a core Clinical Trials Ontology (coreCTO) and an extended Clinical Trials Ontology (extCTO). The updated version, the core Clinical Trials Ontology (Figures 3 and 4) serves as a structured resource integrating basic terms and concepts in the context of all clinical trials (Figure 6 and Figure 7). Therefore, terms and concepts from ClinicalTrials.gov [1] as dedicated data provider to the WHO International Clinical Trials Registry Platform (ICTRP) were integrated (https://prsinfo.clinicaltrials.gov/definitions.html). To ensure alignment with WHO and International Committee of Medical Journal Editors (ICMJE) recommendations, terms outlined in WHO ICTRP were considered (https://www.who.int/clinical-trials-registry-platform/network/who-data-set). The coreCTO is being developed together with external partners from NCBI, FDA and University of Michigan. The extCTO will be used to support implementation of specific applications such as annotation of variables in documents derived from neurodegeneration disease clinical trials and further text mining use cases.

We have developed an application case for the coreCTO and published the initial version of this ontology [2].The current version of the CTO ontology is available in the ROHAN service.



Home | Resources | About

List of all resources in the Terminology Service

Show 25 entries

Resource Name	Short name	Description
Alzheimer's Disease Ontology (ADON)	ADO	ADO is a first attempt to develop an open, public ontology representing relevant knowledge on Alzheimer's disease.
Arbeitsgemeinschaft fuer Methodik und Dokumentation in der Psychiatrie	AMDP	Arbeitsgemeinschaft fuer Methodik und Dokumentation in der Psychiatrie Terminologie
Bipolar Disorder Ontology	BPDO	The SCAI Bipolar Disorder Ontology. This is a draft
Brain Region and Cell- Type Terminology (BRCT)	BRCT	The BRCT stands for Brain Region and Cell Type terminology, 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. This resource has been created for use in the IMI-funded AETIONOMY project (www.aetionomy.org).
Epilepsy Ontology	EPO	The SCAI Epilepsy Ontology. This is a draft
Extended Version of Clinical Trial Ontology	ECTO	PRELIMINARY VERSION FOR DISCUSSION PURPOSE ONLY
Human Disease Ontology	DOID	The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts.
Human Physiology Simulation Ontology (HuPSON)	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.
Multiple Sclerosis Ontology (MSO)	MSO	Multiple sclerosis ontology integrated with Basic formal ontology.
Neuro-Imaging Feature Terminology (NIFT)	NIFT	The Neuro-Imaging Feature Terminology (NIFT) was developed to organize the knowledge domain of measured brain features in association with neurodegenerative diseases by imaging technologies. The purpose is to identify quantitative imaging biomarkers that can be extracted from multi-modal brain imaging data. This terminology attempts to cover measured features and parameters in brain scans relevant to disease progression.
Parkinson's Disease Ontology	PDON	The Parkinson's disease ontology (PDO) as a comprehensive semantic framework covering the whole breadth of the Parkinson's knowledge domain. This ontology with a subclass-based taxonomic hierarchy not only covers the broad spectrum of major biomedical concepts from molecular to clinical features of the disease, but also the different views on disease features held by molecular biologists, clinicians and drug developers. This resource has been created for use in the IMI-funded AETIONOMY project (www.aetionomy.org).
Pathway Terminology System	PTS	Pathway terminology systems integrates biological Events and various Pathway types.
Schizophrenia Ontology	SCHIZO	The SCAI Schizophrenia Ontology. This is a draft
The core Clinical Trial Ontology (CTO)	CTO	The core Ontology of Clinical Trials (CTO) will serve as a structured resource integrating basic terms and concepts in the context of clinical trials. Thereby covering clinicaltrails.gov. CoreCTO will serve as a basic ontology to generate extended versions for specific applications such as annotation of variables in study documents from clinical trials.

Figure 2: Screenshot of terminologies and ontologies integrated into the ontology hub (https://rohan.scai.fraunhofer.de/ols/ontologies)

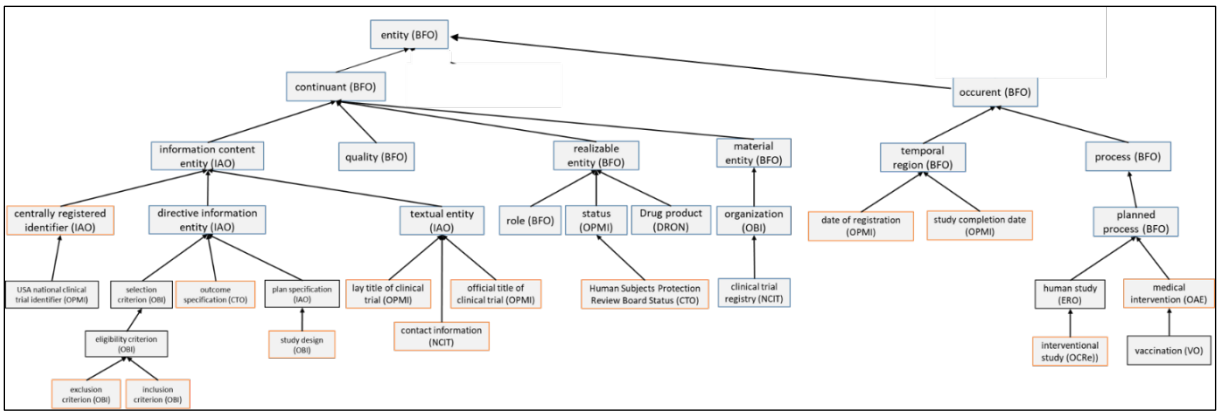


Figure 3: Top hierarchical architecture of coreCTO aligned to basic formal ontology (BFO). Concepts that are defined in WHO ICTRP are framed in red. Name of the OBO ontologies of reused concepts are given in brackets after the concept name.

