A Semantic Data Integration Methodology for Translational Neurodegenerative Disease Research

Sumit Madan[1], Maksims Fiosins[2,3], Stefan Bonn[2,3], and Juliane Fluck[1,4,5]

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schoss Birlinghoven, Sankt Augustin, Germany
[2] German Center for Neurodegenerative Diseases, Tuebingen, Germany
[3] Institute of Medical Systems Biology, Center for Molecular Neurobiology, University Medical Center Hamburg-Eppendorf, Germany
[4] German National Library of Medicine (ZB MED) - Information Centre for Life Sciences, Bonn, Germany
[5] Institute of Geodesy and Geoinformation, University of Bonn, Germany
{sumit.madan,juliane.fluck}@scai.fraunhofer.de

Abstract. The advancement of omics technologies and execution of large-scale clinical studies have led to the production of heterogeneous and big patient datasets. Researchers at DZNE (German Center for Neurodegeneration Diseases) and Fraunhofer SCAI (Fraunhofer Institute for Algorithms and Scientific Computing), located at several sites, are focusing on generation, integration, and analysis of such data, especially related to the field of neurodegenerative diseases. In order to extract meaningful and valuable biological insights, they analyze such datasets separately and, more importantly, in a combined manner. Blending of such datasets, which are often located at different sites and lack semantical traits, requires the development of novel data integration methodologies. We use the concept of federated semantic data layers to disseminate and create a unified view of different types of datasets. In addition to the semantically-enriched data in such data layers, we propose another level of condensed information providing only derived results that is integrated in a central integration platform. Furthermore, the implementation of a semantic lookup platform encloses all semantic concepts needed for the data integration. This eases the creation of connections, hence, improves interoperability between multiple datasets. Further integration of biological relevant relationships between several entity classes such as genes, SNPs, drugs, or miRNAs from public databases leverages the use of existing knowledge. In this paper, we describe the semantic-aware service-oriented infrastructure including the semantic data layers, the semantic lookup platform, and the integration platform and, additionally, give examples how data can be queried and visualized. The proposed architecture makes it easier to support such an infrastructure or adapt it to new use cases. Furthermore, the semantic data layers containing derived results can be used for data publication.

Keywords: Semantic Data Integration, Semantic Data Layer, Translational Research, Neurodegenerative Diseases.

CEUR-WS.org/Vol-2275/paper9.pdf

1 Introduction

Translational medicine in a disease area aims to shorten the time between new scientific findings in laboratories to new therapies for patients. Especially in the area of neurodegenerative diseases and dementia, a fast translation is necessary to reduce the suffering of the patients and their families and the economic burden on the society of a growing ageing population. In industrialized countries, late-stage dementia and complications due to underlying dementia have become the most common causes of death besides heart diseases, malignant growth, and cerebrovascular diseases and currently affects 1.6 million people in Germany [1]. As the population ages, the number of people suffering from dementia and related neurodegenerative disorders (NDD) will substantially rise. To date, all attempts to slow down disease progression by medical (e.g. pharmacological) or non-medical interventions have failed.

To address these challenges, the German Center for Neurodegenerative Diseases (DZNE) was founded in 2009 as an institute of the Helmholtz association. The DZNE has ten sites distributed over Germany that integrate the leading national expertise in the field of neurodegeneration research. DZNE covers a wide range of research topics from fundamental research over clinical to health care and population research. Its broad scope enables the DZNE to follow a translational approach with the ultimate goal to develop novel preventive or therapeutic solutions for neurodegenerative diseases. A current bottleneck in analysing the heterogeneous data generated at the distributed DZNE sites is that different data entities for the same disease or even the same patient are analysed separately and the full potential of a holistic analysis of all data is not leveraged. The key aim of the BMBF-funded project *Integrative Data Semantics for Neurodegeneration research* (IDSN) (www.idsn.info/en/) is the ability to integrate and query data from the different DZNE research fields and combine this with existing disease information and biomedical databases.

