D8.9.2 Curated and tested brain disease ontologies integrated within application services (D8.9.2 – SGA2)

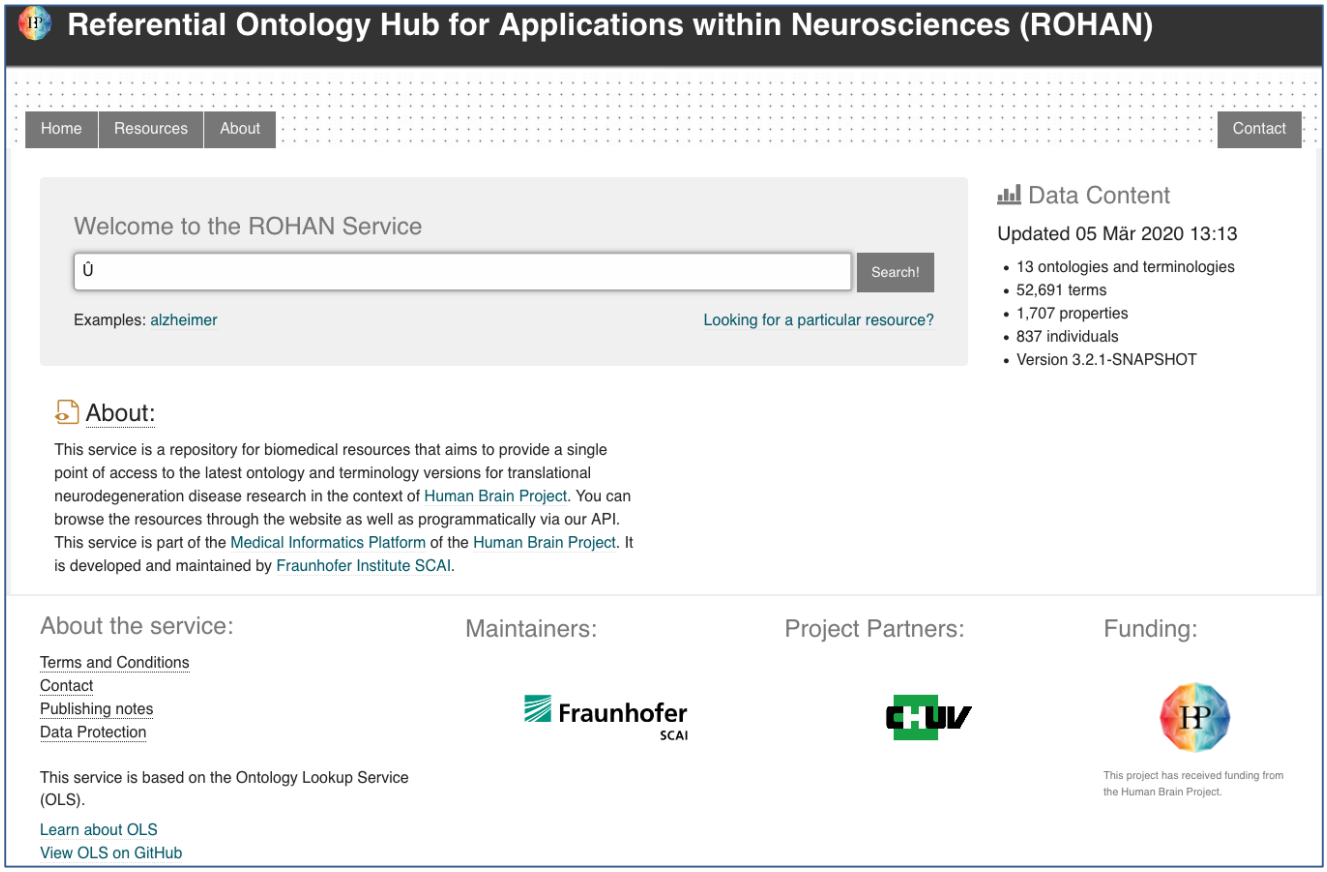

Figure 1: Referential Ontology Hub for Applications within Neurosciences (ROHAN)

Project Number:	785907	Project Title:	Human Brain Project SGA2

Document Title:	Curated and tested brain disease ontologies integrated within application services
Document Filename:	D8.9.2 (D56.2 D126) SGA2 M24 ACCEPTED 200731.docx
Deliverable Number:	SGA2 D8.9.2, D56.2
Deliverable Type:	Demonstrator
Work Packages:	WP8.9
Key Result(s):	KR8.3, KR8.5
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA2 M24 / 31 Mar 2020
Actual Delivery Date:	SGA2 M24 / 27 Mar 2020; Accepted 31 Jul 2020
Author(s):	Sumit MADAN, FRAUNHOFER (P22)
Compiled by:	Sumit MADAN, FRAUNHOFER (P22)
Contributor(s):	Sumit MADAN, FRAUNHOFER (P22), contributed to whole document Johannes DARMS, FRAUNHOFER (P22), contributed to whole document Stephan GEBEL, FRAUNHOFER (P22), contributed to whole document Martin HOFMANN-APITIUS, FRAUNHOFER (P22), contributed to whole document
SciTechCoord Review:	Mehdi SNÈNE, EPFL (P1)
Editorial Review:	Annemieke MICHELS, EPFL (P1)
Description in GA:	Brain ontologies are integrated and accessible through OLS-NEURO. All application services (OLS-NEURO, SCAIView-Neuro) are online and available for use in MIP.
Abstract:	The key objective of the Work Package WP8.9 “Comprehensive ontologies for brain diseases” is to deliver a semantic framework plus relevant services that serve as a central resource for controlled vocabularies and shared ontologies for translational neurosciences. The particular focus of this project lies in the curation of brain disease ontologies that are relevant to the Human Brain Project. The ontologies comprise of the majority of relevant entities and their relationships that are required to describe data and knowledge about brain diseases. Furthermore, the developed ontologies are included in an ontology store – named as Referential Ontology Hub with Application Services for Neurosciences (ROHAN) - that provides end users the browsing and searching ability. In addition, a dedicated literature mining environment for brain research is part of the semantic framework.
Keywords:	Ontology Hub, Neuroscience, Semantic Framework, Brain Disease Ontologies, deep learning, text mining, ROHAN
Target Users/Readers:	computational neuroscience community, computer scientists, Consortium members, ontology experts, HBP community, neuroinformaticians, neuroscientists, platform users, researchers, scientific community, students

Table of Contents

Table of Figures

1. Introduction

The key objective of WP8.9 "Comprehensive ontologies for brain diseases" is to deliver a semantic framework plus relevant services that serve as a central resource for controlled vocabularies and shared ontologies for translational neurosciences. The particular focus of this project lies in the curation of brain disease ontologies that are releveant to the Human Brain Project. The ontologies comprise the majority of relevant entities and their relationships that are necessary to describe data and knowledge about brain diseases. Furthermore, the developed ontologies are available within an ontology store – named *Referential Ontology Hub with Application Services for Neurosciences* (ROHAN) - that allows end users to browse and search these resources. In addition, a dedicated literature mining environment for brain research is part of the semantic framework.

