Noname manuscript No.
(will be inserted by the editor)

Towards Semantic Integration of Federated Research Data

Javad Chamanara · Angelina Kraft · Sören Auer · Oliver Koepler

Received: 01.02.2019 / Accepted: date

Abstract Digitization of the research (data) lifecycle has created a galaxy of data nodes that are often characterized by sparse interoperability. With the start of the European Open Science Cloud in November 2018 and facing the upcoming call for the creation of the National Research Data Infrastructure (NFDI), researchers and infrastructure providers will need to harmonize their data efforts. In this article, we propose a recently initiated proof-of-concept towards a network of semantically harmonized Research Data Management (RDM) systems. This includes a network of research data management and publication systems with semantic integration at three levels, namely, data, metadata, and schema. As such, an ecosystem for agile, evolutionary ontology development, and the community-driven definition of quality criteria and classification schemes for scientific domains will be created. In contrast to the classical data repository approach, this process will allow for cross-repository as well as cross-domain data discovery, integration, and collaboration and will lead to open and interoperable data portals throughout the scientific domains.

At the joint lab of L3S research center and TIB Leibniz Information Center for Science and Technology in Hannover, we are developing a solution based on a customized distribution of CKAN called the Leibniz Data Manager (LDM). LDM utilizes the CKANs harvesting functionality to exchange metadata using the DCAT vocabulary. By adding the concept of semantic schema to LDM, it will contribute to realizing the FAIR paradigm. Variables, their attributes and relationships of a dataset will improve findability and accessibility and can be processed by humans or machines across scientific domains. We argue that it is crucial for the RDM development in Germany that domain-specific data silos should be the exception, and that a semantically-linked network of generic and domain-specific research data systems and services at national, regional, and organization levels should be promoted within the NFDI initiative.

Keywords Semantic Interoperation · Ontology Development · (Meta)Data Harmonization

1 Introduction

Data-intensive science is highly dependent on three core components; 1) software as the specification of methods, 2) data from experiments and observations, and 3) semantics as representative of structure, context, and the domain of knowledge. Nowadays, data is spread over so-called data portals, repositories, and hubs, which are rapidly increasing also with respect to variety, and content. This has created a galaxy of data nodes with multi-level heterogeneity and sparse interoperability.

Many datasets express the same phenomena but in varying syntax, serialization, or resolution. They even may have temporal, spatial, or conceptual overlaps that are not accessible to cross-domain and/or cross-repository queries. With the advancement of data-intensive research as well as machine-aided research e.g., Artificial Intelligence and Machine Learning, the need for multi-disciplinary data integration and analysis is growing.

Javad Chamanara, Angelina Kraft, Sören Auer, and Oliver Koepler
Technische Informationsbibliothek (TIB)
Welfengarten 1B, 30167 Hanover, Germany
Tel.: +49-511-76217951
Fax: +49-511-76214474
E-mail: {chamanara, angelina.kraft, auer, and oliver.koepler} @tib.eu

1.1 Related initiatives

Many solutions have been proposed and implemented to handle this complex heterogeneous environment with the aim of providing a unified access layer for end users. These approaches are mostly relying on metadata and/or data harvesting, e.g., generic services such as B2FIND [15] and DataONE [11], as well as discipline-specific services such as GBIF [4]. There have been solutions such as GeRDI [6] that act as a gateway for the long-tail of researchers who have limited access and/or knowledge to directly work with repositories. Such solutions allow their users to submit the data to the gateway. The gateway then resubmits the data to the designated target repository.

There are discipline-specific efforts such as GFBio [3] that add a layer of semantic integration on top of data repositories to answer higher level queries. BEXIS 2 [5] is system that integrates with systems such as GFBio terminology service [9] to enrich the metadata acquired from users. It also integrates with publishers such as Pangea[1] and GBIF[2] to cover more data lifecycle-related activities and to relieve users from effort-intensive manual quality control, and publishing. Similar to GFBio, it takes advantage of a semantic search that is plugged into the system via ontologies and their mappings to metadata and the data.

OpenPHACTS[17] aims to provide an open pharmacological space by integrating pharmacological data from various data resources and providing tools and services to query this integrated data to support drug discovery research. Another example is the DARIAH project[13], which aims to establish a network of research infrastructures for eHumanities support of research practices based on information and communication technology using virtual research environments (VREs). One example of semantic enhancement is the Semantic Topological Notes (SemToNotes) tool[3] developed at the Institute of Humanities and Computer Science at the University of Cologne as part of the DARIAH-DE Project. This tool enables a topological image annotation and image retrieval, allowing the analysis of spatio-topological relations between semantically enriched image areas.