To achieve coherent data integration, several general and NDD-specific data integration tasks have to be addressed. In general, there is a need to integrate large-scale data coming from high-throughput screening, clinical cohort and/or clinical routine data. Other large-scale data becoming standard in many disease fields are for example automated cellular assays or imaging data. Many more data types will be standard in the future. On the other hand, task-specific data types vary significantly depending on the use case and the disease area, and annotation of data and metadata is needed in such a way that they are interoperable and can be reused. These demands are well described as requests within the FAIR data principles [2].

To cope with these diverse requirements, we present a novel semantic integration methodology for linked biological and clinical data. We realize the architecture by using existing open-source tools in concordance with identified requirements and describe the technical details of the implementation. Key elements of the presented integration platform are (1) a central semantic lookup platform for the vocabulary used within, (2) the modularity of its components, (3) the semantic integration of the different data types and (4) their compliance with the FAIR data principles, (5) a data integration platform for different types of data, and finally (6) query environments allowing for integrative analysis of data by end-users such as clinicians or researchers.

In the following we give an overview of related approaches, describe the IDSN architecture and give examples how to integrate the data and how this data can be queried and visualised.

2 Related Work

Several studies demonstrated the usefulness of data integration in various ways. Daemen et al. [3] were able to enhance the predictive power of clinical decision support models that were used to define personalized therapies for rectal and prostate cancer patients. They used a methodological integration framework while incorporating data from different genome-wide sources including genomics, proteomics, transcriptomics, and epigenetics.

Dawany et al. [4] used a large-scale integration of complex data from high-throughput experiments (HTP) to identify the shared genes and pathways in different cancer types. The integration and normalization methodology was applied on microarray data from more than 80 laboratories with more than 4000 samples that included more than 10 cancer tissue types. For each cancer type an organized list of genes was identified as the potential biomarkers including various kinases and transcription factors.

Iyappan et al. [5] have employed RDF-based technologies to link and publish large volumes and different types of neuroscience related data. They have transformed the data into simple triple format (subject, predicate, and object) that represents relationships between entities. As usual in RDF, the nodes and the edges are encoded using Uniform Resource Identifiers (URIs) that are provided by biological-specific ontologies. They have integrated semantically-enriched data such as PPI networks, miRNA-Target-Interaction networks, transcriptomic data (from GEO and ArrayExpress) and relationships from further biological databases. Although RDF-based technologies are well suited for data interoperability, they are not suitable for every dataset type. Furthermore, they often lack performance and consume large amount of disk space with huge datasets.

In the Open PHACTS discovery platform, RDF and especially Application Programming Interfaces (APIs) is extensively used for designing and development of linked data applications with respect to integrating pharmacology data [6]. The API layer provides output in JSON format for the application developers, hiding RDF that is considered complex for the purpose of user interaction and data presentation. The Open PHACTS discovery platform provides integrated access to more than eleven linked datasets that cover information about chemistry, pathways, and proteins.

The Neuroscience Information Framework (NIF), published in 2008 and initiated by National Institutes of Health Blueprint for Neuroscience Research, is an ecosystem that provides a repository of searchable neuroscience resources [7]. Experimental databases, brain atlases, neuroscience-specific literature, commercial tools and several other data types are supported by NIF. It provides access to the data through a single web-based platform that is mainly available for finding such resources.

3 Architecture

The main purpose of the IDSN architecture presented in Figure 1 is to provide a modular platform for the integrated analysis of data supporting translational neurodegenerative disease research. The data is derived from clinical routine and cohort data or from different high-throughput analysis of biomaterial and is originated at distributed sources. The architecture supports different types of data to be stored and integrated and should be easily extendable to new data types.

In a nutshell the ***primary data***, that is stored in distributed or federated data storages, is analysed locally to derive analysis results and is supplied with semantic annotations using standard ontologies and terminologies to make data interoperable. As a result, a standardized ***semantic layer*** for the different data sources is generated, where standardised vocabulary ensures data interoperability. Links to the original data are incorporated to ensure data provenance.

The used ontologies and terminologies are stored in a separate ***semantic lookup platform***, which is used to retrieve appropriate concepts to annotate data, to provide information about the concepts, and to provide mappings between different ontologies and terminologies. This service is not only used from the federated data storage systems but from the semantic linked data hub as well.