The Work Package contains three Components, entitled "*Ontologies updated and curated*" (Component ID 3071[1]), "Ontology lookup service (OLS-NEURO) for the Human Brain Project infrastructure" (Component ID 3072[2],), and "SCAIView-NEURO, a dedicated literature-mining service for the Human Brain Project" (Component ID 3073[3]).

2. Curation of Brain Disease Ontologies

2.1 Brain Disease Ontologies

In the context of ROHAN, four brain-specific ontologies are being developed, namely, the Association for Methodology and Documentation in Psychiatry (AMDP) ontology, the epilepsy ontology, the schizophrenia ontology, and the clinical trial ontology. **The AMDP ontology** (see Figure 2), created in collaboration with University Clinic RWTH Aachen, mainly contains classes that are used in psychiatry to document psychopathology of patients. The ontology has been created by extracting classes, their definitions, and the associated hierarchy from the AMDP System, which is published as a book[4]. The ontology has been extended with further terms that are provided by University Clinic RWTH Aachen. As the interest of partner RWTH Aachen is in using the ontology for text mining scenarios (e.g. extracting real-world evidences from electronic health records, also called EHRs), the ontology is currently available only in German. We foresee further uptake of this ontology in Germany (e.g. in the course of the project "Commitment" (http://www.sys-med.de/en/alliances/commitment/)) and would be interested in collaborating with other EU partners on generating French, English and Spanish versions of the ontology. Note that a multilingual form of any ontology representing knowledge in the psychiatry domain is absolutely non-trivial.

An application case for this ontology has been developed, in which psychopathogical reports (PPR) are being analysed. Often the medical experts describe psychiatric and somatic symptoms (also called items) and their assessment (for e.g. pathological or normal) in textual psychopathological reports. These reports have been prepared for a text mining task. In a first step, medical experts have annotated the items and their assessment in the reports. Consequently, the items that are reported as pathological have been linked to the AMDP terminology. Figure 3, Figure 4, and Figure 5 show three examples of the annotation extracted from two different pyschopathological reports. The first example (Figure 3) contains items assessed as non-pathological. The second example (Figure 4) contains three items that are assessed as pathological. The third example highlights the linkage of the item "Konzentration" (English: concentration) and its pathological assessment "herabgesetzt" (English: degraded) to the AMDP concept "Konzentrationsstörungen" (ID: AMDP:10) (English: concentration disorders). In total, 150 documents have been annotated by the experts.

[1] https://plus.humanbrainproject.eu/components/3071/
[2] https://plus.humanbrainproject.eu/components/3072/
[3] https://plus.humanbrainproject.eu/components/3073/
[4] Das AMDP-System – Manual zur Dokumentation psychiatrischer Befunde, 10th ed, 2018. Arbeitsgemeinschaft für Methodik und Dokumentation.

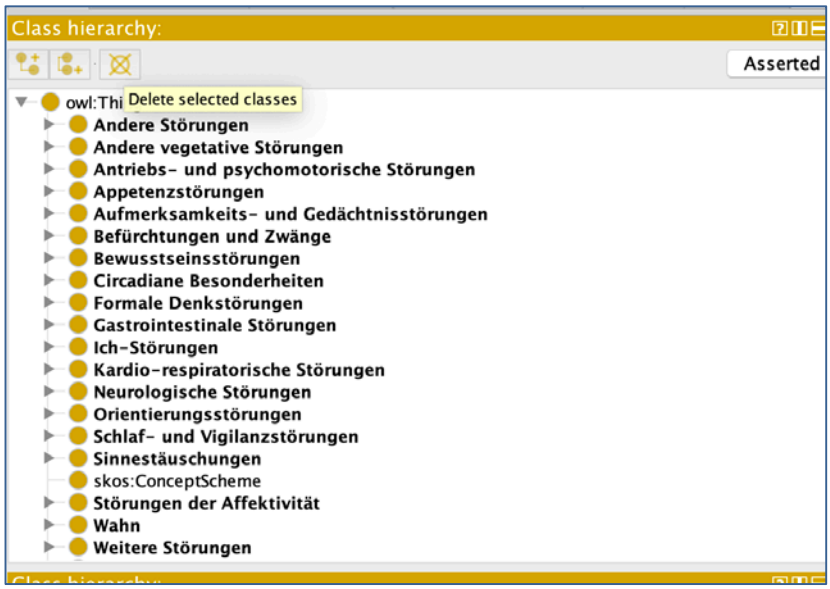

Figure 2: Screenshot of the newly developed AMDP ontology

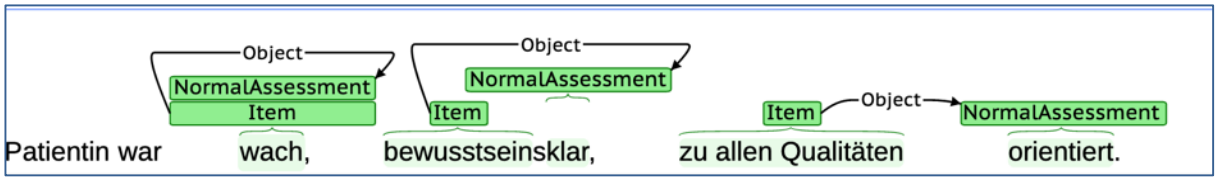

Figure 3: Annotation example 1

A sentence annotated by medical experts with the entity classes *Item* and *NormalAssessment*.

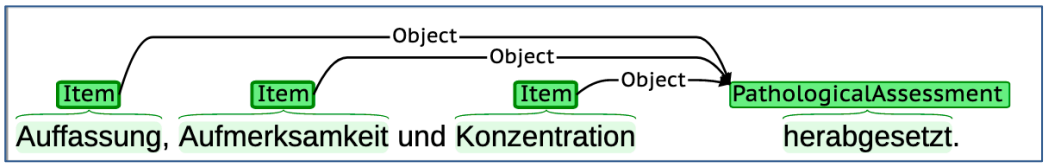

Figure 4: Annotation example 2

A sentence annotated by medical experts with the entity classes *Item* and *PathologicalAssessment*.