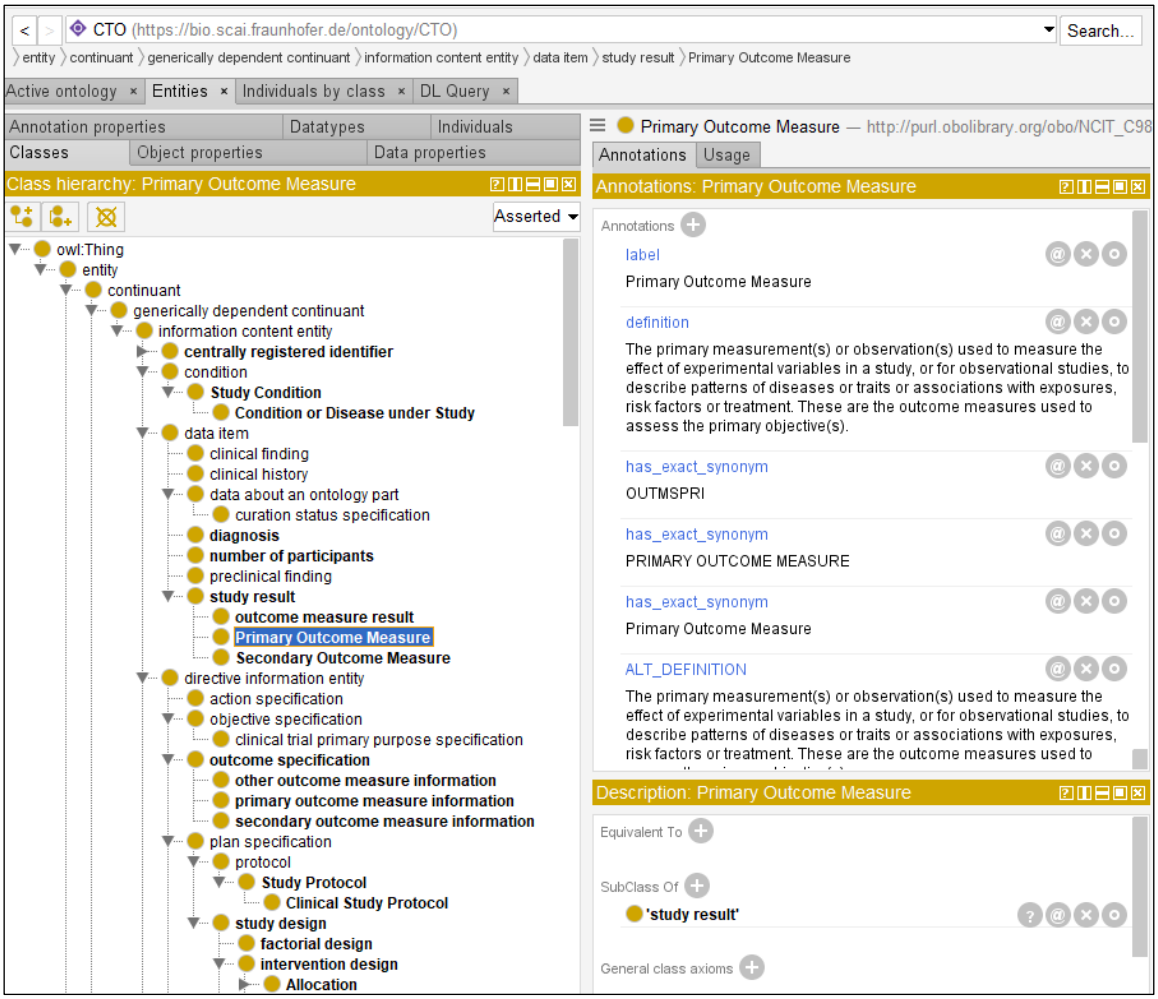


Figure 4: Screenshot of the core Clinical Trials Ontology showing hierarchal structure and annotations of the concepts (displayed with Protégé editor)

2.3. Alzheimer's Disease Ontology

Fraunhofer SCAI has released an Alzheimer Disease Ontology (ADO) [3] in 2014 as the first attempt to represent knowledge semantically. In 2018, Alzheimer Disease Map Ontology (ADMO) [4] was released which is based on systems biology terms. While ADO was aimed to integrate concepts from four domains of AD - preclinical, clinical, etiological, and molecular/cellular mechanisms, ADMO ontological model was made to manage the conversion and integration of AlzPathway (http://alzpathway.org/AlzPathway.html) [4] in OWL format. Even though both ontologies are relevant and complement to each other in its content, they do not follow the OBO foundry principles and are not based on a common upper hierarchy followed by other standard ontologies in the community. This is important when we aim to do further analysis with ontologies like reasoning or to establish the semantic interoperability with other domains. Currently, we are building an integrated version of the ontology which will follow the OBO foundry principles. Concepts from both ontologies (ADO/ADMO) will be considered in the newly written ontology. As internal structure of the integrated Alzheimer Disease Ontology we chose the basic formal ontology (BFO, https://basic-formal-ontology.org/), a hierarchy that is accepted as standard by the OBO foundry (http://www.obofoundry.org/). If concepts exist in other ontologies registered by OBO, these concepts are imported and reused. The integrated Alzheimer Disease ontology will be submitted to the OBO Foundry and a journal publication is in preparation.

2.4. Excerpt of ontology curation guideline

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 (Figure 4). 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.