The semantically enriched data is transferred as semantic layer to a ***data management platform***. The data management platform is a component responsible for unified data access to all semantic data layers. Furthermore, the data management platform ensures that data from all semantic layers correspond to common platform standards (for example, FAIR principles) and that data consistency is checked.

The ***semantic linked data hub***, which fetches and indexes the data from the data management platform, is the central part of the IDSN architecture. It stores the data in various appropriate formats to allow fast data queries and retrieval. In addition, further external data (secondary data) is added within the semantic linked data hub to provide additional information about primary data, or additional links (associations) between primary data elements. Furthermore, external services can be used by the semantic linked data hub to analyse the data or to provide background information. Finally, for visual and further computational analysis graphical user interfaces allow for dedicated data visualisation and interactive analysis.

In this service-oriented architecture, the key functionality of each service is accessible and consumed through a well-defined, rich API that conforms to the popular Representational State Transfer (REST) paradigm. Plugging of available services in this architecture is established through systematic adoption and reuse of the existing APIs, which, to match the needs, can be subjected to extensions. The newly-developed API for the semantic data integration platform, which provides programmatic access to the integrated data, is designed to answer common scientific queries of the users.

Currently, the IDSN platform can handle three main types of bio-medical data, which correspond to the following semantic layers:

- omics data layer contains expression data of small RNAs and RNAs, as well as mutation variants from healthy and NDD subjects.

- pharmacological (assay) data layer contains compound activity rates for the induction of various cellular processes such as apoptosis or the induction of protein expression such as CASP3 induction.
- clinical data layer includes longitudinal clinical routine and longitudinal cohort data from healthy and NDD subjects.

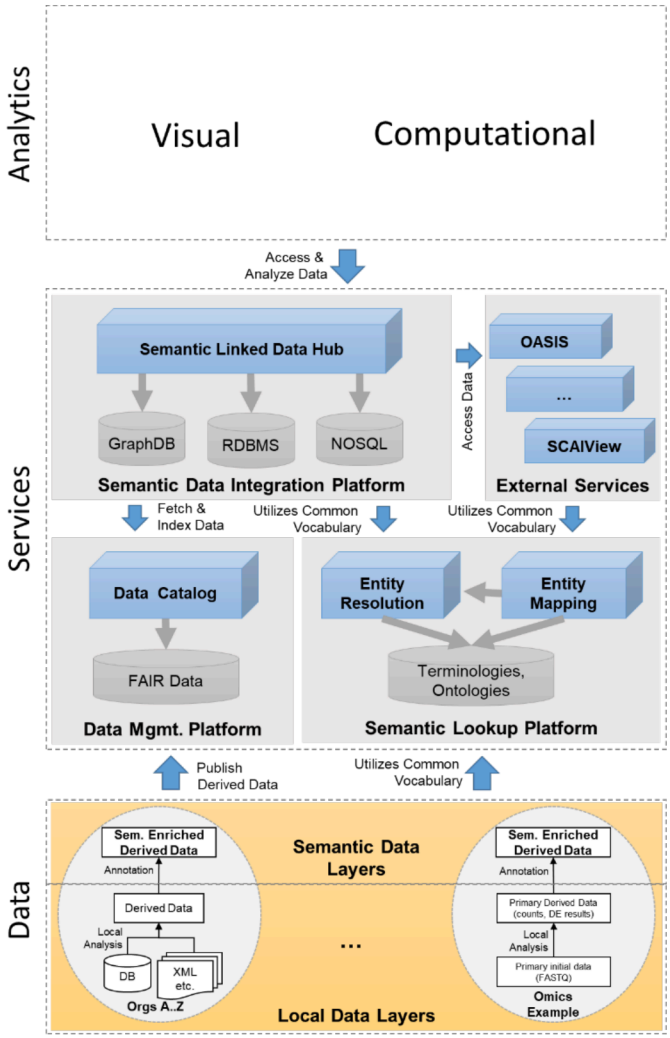

Figure 1. General architecture for semantic data integration to support translational neurodegenerative disease research. The architecture is divided into three layers: data, services, and analytics layer.

Subsequent sections will describe the important modules of the architecture in more detail. As an example for the conversion from source data to a semantic layer, we describe the content of the 'omics' semantic layer in more detail.