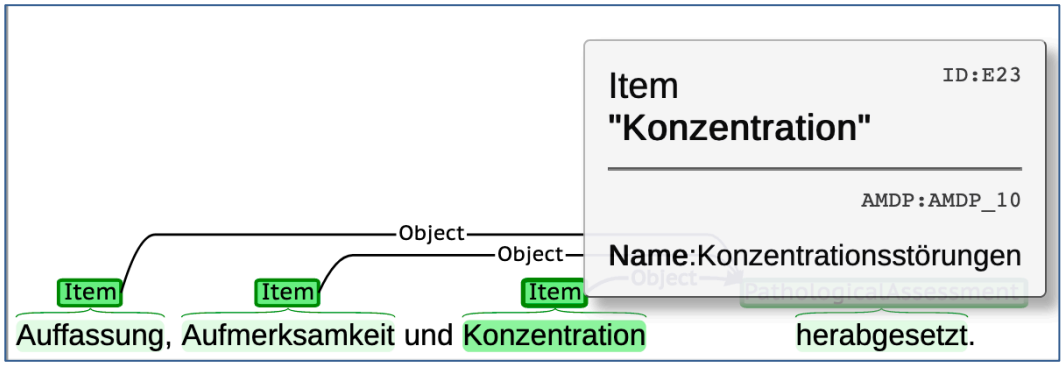

Figure 5: Annotation example 3

An annotated sentence that links the Item "Konzentration" (English: concentration) and its pathological assessment "herabgesetzt" (English: degraded) to the AMDP concept "Konzentrationsstörungen" (ID: AMDP:10) (English: concentration disorders).

A text mining workflow utilising deep learning has been developed to detect the three entity classes in new PPRs. To train a deep-learning model, a pre-trained language understanding model, called GermanBERT[5], has been extended with an additional neural network layer that is fine-tuned with

[5] https://deepset.ai/german-bert

the PPR training set developed by the medical experts. Training has been performed on a GPU compute node by using the Python framework PyTorch. The first scores that have been achieved, yet without any hyperparameter optimisation, are convincing. The models that predict the mentions of the 3 given classes *item*, *normalAssessment*, and *pathologicalAssessment* have achieved an F-Score of 81%, 88%, and 78%, respectively. Further experiments to improve the prediction of the models are ongoing.

The **epilepsy ontology** (see Figure 6) represents a semantic assembly of structured knowledge on various aspects of epilepsy. The foundation of the ontology is built by representing entities and definitions [6] from the latest International League Against Epilepsy (ILAE). Furthermore, other domain-related ontologies like Epilepsy Syndrome Seizure Ontology (ESSO)[7], Epilepsy and Seizure Ontology (EPSO)[8], Epilepsy Ontology (EPILONT)[9], and information from scientific literature were incorporated in this ontology. The structure of the ontology is based on 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 epilepsy ontology is based on the Protege ontology web language (OWL) editor; hence the ontology is directly available in OWL format, a standard format for publishing and sharing ontologies.

Figure 6: Screenshot of the hierarchy of the newly developed epilepsy ontology

The **schizophrenia ontology** (see Figure 7) puts the focus on modelling the domain of the schizophrenia disease. It contains the schizophrenia subtype classification and associated sub-concepts as well as integrates schizophrenia-related concepts from existing psychiatric and mental disease-related ontologies such as mental-functioning ontology or emotion ontology. Furthermore,

[6] https://www.ilae.org/guidelines/definition-and-classification
[7] https://bioportal.bioontology.org/ontologies/ESSO
[8] https://bioportal.bioontology.org/ontologies/EPSO
[9] https://bioportal.bioontology.org/ontologies/EPILONT

the ontology has been further enriched and updated with the information available from scientific journal articles and schizophrenia related websites.

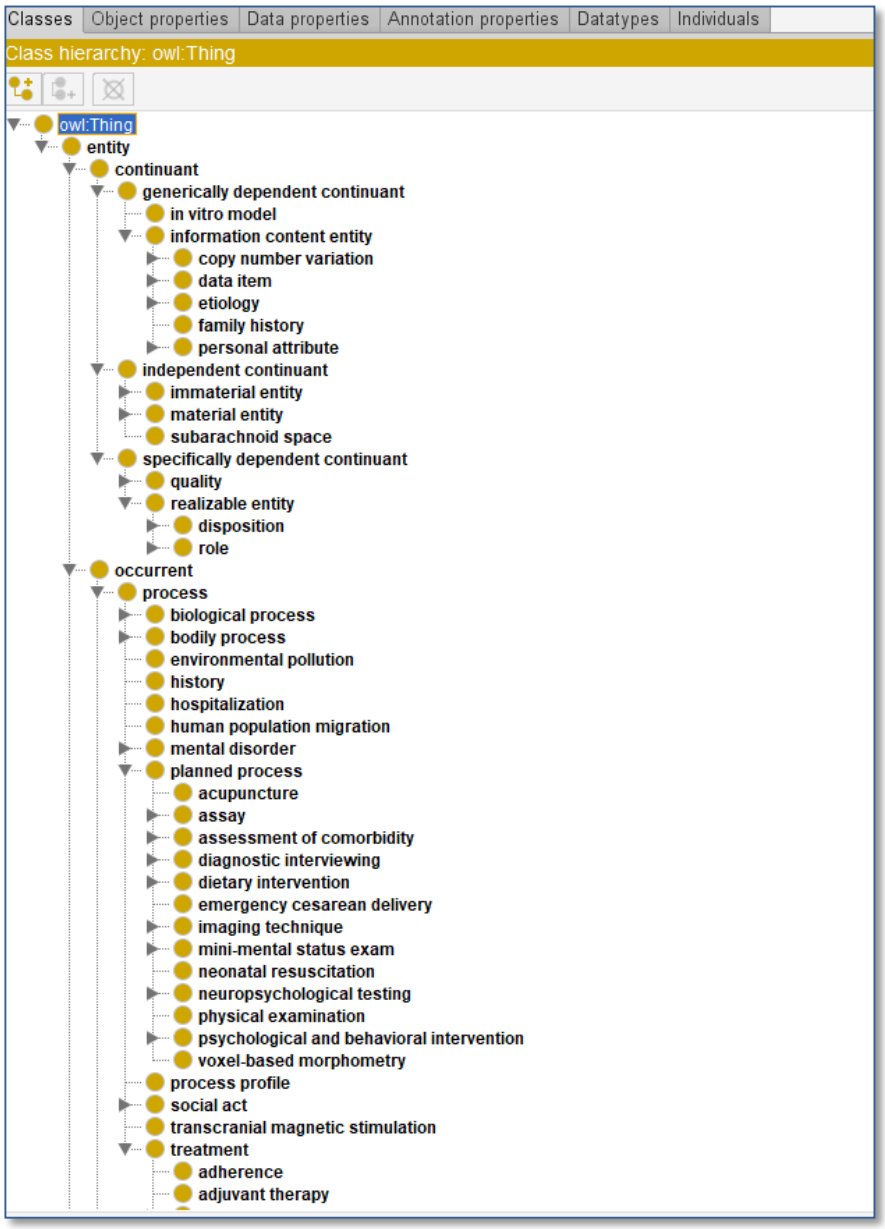

Figure 7: Screenshot of the hierarchy of the newly developed schizophrenia ontology