2.5. Data catalogues, mappings, knowledge graphs and services

Fraunhofer SCAI provides and updates the web services below:

AData Viewer (https://adata.scai.fraunhofer.de): Harmonizing all the public AD cohort datasets available

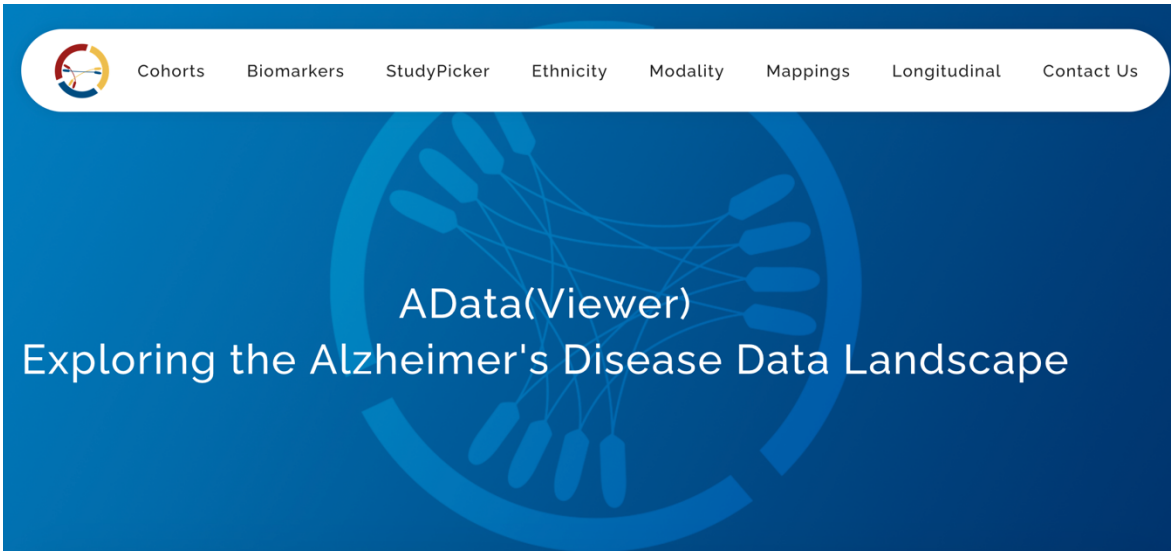


Figure 5: Screenshot of the AData Viewer frontpage

PathMe Viewer (https://pathme.scai.fraunhofer.de): Merging and exploring mechanistic pathway knowledge

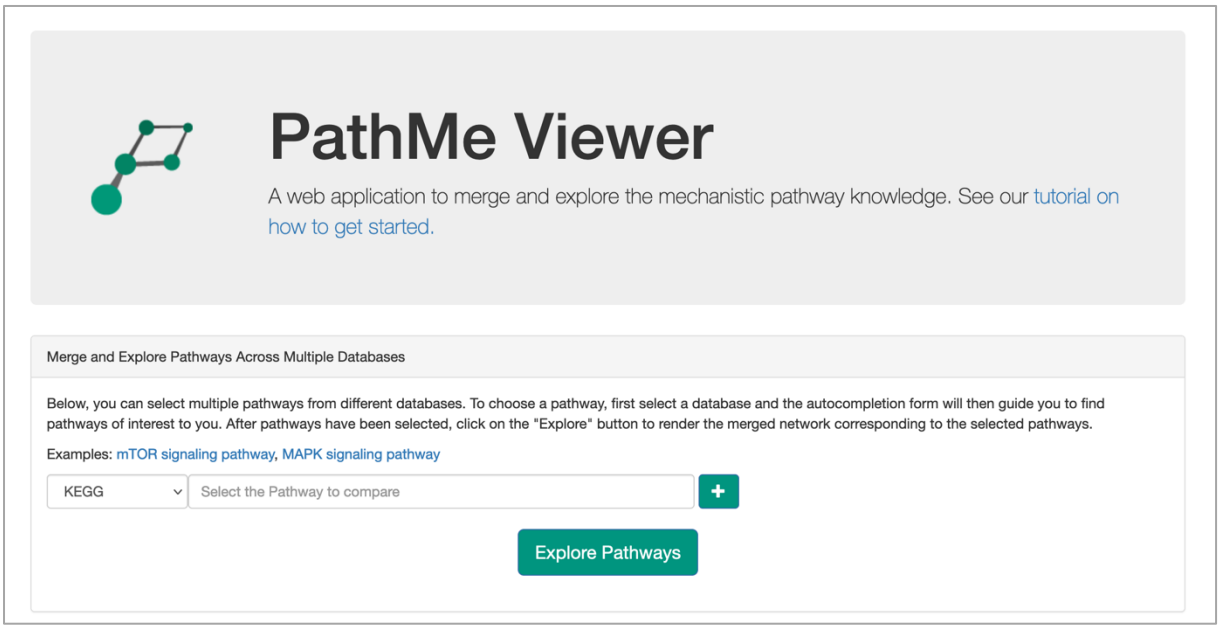


Figure 6: Screenshot of the PathMe Viewer frontpage

ComPath (https://compath.scai.fraunhofer.de): an integrative and extensible web service for exploring, comparing, and curating pathway databases