The varietyin data formats and volumes, the metadata schema, the services expected to be provided by repositories, the extent of integration with other systems in research and publishing environments, and the need for broader data life cycle support has led many research institutes as well as funding agencies to increase their efforts towards the realization of full-fledged, wide-spectrum, and sustainable solutions [2].

In the course of the digital transformation in the German science ecosystem, some central challenges arise with regard to heterogeneity and quality standards of research data. These are in particular related to the decentralization of the actors involved, e.g., scientists at universities, research institutions, companies, and professional associations. The heterogeneity in technical, semantic, and organizational layers increases the depth of the challenge. Already in academic education, a stronger thematization of aspects of digitization is necessary. This counteracts the isolation of individual scientists and groups who are interested in greater depth in aspects of data acquisition, storage, evaluation, high-performance computing (HPC), or machine learning. So far, these groups have hardly succeeded in ”taking along” the entire community of their corresponding scientific domains and in initiating comprehensive change processes.

1.2 National Research Data Infrastructure (NFDI)

As a national answer to the establishment of the European Open Science Cloud (EOSC) [1] and the requirement to increase the FAIRness (findable, accessible, interoperable, reuseable) [12,16] of research data by the European Commission, the German government presented its plan to establish the National Research Data Infrastructure (NFDI). This initiative has come to life by the decision of the Joint Science Conference (GWK) on 16 November 2018 and will be driven and coordinated by the German Research Foundation (DFG) with a planned start in early 2020 and a running time of at least 10 years[4]. With funding of approximately 85-90 million Euro per year in the end phase of NFDI, this is one, if not the largest national call for research data development among European countries so far. NFDI is expected to fund about 30 consortia representing various scientific disciplines, which will be selected in three rounds of calls and will be evaluated at regular intervals by an expert committee coordinated within the DFG.

It is anticipated that these disciplines will not come up with a single central ”one-fits-all” solution. Rather, it is more realistic to assume that they realize solutions that comply with their specific requirements, hence, a large diversity is expected. Although there would be a layer of shared infrastructure and principles, requirements such as metadata standards, data usage patterns, data protection policies, funding, and administrative

[1] `https://www.pangaea.de/`

[2] `https://www.gbif.org/`

[3] `https://github.com/UB-Heidelberg/SemToNotes`

[4] `http://www.dfg.de/foerderung/programme/nfdi/`

domains will likely lead to a distributed yet cooperative network of data management systems. Research data is a constitutive and complex element of research workflows and the associated processes. As such, one of the main obstacles of the NFDI initiative will be the introduction of FAIR principles to all levels of research data management (RDM), with a focus on the data exchange between the existing data repositories and the verification of data quality.

In this article, we introduce an approach that envisions a network of research data management and publication systems with semantic integration at three levels, namely, data, metadata, and schema. This approach is proposed as an integral layer that will contribute to data harmonization in the NFDI framework. Some elements of the solution are already prototyped and are under experimental usage at TIB[5].

2 The Proposed Approach

In the spirit of the NFDI initiative and the European Open Science Cloud (EOSC), we argue that it is crucial for Germany to either lower or discourage the growth of isolated data silos. These data silos may be established because of separate funding, domains of activities, or technologies utilized. Rather a network of interconnected domain-specific data management systems at national, regional, and organization levels should be promoted. Currently, the open data world (e.g. (meta)data standards like DCAT, CKAN, LOD, schema.org, etc.) and the research data world (e.g. DDI, DataCite, etc.) are still disparate. It is our goal to close this gap and use open data standards to facilitate FAIR research data infrastructures.

In this paper, we describe our recently begun effort to approach this gap by proposing a proof-of-concept that relies on federation and harmonization; federation in maintaining data, harmonization in describing and accessing data.

The proposed approach also suggests an infrastructure for RDM nationwide. The infrastructure will be utilized by different actors, e.g., universities, institutes, companies, and governments, as well as multiple scientific disciplines. As it is meant to cover a wide variety of RDM related activities, we target the following five high-level objectives:

1. to establish an agile, iterative, and community-driven method for ontology development by and with all stakeholders,
2. to establish a community-driven definition of quality criteria and classification schemes for scientific domains and engage with the existing ones,
3. to describe the structure of typical data formats and creation of mappings on the ontologies built,
4. to realize open and interoperable data portals for the scientific domains, and
5. to promote community participation, re-use, and transfer of knowledge.