To model the field of clinical trials, the Clinical Trial Ontology for Neurodegeneration Diseases (CTO-NDD) was developed for the IMI-funded AETIONOMY[10] project. During the ROHAN project, this ontology has been updated. The updated version contains now two ontologies, a coreCTO and an extendedCTO. The **core Clinical Trial Ontology (CTO)** will serve as a structured resource integrating basic terms and concepts in the context of clinical trials, thereby covering clinicaltrails.gov. It is noteworthy to mention that the coreCTO is being developed in a joint effort of international partners, including the NCBI (National Center of Biotechnology Information), the FDA (Food and Drug Administration) and the University of Michigan. Whereas, the **extended version of CTO** uses the coreCTO as a basic ontology to support implementation of specific applications, such as annotation of variables in clincal study documents from neurodegeneration disease related clinical trials.

The current version of the ontologies is available at:

- AMDP ontology: https://rohan.scai.fraunhofer.de/ols/ontologies/amdp

[10] https://www.aetionomy.eu/

- Epilepsy ontology: https://rohan.scai.fraunhofer.de/ols/ontologies/epo
- Schizophrenia ontology: https://rohan.scai.fraunhofer.de/ols/ontologies/schizo
- Clinical Trial Ontology (core and extended version): https://rohan.scai.fraunhofer.de/ols/ontologies/cto and https://rohan.scai.fraunhofer.de/ols/ontologies/ecto

2.2 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"[11]. Hence, where possible, we have adopted widely-used best practices. Furthermore, to make the curation reproducible, several annotations have been added to the ontologies. For example:

- **rdfs:label** annotation defines the name of each concept.
- **iao:definition (IAO:0000115)** property contains the definitions for a concept.
- **rdfs:isDefinedBy** contains a reference to the source of the definition, if the definition couldn't be imported from an existing ontology. This annotation is added to **iao:definition** itself, so each defintion can be traced to the defining source**.** This is necessary as a single concept may have multiple defintions.
- **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. All guidelines and definitions, as well as the original OWL representations of the ontologies, will be published in dedicated papers. All of our work output is open access and can be freely used by the scientific community.

3. Application Services

3.1 OLS-Neuro, an Ontology Lookup Service for Translational Neurodegeneration Research

The OLS-Neuro service is based on the software 'Ontology Lookup Service'[12] (OLS) that is developed by European Bioinformatics Institute (EBI) for exploring ontologies. The service provides a web-based user interface for exploring and visualising the ontologies (Figure 8), and additionally, a flexible RESTful API to access the resources programmatically. It also provides a utility that allows to regularly update ontologies with ease. Furthermore, the service includes a search index for terms and synonyms with autocomplete functionality (Figure 9). To manage ontologies, the service also uses a flexible configuration system.

Availability: ROHAN OLS-Neuro is available at http://rohan.scai.fraunhofer.de/ols.

[11] http://www.obofoundry.org/
[12] Source code is available at https://github.com/EBISPOT/OLS