3.1 Semantic layer for omics data

The main purpose of a semantic layer is to provide meaningful semantics to data, in order to enable the connection between primary and secondary data. Another key aspect of the introduced semantic layer is that it stores derived data, which represents the analysed and, in some cases, interpreted data by experts. We focus extensively on the usage of derived data for several reasons:

1) Raw data consumes a lot of space and it is costly to store it redundantly on several locations. 2) Often the raw data needs to be manipulated (for e.g. cleansing, transforming, standardizing, harmonizing, normalizing) to prepare it for the combined or comparative analysis. 3) The development and execution of analysis pipelines for processing raw data needs expert knowledge, which is generally available on the local sites. 4) Raw datasets are not necessarily interoperable as their metadata annotations might not use the common vocabulary or, even worse, they might not even exist.

As result, a semantic layer reduces complexity and facilitates user data access, search, and understanding of data. Note that under "semantic layer" we understand data structure for a single type of data such as RNA sequencing (RNA-seq) and CAP gene expression analysis (CAGE) for RNA expression or whole exome sequencing (WES) for protein-coding genes (about 1% of the genome) in order to find mutation variants. In the case of the RNA source data in FASTQ format, in a first step, counts of expressed RNA or small RNA are calculated. This is done with available tools such as OASIS (https://oasis.dzne.de/) [8] for the calculation of small RNA counts. In a second step, differential expression scores (p-values) between healthy and diseased subjects are calculated.

The derived data information, the RNA counts as well as the p-values for differential expression together with fold change information, is stored in the semantic layer instead of the initial FASTQ data (for an overview, see Table 1). In addition, the RNA entities are normalized to their corresponding genes in HGNC (https://www.genenames.org/) and Ensemble (http://ensemblgenomes.org/). Furthermore, all metadata annotations are normalized as well and stored in the semantic layer. Examples of metadata annotations are the organism, tissue, cell type or disease type. For the normalization of annotations, vocabularies from the semantic lookup service are used (cf. next subsection "Data annotation").

In the case of WES data, genetic variants are normalized to dbSNP (https://www.ncbi.nlm.nih.gov/snp) entities. The variant frequency is calculated with the help of external reference data sources. Furthermore, for the calculation of the disease burden for genetic variants, the CADD (Combined Annotation Dependent Depletion) [9] score is used. CADD can quantitatively prioritize functional, deleterious, and disease causal variants across a wide range of functional categories including effect sizes and genetic architectures. It can be used to prioritize causal variation in both research and clinical settings. The variant frequency as well as the CADD score is stored in the semantic layer for WES data. In addition to the internal data, information from external resources are integrated. Gene and variant -disease relationships are integrated

from DisGeNET (http://www.disgenet.org) which assembles this information from databases as well as from literature. Furthermore, miRTarBase and SCAIView (https://www.scaiview.com/) are used to integrate miRNA-gene relations.

For data annotations the designed semantic layers utilize the controlled neuro-specific vocabularies and mappings available in the semantic lookup platform. This includes consistent annotation of the data with particular semantic terms, which are common for different data types as well as metadata. The annotations are stored as key:value descriptions using controlled vocabulary for both key and value terms. An example key:value pair is HGNC:BCL2, where the key HGNC is the reference to the terminology and BCL2 the HGNC label for the gene BCL2. Currently, the semantic lookup platform contains more than 20 pre-defined annotation keys, however, new (not normalized) keys are allowed as well. Using predefined keys allows user to preserve semantic meaning of annotations. Those not-normalized keys and values are uploaded into the semantic lookup platform as additional terminology.

Table 1. Types of omics data in the IDSN platform. From different primary data, expressed small RNAs and genes and gene variants are identified and normalized to the corresponding concepts from mirBase, HGNC, Ensembl or dbSNP. Small RNA count, RNA count and variant counts as well as differential expression for RNA and CADD score for variants are calculated and stored in the semantic layers. Within the semantic integration platform further relationships such as gene-variant, miRNA-gene or gene-disease relations are added from external resources.