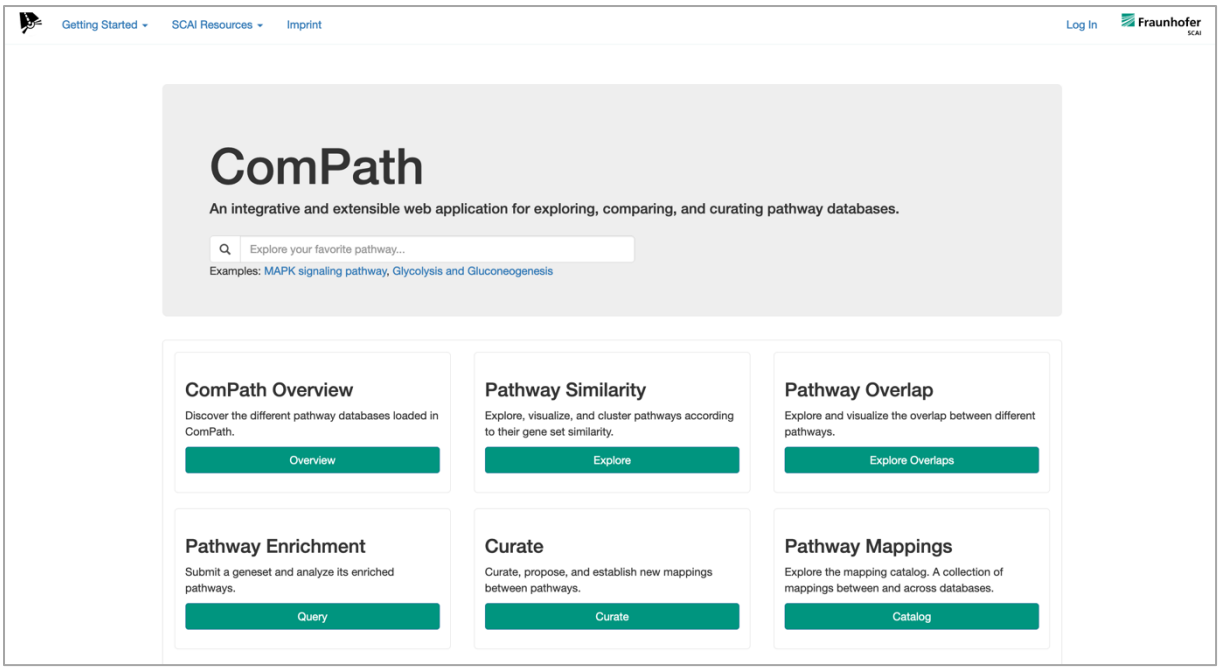


Figure 7: Screenshot of the PComPath frontpage

NeuroMMSig server (https://neurommsig.scai.fraunhofer.de): Interface for users to submit data to the mechanistic subgraphs in the Alzheimer's Disease, Parkinson's Disease and Epilepsy knowledge assemblies built in Fraunhofer SCAI.

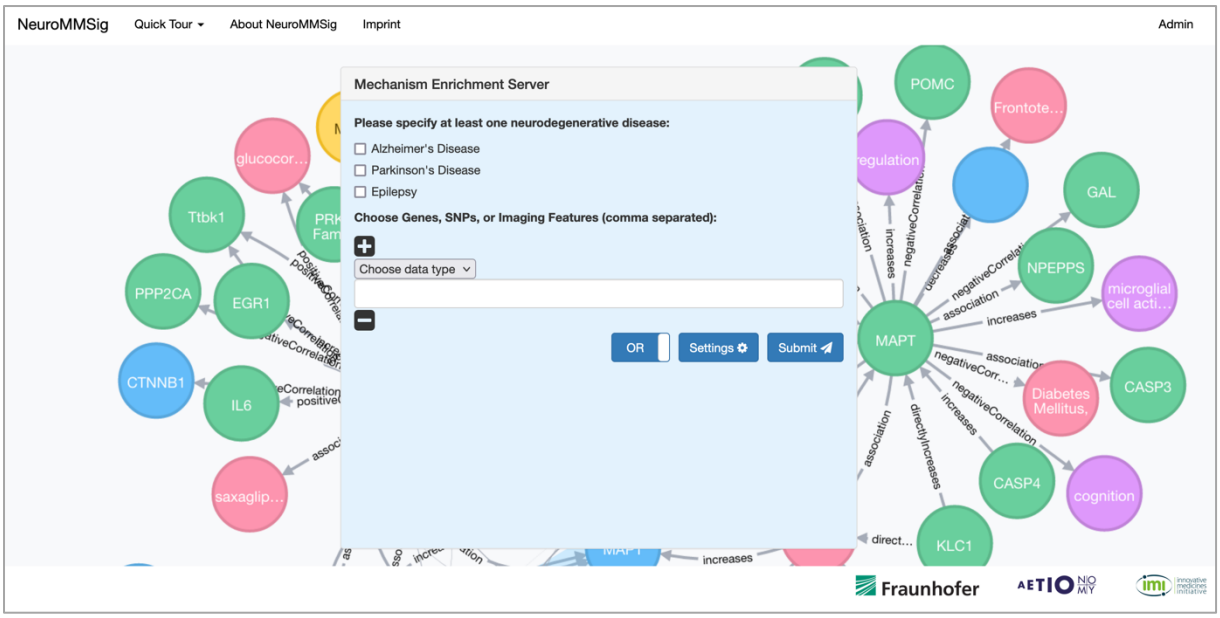


Figure 8: Screenshot of the NeuroMMSig frontpage

BiKMi - Biomedical Knowledge Miner (https://bikmi.pharmacome.scaiview.com/): provides tools to access and validate knowledge encompassing all of the latest information pertaining to Alzheimer's Disease.