In analogy with IaaS that provides infrastructure as a service, our proposed solution would provide DIaaS; Data Infrastructure as a Service. DIaaS is a distributed network of so-called service nodes that operate in their respective administrative domains e.g., institutes, universities, and government to serve their domain-specific audience e.g., biodiversity, material science, and chemistry with data management services. Each node can run one or more services. Each participating administrative domain may run one or more nodes in order to provide a set of designated services to different disciplines. The participants may provide the services as the source of Truth or as a point of availability.

A catalog service is used to provide metadata and semantic descriptions to facilitate data discovery. A repository service hosts the actual data and is reachable from the corresponding catalog(s). An archiving service preserves data for its lifetime and provides unique identification based on a permanent identification scheme such as DOI as well as data versioning for citation and reproduction purposes.

It is possible and encouraged for the participants to configure the network, the nodes, and the services according to their requirements. For example, a single catalog can represent all the data of a set of designated repositories. A university's catalog service may represent the catalogs of its departments, while each departmental catalog represents its respective department's repository. A regional data management system may harvest data and metadata from local partners and act as an aggregator. Also, it is foreseeable that the datasets of a repository are discoverable by many catalogs.

By achieving these objectives we establish a network of semantically harmonized RDM systems that are built by and for their corresponding communities. In addition, this approach allows for cross-repository as well as cross-domain discovery, integration, and collaboration. Harmonizing RDM services requires the application of semantic description and networking of data in various scientific domains and syndication via relevant subject portals and aggregators.

Fig. 1 illustrates a sample topology of such a hierarchical and peer-to-peer network of research data infrastructures. This topology shows how exemplary regional

[5] https://www.tib.eu/en/

and thematic repositories can be kept standalone yet cooperating. A regional body such as the Leibniz University Hanover may choose to ingest metadata from engineering repositories, while Hanover Medical School is interested only in the repositories active in the field of medicine.

These bi-/multi-lateral cooperation patterns can be easily established to broaden the visibility of data to a larger and more diverse audience. At the same time, the hierarchical structure allows for metadata and/or ownership propagation when there is an administration demand for it. Higher level nodes such as the ministry of education and research in Germany may opt to provide discovery service for all its subsidiaries without requiring the data to be centralized at a ministry-level repository.

Scientists play important roles here. The whole solution is designed based on the assumption of the existence of an actively contributing community. Individual scientists may produce and/or consume data as well as metadata. They are additionally, and more importantly, involved in the ontology definition, development, and application loop. Therefore, the workflows of searching for proper data, submitting data, describing data with metadata, and enriching it with semantics involves the scientists. This establishes a positive feedback cycle that not only the scientists and their audience benefit from, but also produces a set of rich, stable, and agreed-upon ontologies.

This degree of flexibility comes at a price; data integration and data ownership! The ownership of research data becomes an issue because of the possible circulation of data between repositories/archives owned by different administrative domains. One possible solution to maintain data sovereignty is to containerize data. However, in this paper, we focus on the data integration issue only.

Data integration deals with heterogeneity in format, syntax, and semantics levels. Any collaborative and distributed data management infrastructure needs to suggest a solution for this heterogeneity. We approach this challenge at different stages. We define a dataset as a three components package: data, schema, and metadata. The schema defines the structure of the data so that other human and machine users can understand and consume the data. The metadata explains different aspects of the content of the dataset. It has four components: domain-agnostic metadata attributes that define the bibliographic information of the dataset, domain-specific metadata attributes that describe the content of the dataset, domain-specific metadata that describes a value or an object inside the dataset, and any other metadata that specifies the policies and rules governing the dataset e.g., access and publishing.

2.1 The Leibniz Data Manager

In the following, we describe semantic integration and description as well as the data ownership. It is notable that the solution offered here is under development at the L3S/TIB joint lab, Leibniz University Hanover, Germany. It is an iteratively developed product based on a customized distribution of CKAN[6] called Leibniz Data Manager[7] (LDM). LDM is able to play any of the abovementioned service roles alone or in combination with a set of plug-ins curated by TIB.