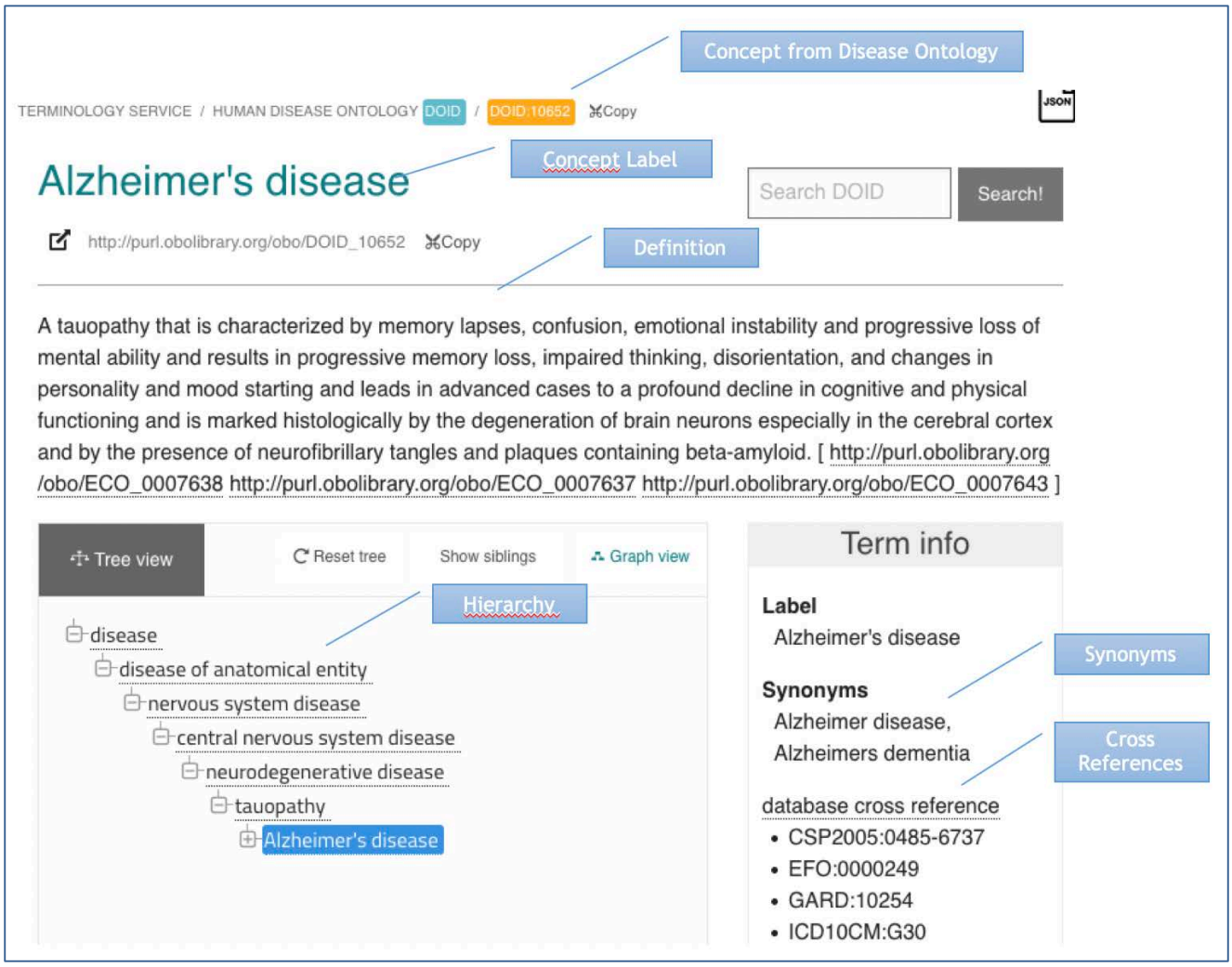

Figure 8: Screenshot of a concept page from OLS

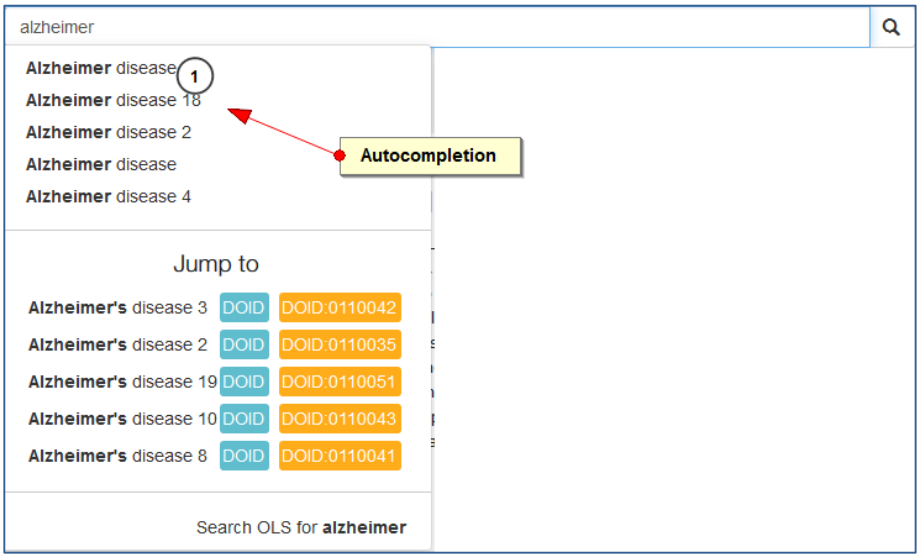

Figure 9: Screenshot of the autocomplete functionality of OLS

3.2 SCAIView-Neuro, a Semantic Search Engine

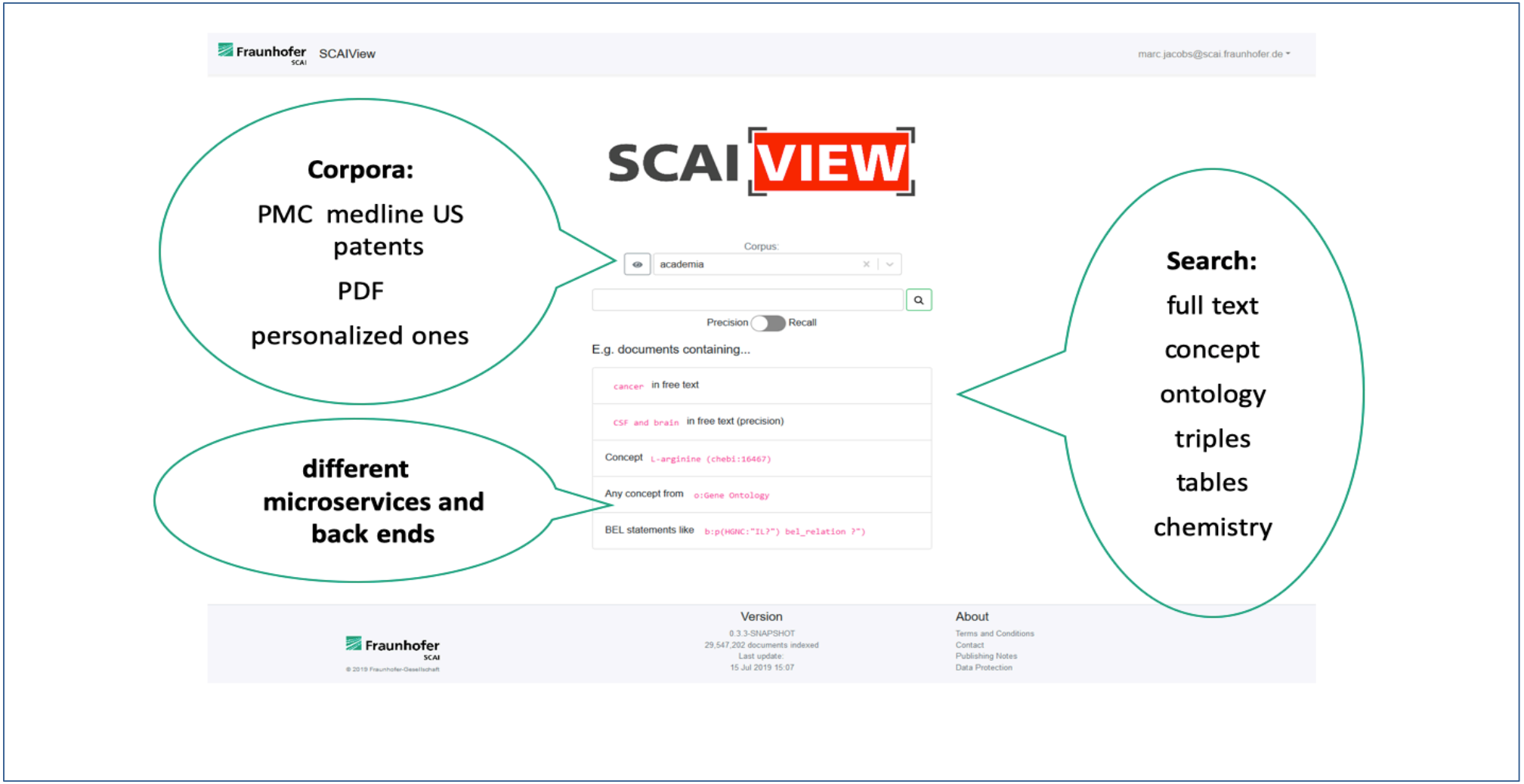

Figure 10: User Interface and the functionality of the semantic search engine SCAIView

SCAIView is an advanced semantic search and knowledge retrieval 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 identifies and displays information about genes, drugs, SNPs and other life science entities. Various new functionalities have been developed together with corresponding view components.