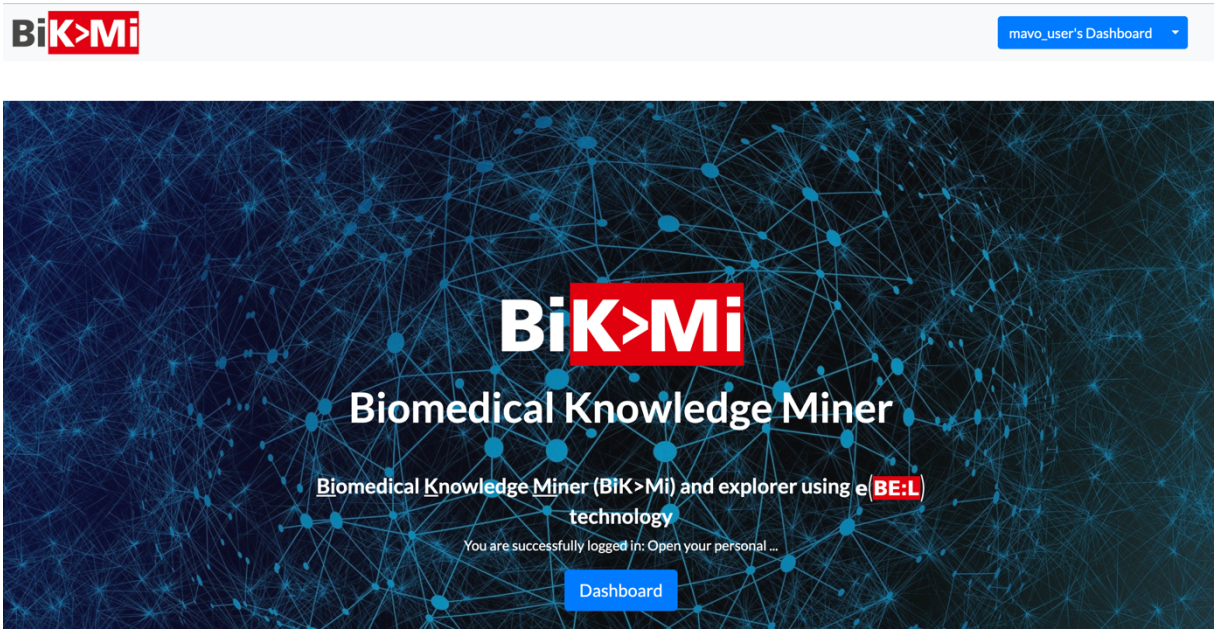


Figure 9: Screenshot of the BiKMi frontpage

Data Steward Tool (DST) - an application that enables semi-automated semantic integration of clinical data into ontologies, global data models, and data standards. The applicability of the tool has been tested in the field of dementia by establishing a Clinical Data Model (CDM). A manuscript describing the DST has been submitted to the Bioinformatics Journal (the preprint is available here: https://arxiv.org/abs/2111.09313).

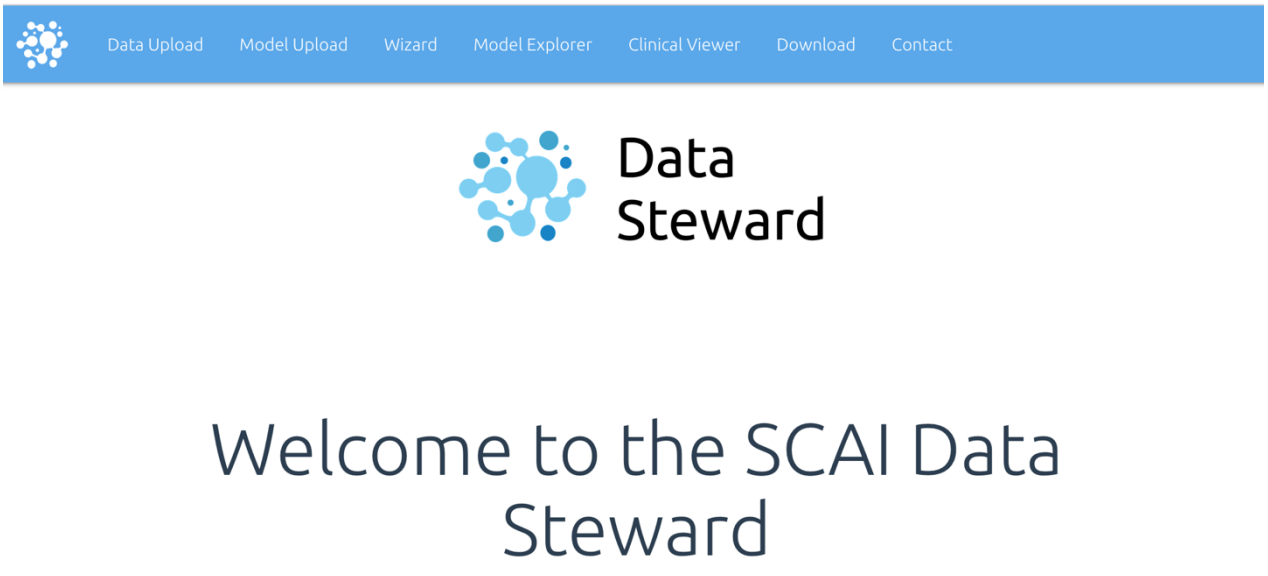


Figure 10: Screenshot of the Data Steward Tool

Knowledge Graph Services

Knowledge graphs have been generated in a disease-context manner. All are available (with guest accounts) at: **https://graphstore.scai.fraunhofer.de**

2.6. SCAIView-Neuro (Semantic Search Engine)

SCAIView is an advanced text mining system and semantic search engine that addresses questions of interest to general biomedical and life science researchers. Most of the current knowledge exists as unstructured text (publications, text fields in databases) and SCAIView provides users with full-text and biomedical concept searches, which are supported by large biomedical terminologies and outstanding text mining technologies. Using machine learning and rule-based named entity recognition, SCAIView extracts information about genes, drugs, SNPs and other life science entities. SCAIView provides various functionalities and further have been developed within TVB-Cloud together with corresponding view components (see D6.1 and D6.2). It is planned to deploy SCAIView in the central TVB-Cloud platform Virtual Research Environment (VRE, https://vre.charite.de/) hosted at the CHARITE.