LDM is provided as a multi-layer distribution. A base version and a set of satellite distributions. The base is maintained by TIB under a liberal open source copyright. LDM base is a distribution with features that are generic and wide enough to cover most of the requirements of an RDM. On top of the base distribution, LDM offers a set of discipline-specific distributions. These distributions are tailored to the specific needs of a user community or discipline, e.g., chemistry, ecology, and biodiversity. These tailored distributions may offer extra features not available in the base distribution. They may also customize a set of features to adapt to the nature of the work in the target discipline. For example, in chemistry, it would be useful to search the datasets by the specification of a molecular structure or by providing a segment of their NMR spectrum. Also, visualizing a molecule found in a dataset as a 3D object based on its InChi code would greatly improve the understanding of the user with respect to the dataset.

LDM base provides faceted search and semantic tagging. In addition to well-known data types, LDM facilitates visualization of drawings and AutoCAD files for e.g., material and engineering sciences. Also, it supports Jupyterlab[8] to allow scientists online programming access to the data.

LDM base is additionally equipped with a set of cross-cutting features that need to be configured and/or adapted to the requirements of the specific distribution. One of these features is the semantic tagger, which allows the data owners/ data curators to annotate the datasets with tags that are obtained from a terminology service. These tags may come from controlled vocabularies, thesaurus, or ontologies. However, each domain of science may have its own set of vocabularies,

[6] https://ckan.org/

[7] https://projects.tib.eu/datamanager/

[8] https://jupyterlab.readthedocs.io/en/stable/

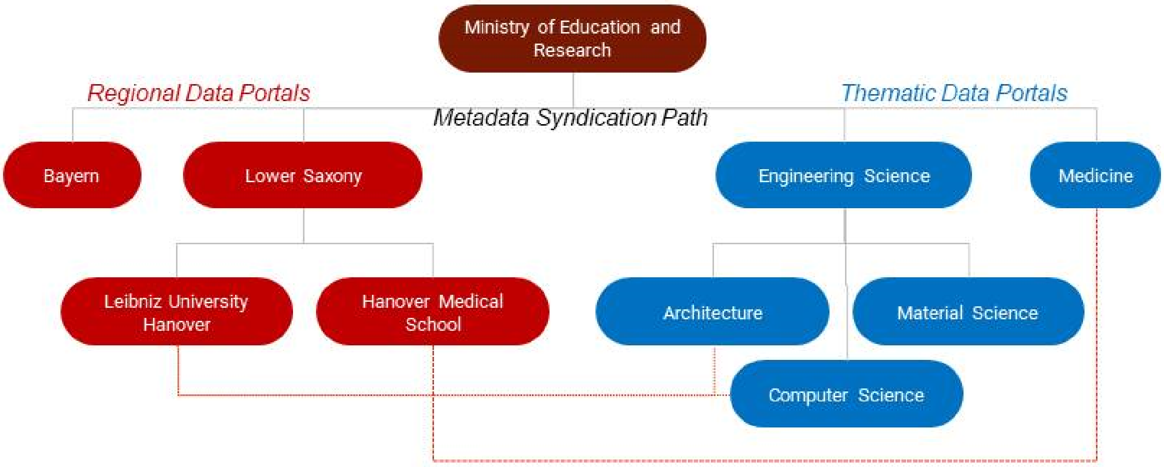


Fig. 1 An exemplary topology of the network of research data repositories. This topology shows how regional and thematic repositories can be kept standalone yet cooperating. Higher level nodes provide discovery function for all their sub-repositories.

thesaurus, or ontologies. Therefore, the semantic tagger must be configured to use a proper server.

LDM utilizes the CKAN's harvesting functionality to exchange the domain-agnostic part of the metadata using the DCAT vocabulary[9]. The domain-specific part of the metadata will be defined by the participating consortia and plugged into their LDM deployments. Here is where the metadata heterogeneity emerges; various metadata and different data schemes (possibly for conceptually similar or compatible data). The answer to this challenge is to employ semantic integration and encourage cooperative annotation and curation.

2.2 Schema Integration

Although data can be of any format and structure, many research data benefit from some sort of schema. These schemas define the meaning of the values acquired during an experiment, an observation, a simulation, or a computation. The meanings are usually captured in so-called variables or parameters, which have labels, units of measurement, and in some causes ranges and domains.

The process of matching and identifying these variables as well as determining how they are convertible to other compatible ones falls under the concept of schema integration. It is done via linking data schemas to their corresponding concepts in ontologies. This, in turn, brings the whole power of semantic annotation, querying, and inference into the schemas. Therefore, one can, for example, query for datasets that contain a specific variable with all its sub-variables. It is also possible to automatically convert data from a measurement system to another and run user queries against the harmonized data.