Primary Data	Data type measured	Controlled vocabulary	Derived data	Secondary data	External source
small RNA-sec	small RNA	miRbase, Ensembl	counts, differential expression (p-value)	miRNA-gene relations	miRTarBase SCAIView
RNA-seq	RNA	HGNC, Ensembl	counts, differential expression (p-value)	gene-disease relations	DisGeNET
CAGE	RNA	HGNC, Ensembl	counts, differential expression (p-value)		
WES	mutation variant	dbSNP	variant calling, variant annotation (CADD)	gene-variant relations variant-disease relations	dbSNP, DisGeNET

For the annotation of data sets, an annotation tool that integrates the semantic lookup platform was developed. The annotation of data is designed as a semi-automated process: the system automatically suggests a ranked list of normalized concepts for existing annotations based on the Levenshtein distance between database entries and controlled vocabulary. These suggestions are provided within a user interface to the users

for manual curation. It enables editing, adding, and deleting as well as searching for concepts within the integrated semantic lookup platform interface.

We also annotate biological conditions that are part of the metadata. Biological condition annotation allows to group samples of a dataset in such a way, that samples of one group correspond to particular biological condition. Examples of biological conditions are healthy and diseased patients, or several diseases, or several stages or conditions of a particular disease. Annotation of biological conditions allows to perform some data analysis automatically for the semantic layer or within the semantic data integration platform. For example, differential expression analysis of small RNA datasets based on annotation of biological conditions can be directly computed using OASIS.

3.2 Semantic lookup platform

In translational bio-medical research, controlled bio-medical vocabularies such as terminologies, ontologies, taxonomies play an important role for annotation, integration, and analysis of biological data. These vocabularies are essential for data interoperability across departments and institutes. The semantic lookup platform extends the proposed architecture by providing access to semantics through bio-medical vocabularies facilitating data integration and interoperability. It provides a coverage on terms within the semantic data layers, a detailed description for each term, and provide mappings to same concepts from different vocabularies.

After reviewing several existing open-source software projects (such as AberOWL, Ontobee, BioPortal, Ontology Lookup Service, Ontology Cross-reference Service), we chose two services: as entity resolution service (ERS) we selected the Ontology Lookup Service (OLS) [10] and as entity mapping service (EMS) we chose the Ontology Cross-reference Service (OXO) (both developed by the EMBL-EBI).

Both services provide a web-based user interface for exploring and visualizing the vocabularies (ERS) and mappings (EMS) and, additionally, a flexible REST-based API to programmatically access these resources. Additionally, they both provide a utility to regularly update vocabularies and mappings. Furthermore, the ERS includes a search engine for terms and synonyms with autocomplete functionality. To manage vocabularies, the ERS also uses a flexible configuration system.

An important extension of the provided services is the incorporation of terminologies that enable to annotate and map all entities in the different semantic layers. Mainly these are relevant life science instances such as genes, SNPs, miRNAs, organisms, cell lines, and terminologies for the description of neuroscience-relevant clinical conditions.

The designed semantic data layers utilize the vocabularies available in the semantic lookup platform. Terms in such controlled vocabularies have several characteristics that make them suitable for annotation and curation of data. A single term often represents a formal specification of a biomedical concept. They are defined and standardized by assigning a persistent identifier (with an IRI), a unique primary label, and a textual description. They can also include further metadata such as abbreviations, synonyms, and cross-references. Additionally, these terms can be hierarchically organized and put

in a relationship with each other. Using such vocabularies allows the alignment of datasets, makes datasets semantically meaningful, and facilitates data understanding by end users.

An advantage of using hierarchical vocabularies or ontologies within the semantic lookup platform is the possibility to search by parent terms. For example, a search for "neurodegenerative disease" will find all samples annotated by any subclass of this disease category or search for "brain" will find samples annotated by one of the brain parts.

3.3 Data management platform

The importance of a good data management to support scientific discovery and innovation is highly emphasized by Wilkinson et al. [11], for which they have developed the FAIR (findable, accessible, interoperable, and reusable) guiding principles to manage scientific data. An additional aspect is that it is also important for the various DZNE departments to discover datasets published by other departments with the goal that these datasets can be evaluated and re-used for further experiments. Thus, we incorporated such a data management platform that provides services to catalog and search (derived) datasets in the proposed architecture.