Corpus selection and advanced search utility: SCAIView provides access to various corpora, for instance, it includes documents from PubMed, PubMed Central databases. Using the SCAI text mining pipeline, documents from these corpora are annotated with concepts that are derived from the various integrated ontologies. The search and query utility allows a user to execute 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.

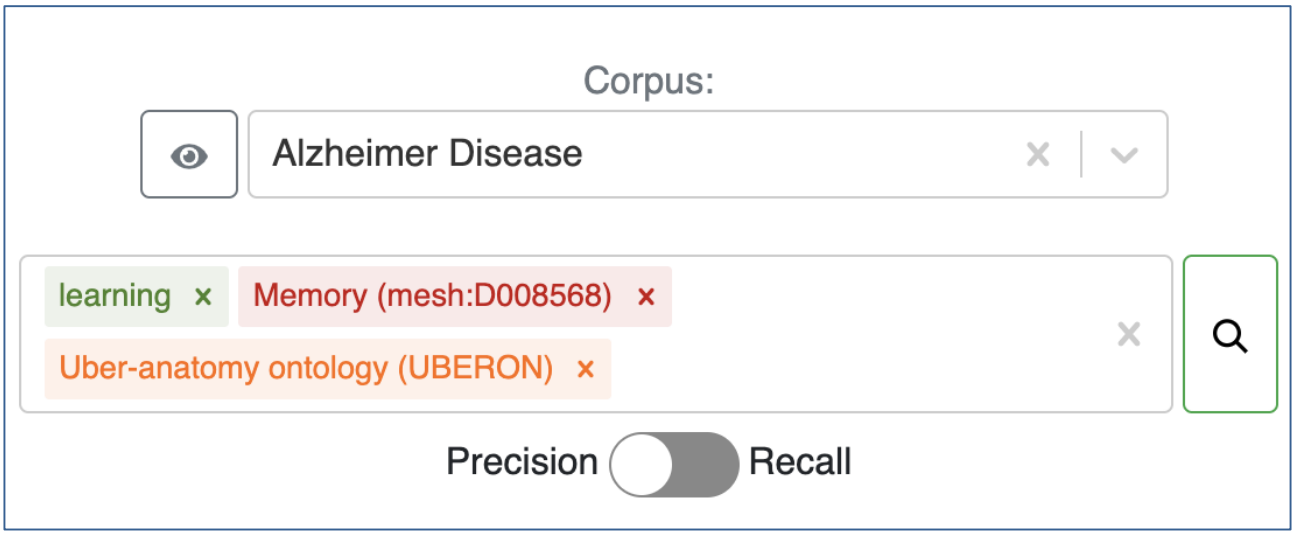

Figure 11: Query example

Query example: one can search for 'learning' as a free text (green) term, memory as a concept (red) and 'Uber-anatomy ontology' as an ontology (orange) in the 'Alzheimer's Disease' literature collection. As a result, SCAIView will deliver documents from PubMed database that contain the queried terms and concepts.

Document list and network views: Two further views are available to visualise the results of a search query. Figure 12 shows the document list view that highlights that the search query retrieved 843 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. The network view (see Figure 13) for a search query gives the user the ability to overview and investigate the associated MeSH terms tagged in the retrieved document list.

Full text view: Visualisation of a full-text view of a document that highlights text mining results (annotations) using different colours for each ontology or terminology. It has been developed specifically to visualise the full-text of a single document. Figure 14 shows a screenshot of a PubMed article with its corresponding text mined results.

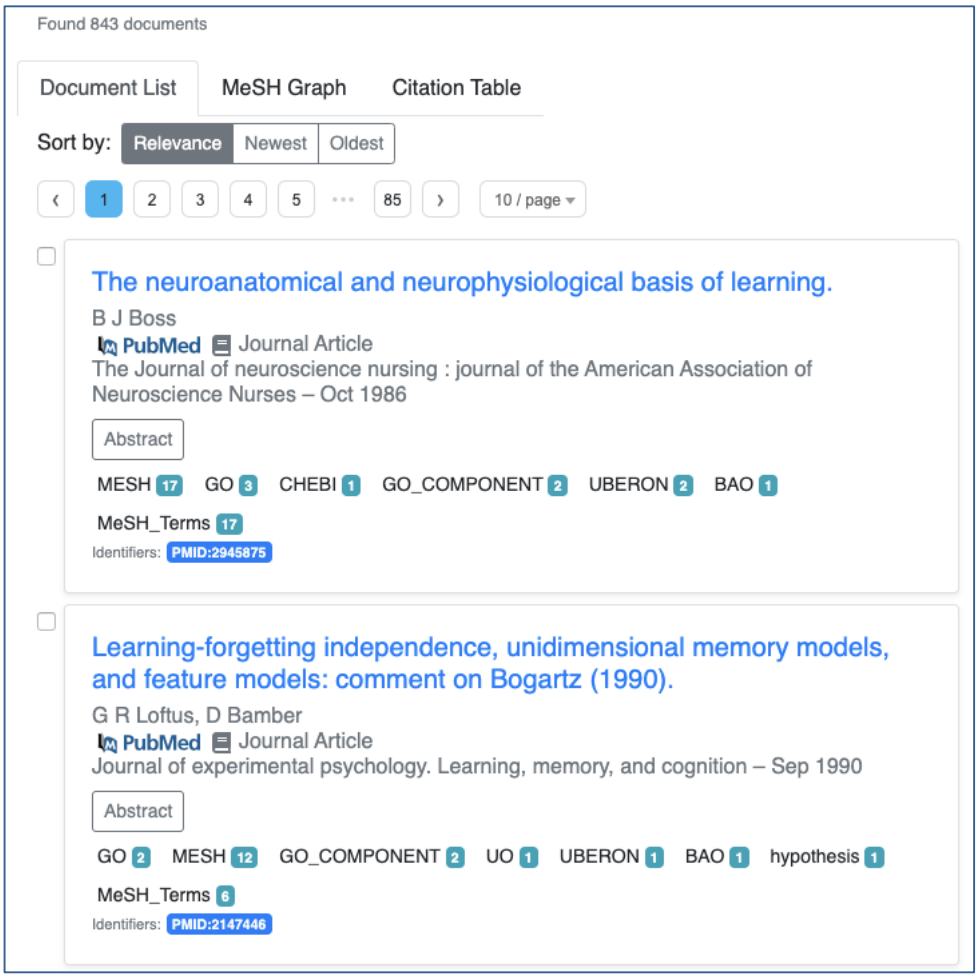

Figure 12: Document list view of SCAIView.