Corpus selection and advanced search utility: SCAIView provides access to various corpora, for instance, it includes documents from PubMed, PubMed Central databases (but also arbitrary text corpora). Using the SCAI text mining pipeline, these corpora are annotated with ontological concepts that occur in the documents. The search and query utility allows a user to 'fire' free text search, search for a specific ontological concept, or search documents that are tagged with a specific ontology in all available corpora. Figure 10 and Figure 11 show two screenshots of the corpus selection page and an exemplary search query.

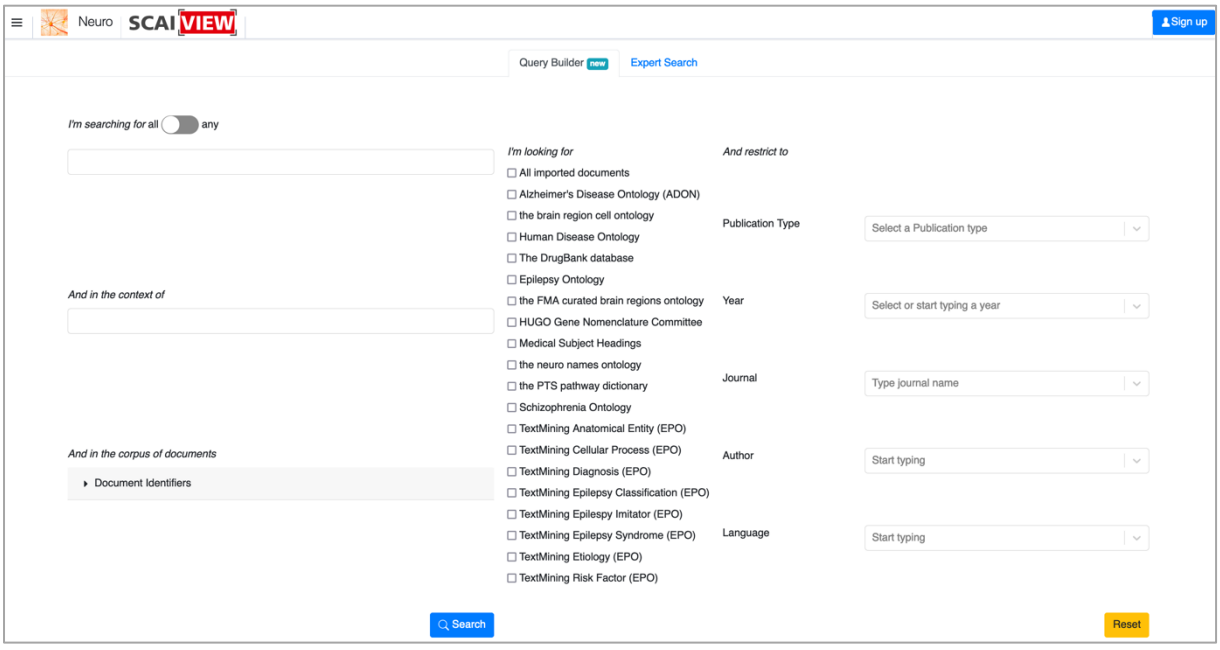


Figure 11: User Interface and the functionality of the semantic search engine Neuro SCAIView

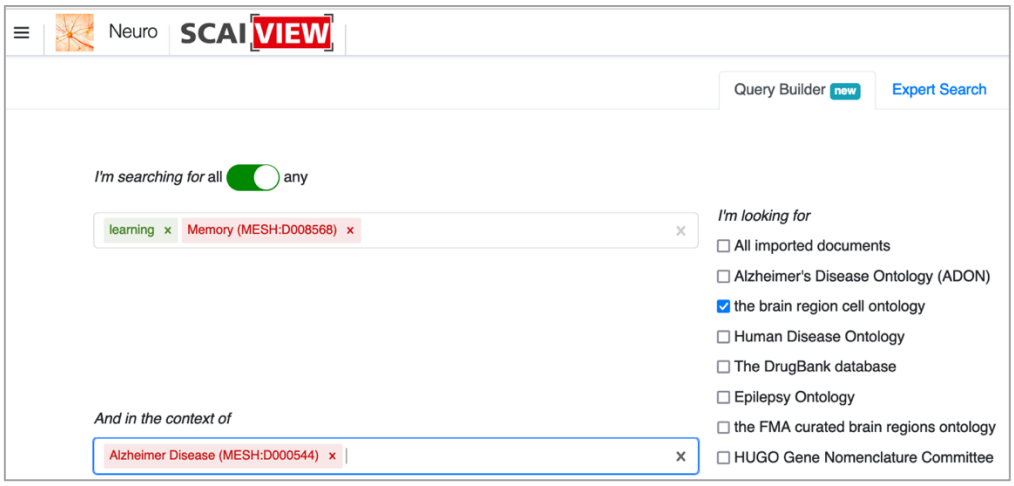


Figure 12: Query example: one can search for 'learning' as a free text (green) term, memory and Alzheimer Disease as a concept (red) and 'the brain region cell ontology as tag in https://neuro.scaiview.com/

Document list view is available to visualize the results of a search query. Figure 12 shows the document list view that highlights the search query and the 1818 retrieved documents. Each document is represented by its title, abstract, authors, and the text mined annotations. The view also allows the user to change the sorting of the results.