We use the open source software DKAN (https://getdkan.org/) to catalog and publish the biomedical datasets generated at different DZNE sites. The data management platform generates a formal citation for each added dataset. To cite the data, it supports the popular Open Data Metadata Schema (https://project-open-data.cio.gov/v1.1/ schema/) that is based on the Data Catalog Vocabulary (DCAT) (https://www.w3.org/ TR/vocab-dcat/), a W3C recommendation, which is designed to facilitate interoperability between data catalogs. It provides a persistent identifier as soon as a dataset is published. Additionally, the datasets include (neuroscience-specific) metadata, licenses, authors, and version information, all of which is cataloged, indexed, and searchable through a web-based user interface. The software also offers several REST-API endpoints to communicate with other services while allowing browsing the datasets, accessing metadata, and retrieving the datasets.

3.4 Semantic Data Integration Platform

The semantic data integration platform is the central part of the proposed architecture. It interlinks between different types of biomedical derived data together with annotations and secondary data. The primary goal of the integrated semantic data hub is to enable end users to answer their research questions. For example, researchers of neurodegenerative diseases may be interested to investigate the role of a particular gene in different types of diseases. Clinical doctors may be interested in the interpretation of genetic tests of a particular patient.

The fundamental data structure for the indexed data in the semantic hub is represented as a graph that we consider as essential for the analysis of the integrated data. The nodes in the graph represent the entities and edges represent the relations that are

used to connect entities with each other. Furthermore, nodes and edges may have additional properties or metadata such as the context information or the provenance attached to them. The platform also allows flexible data modeling to integrate heterogeneous datasets that are not (fully) suited for the graph-based structure such as clinical routine data. Hence, the platform design additionally covers the combination of two further database types to integrate relational and document-based data.

During the indexing process in the semantic hub, data is being connected and aligned with secondary data. This data has been incorporated from external resources and is necessary to link the different entity types. Examples are the regulation of genes by miRNAs that have been extracted from miRTarBase and from SCAIView. Other external resources are listed in Table 1. As such associations are of graph-nature, they fit perfectly in the graph database of the platform.

The REST-based API, as for the other services, is a key component in the data hub for common scientific queries. According to the needs of the queries, the implemented interfaces access, filter, and combine integrated data from databases. In such a way, the API can wrap the well-optimized queries built specifically for either graph, relational and/or document-based databases. This enables us to provide a high-performance platform with fast responses. During the development we also focused on the requirements of the developers who build dedicated (web-based) user interfaces or apply analytical approaches over the integrated data. Furthermore, the platform also communicates with the semantic lookup platform to retrieve entity-based information, or with further external services such as SCAIView, OASIS, NeuroMMSig (https://neurommsig.scai.fraunhofer.de/) to retrieve secondary data relevant for the asked scientific questions.

4 The Small RNA Expression Atlas as visualisation and computational analysis use case

The Small RNA Expression Atlas (SEA) is a web application that allows for the search of known and novel small RNAs across ten organisms using standardized search terms and ontologies (http://sea.ims.bio/) [12]. It is based on the IDSN semantic hub, however, for one particular primary data type (small RNAs). In contrast to proprietary patient data that is not publically available, SEA incorporates publicly available datasets from GEO. For the generation of the semantic layers, all data is semantically annotated with the support of the annotation tool and the semantic lookup platform. Furthermore, derived data is obtained by using the OASIS web application. The derived data includes smallRNA counts as well as pathogen expression. Future analysis incorporated within the semantic layer includes differential expression and classification relevance scores, p-values for DE and Gini indices for classification respectively. SEA supports interactive result visualisation of the data within the semantic integration platform. It allows for querying and displaying sRNA expression information, primary and derived data visualization, as well as visual analysis for disease-specific biomarker detection based on relevance scores. In addition, it supports the re-analysis of selected data and contains a user model for user-specific data management.

5 Conclusion

In this manuscript we discussed a novel semantic data integration architecture with primary focus on neurodegenerative disease research. The architecture allows to create unified view of distributed biological data using the concept of federated semantic data layers, and to integrate data together with derived and secondary data in a central integration platform.