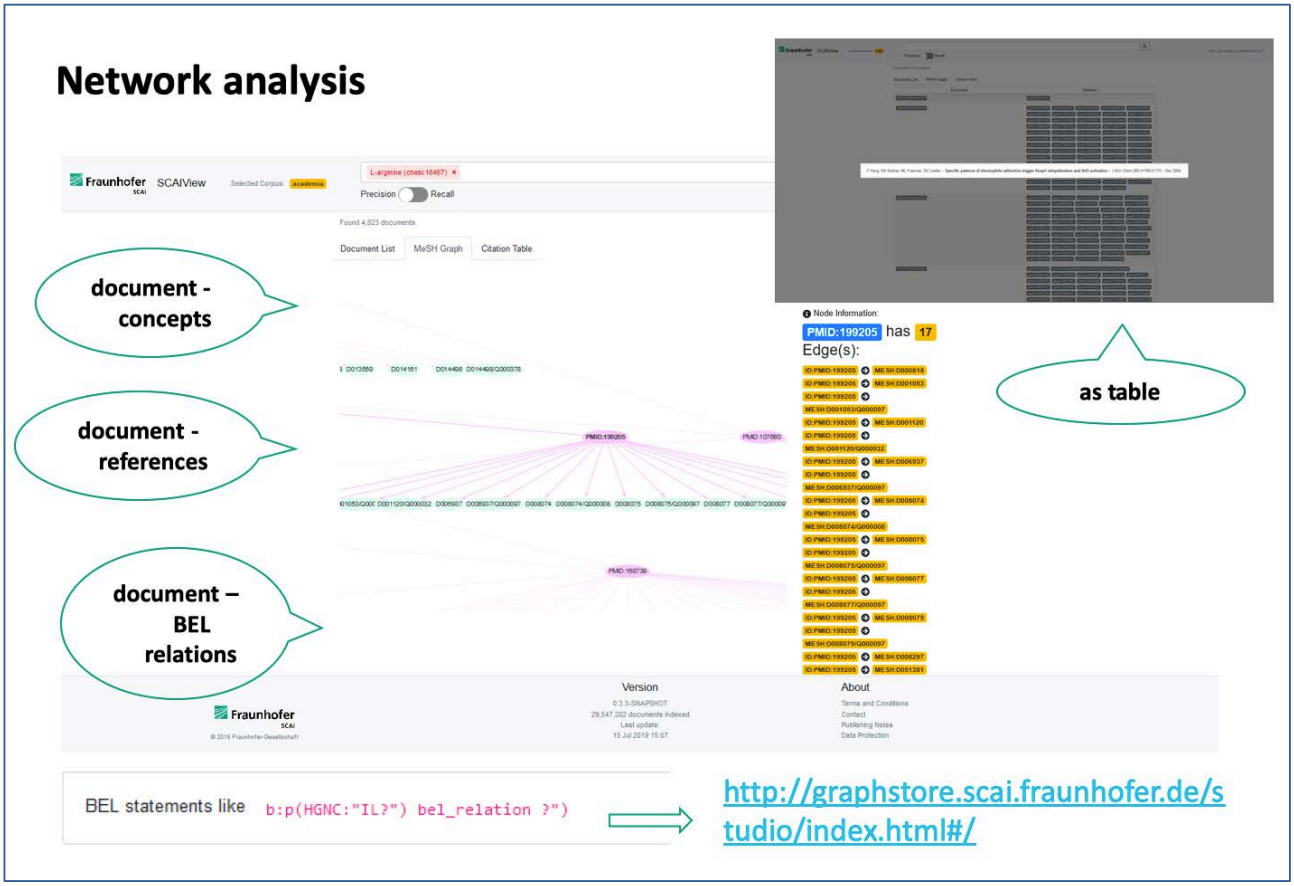

Figure 13 Network view of SCAIView

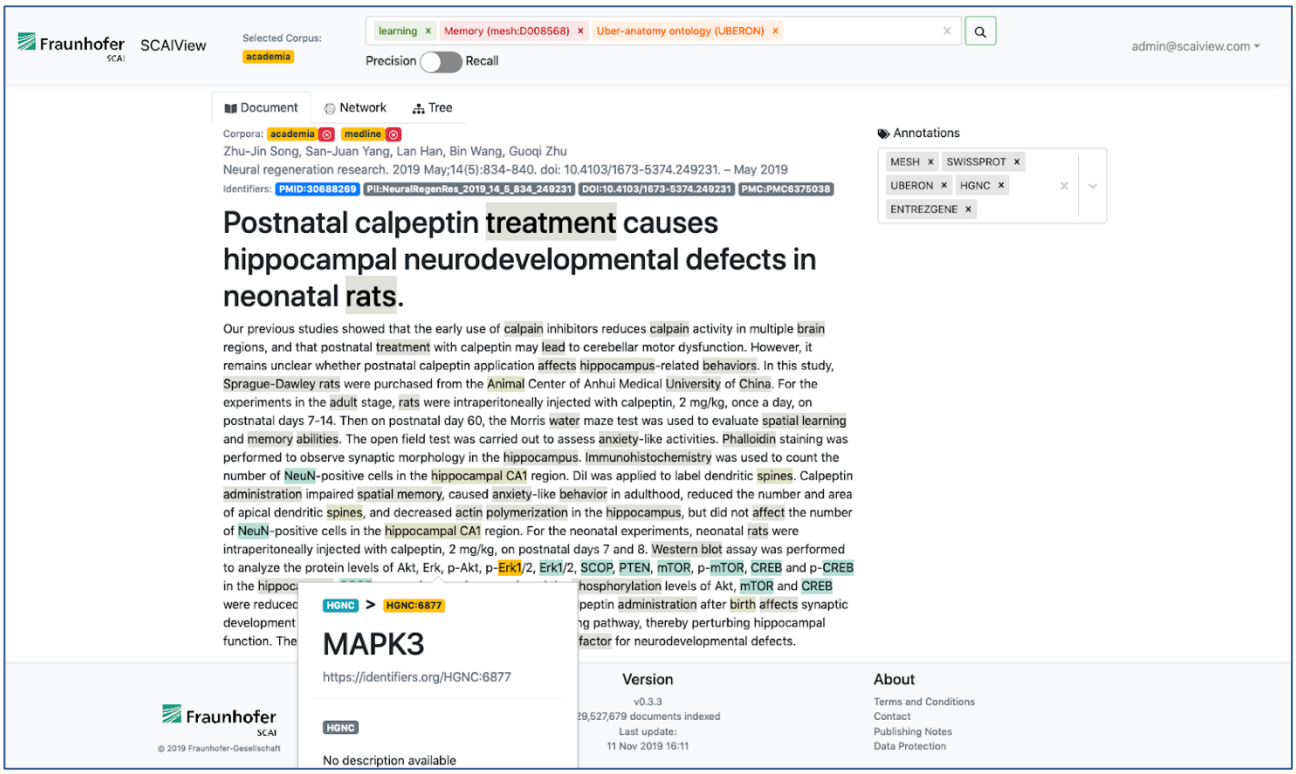

Figure 14: Full-text view of the PubMed article with highlighted annotations and the concept hierarchy viewer

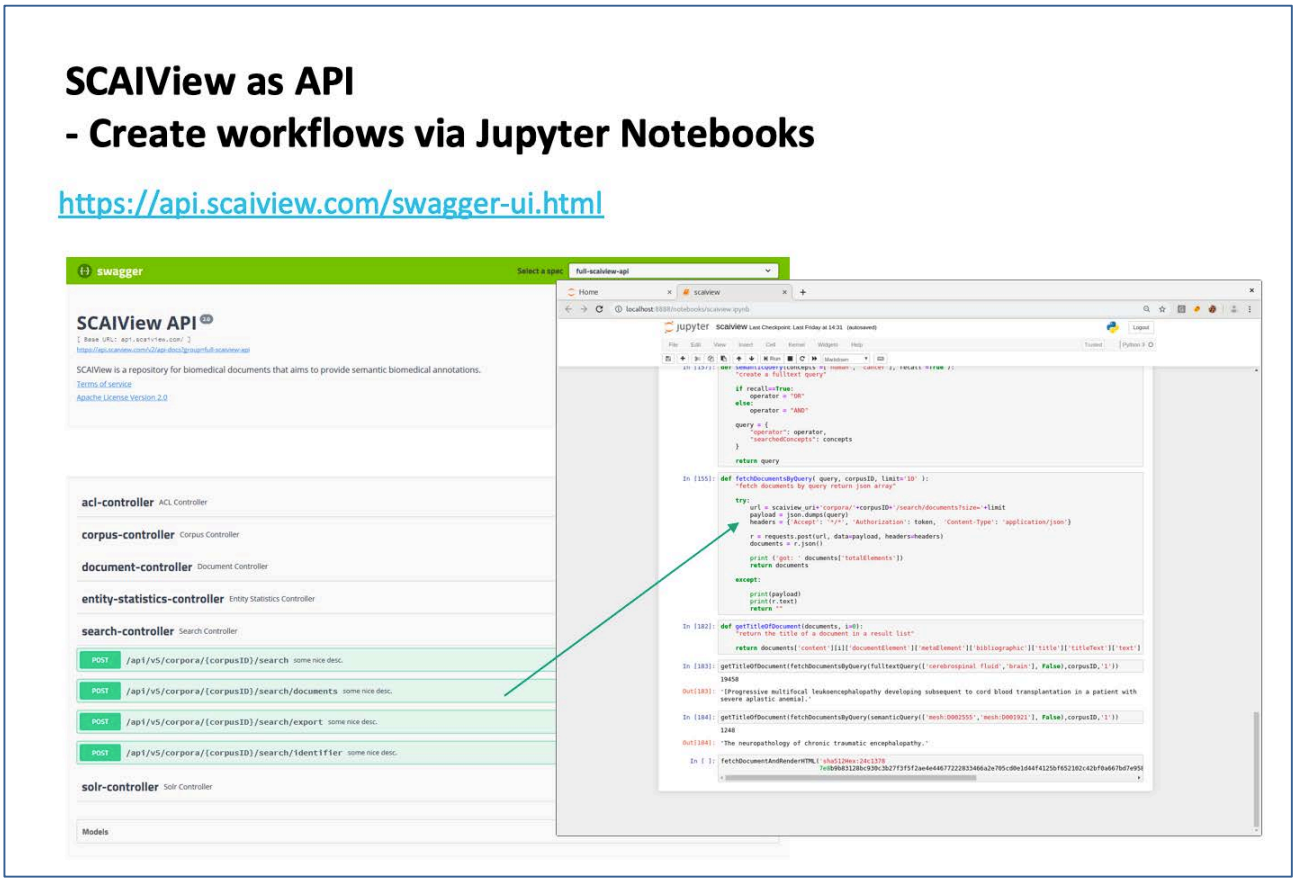

Figure 15: Exemplary usage of SCAIView API in a Jupyter Notebook.

SCAIView API: The API of SCAIView (available at https://api.scaiview.com/swagger-ui.html) provides programmatic access to the above-mentioned functionalities. The API is also consumed by the several SCAIView view components. The search-controller API can be used to search and retrieve documents for a given search query and a corpus ID, whereas, the document-controller API is being

used to retrieve title, abstract, annotations, authors etc. by providing a document ID. Various other SCAIView functionalities can be accessed through further API controllers. We have also created a Jupyter notebook especially to demonstrate the SCAIView functionalities by including examples of full-text queries, semantic queries and further API methods (see Figure 15).

Authentication and Authorisation: The access to SCAIView UI and API is secured with the OpenID Connect (OIDC) authentication standard. Only authorised 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.

Availability: SCAIView and SCAIView API are available at https://ui.scaiview.com and https://api.scaiview.com/. Registration is needed to access the SCAIView service and the API.

Both services can also be accessed through Medical Informatics Platform (MIP). A deeper integration of OLS-NEURO in MIP to harmonize the clinical variables is the next step. In future discussions with the HBP Knowledge Graph team at EPFL (P1), we intend to bridge between the "Referential Ontology Hub for Applications within Neuroscience (ROHAN)" and the metadata graph developed by partner EPFL. Despite the substantial difference between the Knowledge Graph in its current form and the concept we present here, we are confident that an invocation of OLS-NEURO and SCAIView-NEURO services by the Knowledge Graph should be possible. As a consequence, the OLS-NEURO system and its link to text-mining services bear the potential to dramatically increase the usability of the Knowledge Graph services.