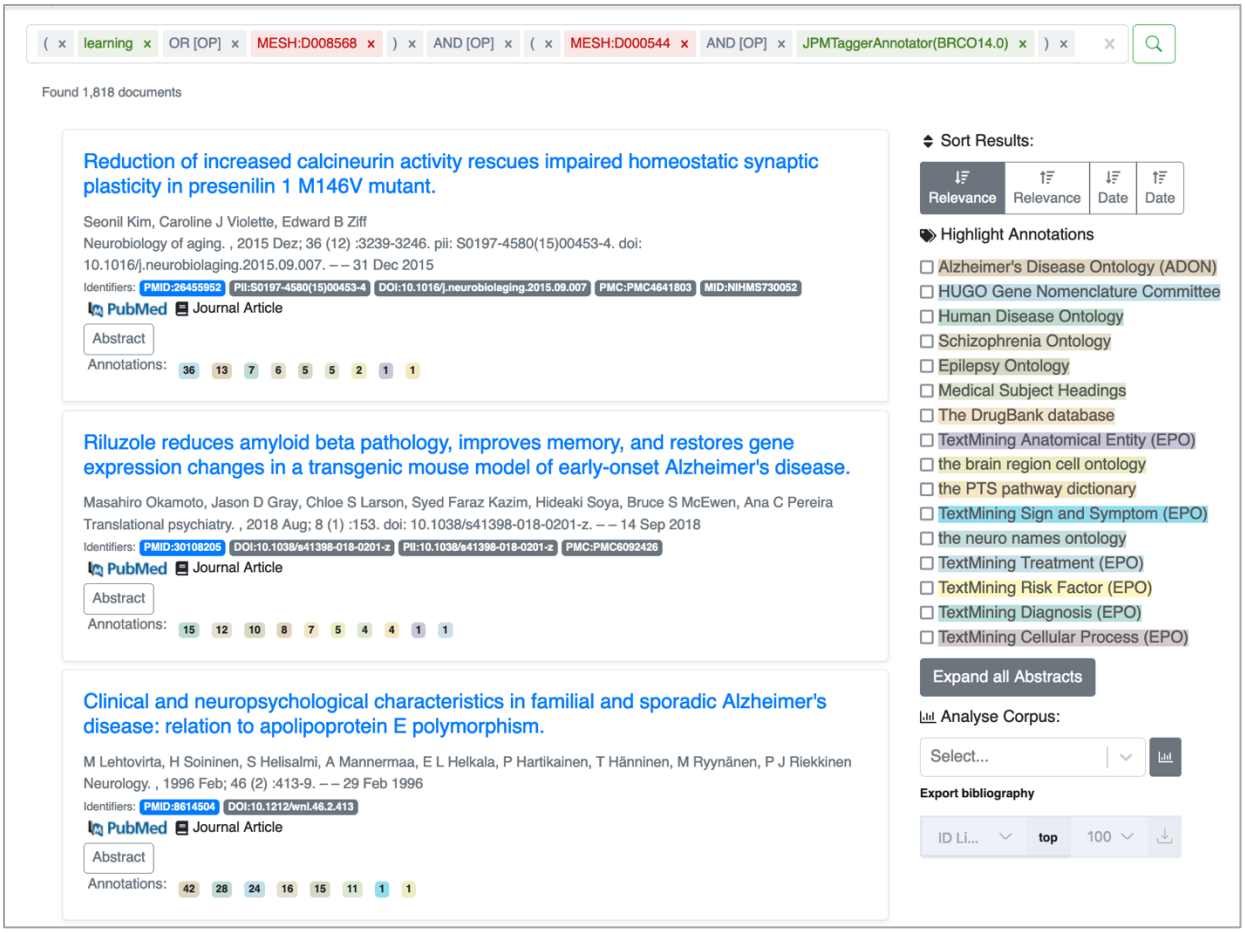


Figure 13: Document list view of SCAIView

Full text view: Visualization of a full text view of a document that highlights text mining results (annotations) using different colors for each ontology or terminology. It has been developed specifically to visualize the full-text of a single document. Figure 13 shows a screenshot of a PubMed article with its corresponding text mined results.