Using of semantic concepts provides data with semantic meaning, which facilitates querying by end users, as well as allows interoperability of different types of biological data. The semantic lookup platform provides all necessary semantic concepts.

The architecture demonstrated its efficiency serving as basis for smallRNA Expression Atlas (SEA). SEA allows semantic integration of a big amount of publicly available smallRNA data, linked storage of smallRNA and pathogen information together with DE and classification results as well as smallRNA-gene and smallRNA-disease associations from external databases.

In this manuscript, we focused primarily on the integration of omics data. For two other types of data: pharmacological assays as well as clinical information, a similar approach was used. The resulting semantic-aware architecture will represent the basis for DZNE data integration, which will allow querying across the various highlighted data types.

Acknowledgements

The project IDSN is supported by the German Federal Ministry of Education and Research (BMBF) as part of the program "i:DSem – Integrative Data Semantics in the Systems Medicine", project number 031L0029 [A-C].

References

1. Bickel, D.H.: Die Häufigkeit von Demenzerkrankungen, https://www.deutsche-alzheimer.de/fileadmin/alz/pdf/factsheets/infoblatt1_haeufigkeit_demenzerkrankungen_dalzg.pdf, (2014).
2. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3, 160018 (2016).
3. Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J.A.K., Sempoux, C., Machiels, J.P., Haustermans, K., Moor, B. De: A kernel-based integration of genome-wide data for clinical decision support. Genome Med. 1, (2009).

4. Dawany, N.B., Dampier, W.N., Tozeren, A.: Large-scale integration of microarray data reveals genes and pathways common to multiple cancer types. Int. J. Cancer. 128, 2881–2891 (2011).
5. Iyappan, A., Kawalia, S.B., Raschka, T., Hofmann-Apitius, M., Senger, P.: NeuroRDF: Semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer's disease. J. Biomed. Semantics. 7, 45 (2016).
6. Groth, P., Loizou, A., Gray, A.J.G., Goble, C., Harland, L., Pettifer, S.: API-centric Linked Data integration: The Open PHACTS Discovery Platform case study. J. Web Semant. 29, 12–18 (2014).
7. Gardner, D., Akil, H., Ascoli, G.A., Bowden, D.M., Bug, W., Donohue, D.E., Goldberg, D.H., Grafstein, B., Grethe, J.S., Gupta, A., Halavi, M., Kennedy, D.N., Marenco, L., Martone, M.E., Miller, P.L., Müller, H.-M., Robert, A., Shepherd, G.M., Sternberg, P.W., Van Essen, D.C., Williams, R.W.: The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience. Neuroinformatics. 6, 149–160 (2008).
8. Rahman, R.U., Gautam, A., Bethune, J., Sattar, A., Fiosins, M., Magruder, D.S., Capece, V., Shomroni, O., Bonn, S.: Oasis 2: Improved online analysis of small RNA-seq data. BMC Bioinformatics. 19, (2018).
9. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M., Shendure, J.: A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).
10. Jupp, S., Burdett, T., Malone, J., Leroy, C., Pearce, M., McMurry, J., Parkinson, H.: A new ontology lookup service at EMBL-EBI. In: CEUR Workshop Proceedings. pp. 118–119 (2015).
11. Allen, A., Aragon, C., Becker, C., Carver, J.C., Chis, A., Combemale, B., Croucher, M., Crowston, K., Garijo, D., Gehani, A., Goble, C., Haines, R., Hirschfeld, R., Howison, J., Huff, K., Jay, C., Katz, D.S., Kirchner, C., Kuksenok, K., Lämmel, R., Nierstrasz, O., Turk, M., Van Nieuwpoort, R. V., Vaughn, M., Vinju, J.: Lightning talk: "I solemnly pledge" A manifesto for personal responsibility in the engineering of academic software. CEUR Workshop Proc. 1686, 160018 (2016).
12. Rahman, R.-U., Sattar, A., Fiosins, M., Gautam, A., Magruder, D.S., Bethune, J., Madan, S., Fluck, J., Bonn, S.: SEA: The Small RNA Expression Atlas. bioRxiv. 133199 (2017).