Scientists of related domains obtain the capability to search across datasets and merge them according to the semantic compatibility of the datasets' schemas. For this purpose, domain-specific classification schemas (such as SKOS thesauri[10]), as well as the required mappings to describe the data structures on a domain ontology, would have to be developed and linked in the corresponding DCAT descriptions. The development effort would be managed by cooperative curation of members of cooperating entities. Terminology services and ontologies integrated into EOSC ecosystem are valuable assets here.

One example of such a co-operation between scientists and infrastructure providers is the development of a parameterized technology process model at Leibniz University Hanover which analyzes existing standards data structures used in mechanical engineering and tailored forming and develops a classification system for the components of the process chain. Based on the developed classification system, a production-specific vocabulary can be developed and evaluated. An instance of the LDM is provided for the goal of ensuring systematization, transparency, and long-term archiving of extensive process data in different formats and for different software versions that support these formats. For this purpose, based on the production-specific vocabulary, a central metadata directory supported by semantic, machine-readable vocabularies is set up.

By adding this concept of the semantic schema to LDM, it greatly contributes to the FAIR paradigm.

[9] https://www.w3.org/TR/vocab-dcat/

[10] https://www.w3.org/2004/02/skos/

Each variable and all its attributes and relationships are findable. Variables can be accessed and processed by human or machine. They are interoperable not only for consumption but also for building an open and interconnected set of domain-specific variables. Each variable can be uniquely identified and reused in different datasets.

2.3 Schema Description

Ontologies establish a common understanding of data and capture domain-specific semantics by defining concepts, associated attributes, and relations. We utilize semantic classification and description of data using domain-specific vocabularies and ontologies in order to enhance discoverability, interoperability, and harmonization of datasets. These vocabularies and ontologies have to evolve community-driven, iterative, and evolutionary. This requires appropriate technical and organizational support as well as utilizing existing robust solutions. Similar to schema.org, the global collaborative vocabulary creation initiative, Fraunhofer and TIB have jointly developed VoCol [7,8] to allow collaborative and online development of ontologies. This is a web-based visual authoring tool for collaborative crafting of ontologies. The ontologies are maintained in their respective GitHub repositories to benefit from all versioning, tagging, branching, and forking capabilities that GitHub provides. The VoCol web interface provides easy mechanisms for integrating with GitHub and performing commits.

While metadata of the datasets of different repositories can be exchanged via DCAT or W3C Data on the Web Best Practices, ontologies are used to provide harmonized meaning to the content of data. Therefore, it is possible to map data to different domain-specific vocabularies to achieve deeper content-based indexing of the data. Ontologies, in addition, enable data integration, e.g., data networking and federated access as well as new explorations possibilities, e.g., semantic search, cross-repository data discovery, and visualization. Fig. 2 presents an example of such a mapping from a running project. In the medical domain, a disease can be related to symptoms, which are treated with a specific medication. Within an ontology, the concepts of disease, symptom, and medication are described by attributes and mapped to datasets from different sources and with different detail level. A medical data analyst can now use the concepts of the ontology to build an integrated view of these enriched datasets and determine the *ICD-10* code and *Treatment* of the disease, even if the disease is named differently in the datasets. These data are taken from one of the projects that L3S Research Group[11] is involved in.

2.4 FAIR Principles

In the area of research data, FAIR principles have gained great popularity in recent years. However, the principles are often not technically clear enough, hence, leave many implementation possibilities open. This opens up the door for the emergence of technically incompatible yet nevertheless FAIR-compliant repositories for research data. The NFDI initiative will fund consortia from various scientific disciplines. It is anticipated that most consortia will target the specific requirements of their disciplines, which is necessary. However, this pure focus on individual disciplines, e.g., according to the DFG classification system, faces the obvious risk of leading to even further fragmentation and more data silos, i.e., less integration.

A major challenge for the formation of a common, shared NFDI for Germany and beyond that in Europe and worldwide, is the establishment of shared infrastructures and principles, which tackle also generic requirements, such as vocabulary and metadata standards, reference models, data usage patterns, data protection policies, licensing, and administrative domains, which will lead to a distributed yet cooperating network of scientific data management systems. Finding a "common ground" regarding technical, social, cultural, and economic aspects of research data management needs to be addressed from the very start of the NFDI funding. Here, tools like the LDM can aid in a transferable FAIR implementation on the basis of many years of experience and developments in the field of Open Data, semantic technologies, and W3C data on the web best practices.