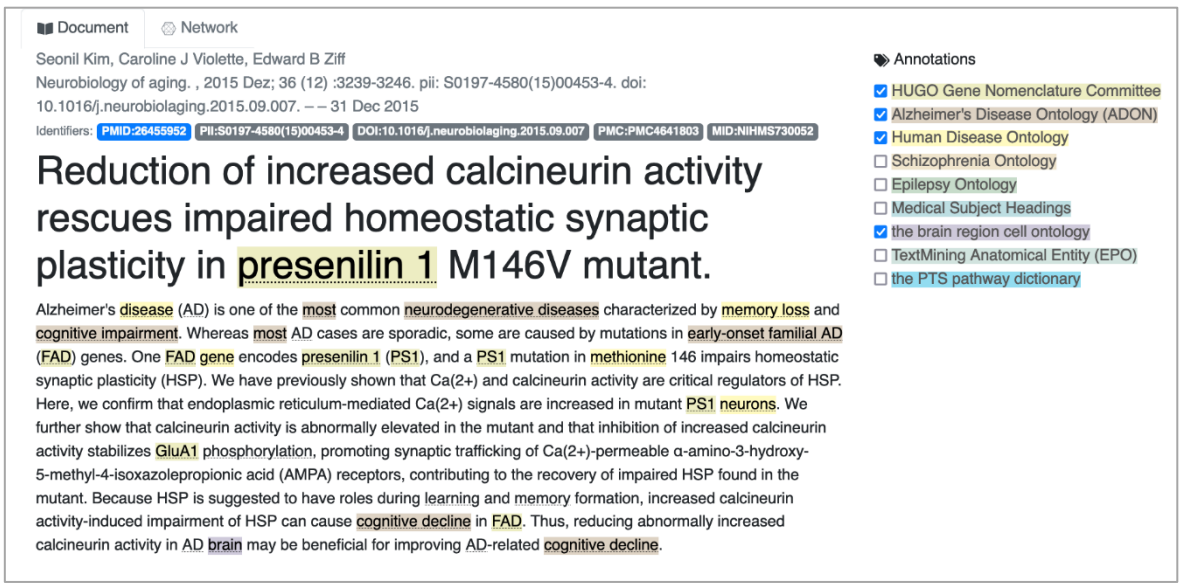


Figure 14: Full-text view of the PubMed article with highlighted annotations

Authentication and Authorization: The access to SCAIView GUI and API is secured with the OpenID Connect (OIDC) authentication standard. Only authorize users can access, upload and modify their corpora, documents and associated metadata. Furthermore, these services are also secured on the transport level via Hypertext Transfer Protocol Secure (HTTPS) technique. Hence, SCAIView is capable to offer a secured access to sensitive documents.

Stability and Up-to-dateness: We are ensuring the reproducibility of our software deployment by having all of our software compiled into images to be run in containers. Our services have transitioned to the 'kubernetes' open-source platform allowing us to easily monitor and scale our applications as needed. Document collections as well as ontology and terminology resources are continuously updated in our SCAIView instance to bring literature and annotations up to date.

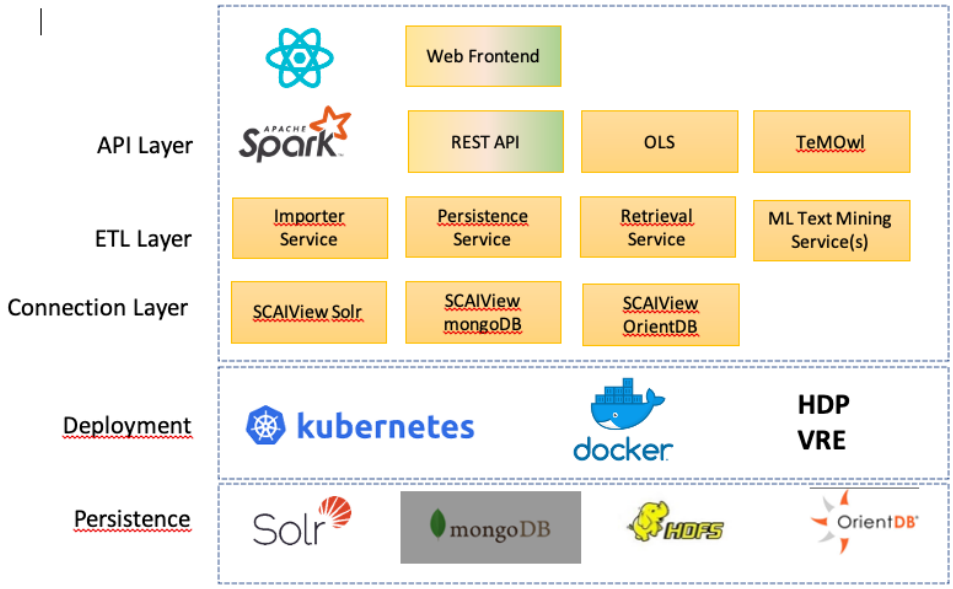


Figure 15: SCAIView microservice architecture

The microservice-based system allows us to reuse parts for different purposes. Data itself can be easily processed, shared and accessed. Additionally, the system allows us to realize FAIR (Findable, Accessible, Interoperable, and Reusable) principles.

3. Conclusions

The wealth of medical data and scientific publications is constantly growing. Updating and managing data that are collected on a regular basis poses a new challenge. Keeping the state of knowledge in resources such as ontologies, terminologies and knowledge graphs continuously up to date is not trivial. We developed strategies for updating and curation of our domain knowledge with continuously updated high-quality content. We use software development tools as well as data mining and curation approaches to ensure continuous integration of new knowledge and the fulfillment of FAIR data principles.

4. References

1. Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M., and Ide, N.C. (2011). The ClinicalTrials.gov Results Database — Update and Key Issues. N. Engl. J. Med. *364* https://doi.org/10.1056/nejmsa1012065
2. Lin, A.Y., Gebel, S., Li, Q.L., Madan, S., Darms, J., Bolton, E., Smith, B., Hofmann-Apitius, M., He, Y.O., and Kodamullil, A.T. (2021). CTO: A Community-Based Clinical Trial Ontology and Its Applications in PubChemRDF and SCAIView.
3. Malhotra, A., Younesi, E., Gündel, M., Müller, B., Heneka, M.T., and Hofmann-Apitius, M. (2014). ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimers. Dement. *10*, 238–246 https://doi.org/10.1016/J.JALZ.2013.02.009
4. Henry, V., Moszer, I., Dameron, O., Potier, M.C., Hofmann-Apitius, M., and Colliot, O. (2018). Converting Alzheimer s disease map into a heavyweight ontology: a formal network to integrate data. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) *11371 LNBI*, 207–215 https://doi.org/10.1007/978-3-030-06016-9_20