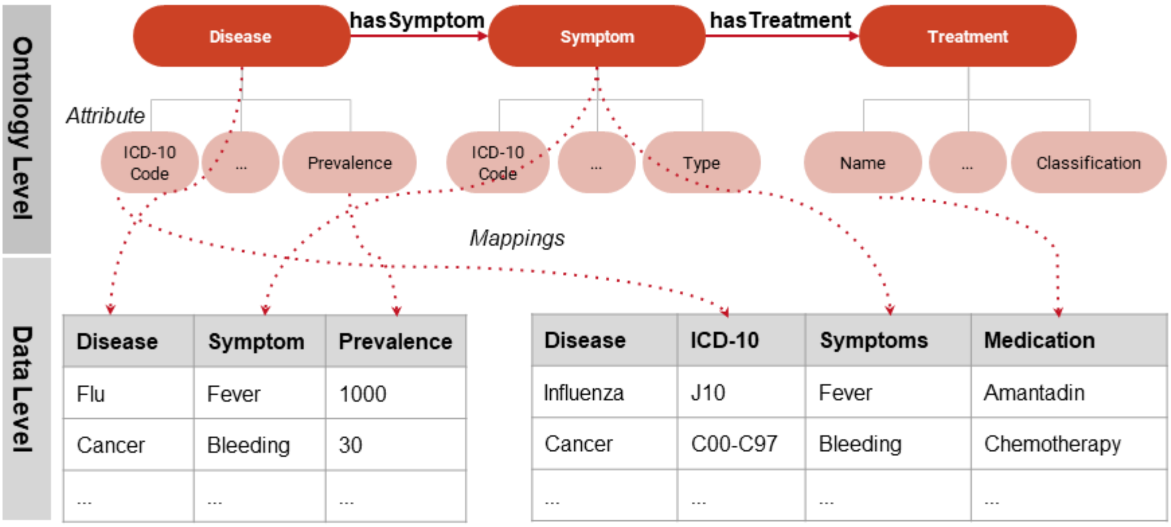
Nevertheless, this should also take place with the integration of community projects such as GO FAIR[14] or FORCE11[10], which demand the FAIR principles and certify their compliance.

3 Summary

In summary, our recently initiated work on the concept semantic integration aims at:

1. the establishment of a common understanding for the structuring of research data;
2. ensuring completeness, coherence, and consistency of data and metadata;

[11] `https://www.l3s.de/en`



Disease	Symptom	Prevalence
Flu	Fever	1000
Cancer	Bleeding	30
...	...	...

Disease	ICD-10	Symptoms	Medication
Influenza	J10	Fever	Amantadin
Cancer	C00-C97	Bleeding	Chemotherapy
...	...	...	...

Fig. 2 An example of a mapping of research data to a relevant ontology in the medical domain. Using a shared ontology, it is possible to establish mappings between conceptually compatible data schemas. This exemplary ontology and the segments of the datasets are from a project with Medical University of Hanover that the L3S research group is involved in.

3. recognizing and managing heterogeneity in data and metadata schemas; and
4. the support of the research data management processes in the complete research cycle within the NFDI initiative.

Table 1 is a short summary of the objectives and the tools, utilities, or standards we use to implement them. The individual data portals will be established by installing and running instances of LDM. The collaboration between the LDM nodes is managed by DCAT as minimum interchanged metadata. The semantic integration is enhanced by the ontologies developed by the communities. These ontologies are also published and shared under open copyrights for maximum outreach. The essence of our approach to RDM is made of these building blocks:

1. a set of independent yet cooperating repositories built on top on LDM;
2. a (set of) discovery services, built on LDM, that harvest metadata from the repositories to improve and increase the findability of the datasets;
3. a collection of tailored harvesters that are able to retrieve metadata from the repositories and map them to a harmonized metadata compatible with, e.g., DCAT;
4. a set of community-driven easy to build and edit vocabularies, developed on VoCol, to act as semantic annotation of metadata items such as tags, names, processes, and units; and
5. a set of well-described data structures and typical data formats to be used in the consortium of collaborating repositories.

The focus of vocabulary and ontology development is on aspects such as agility, iterative and community-driven development, and interoperability instead of more official processes such as standardization. For this, we utilize WebVOWL[12] for visual authoring of ontologies, VoCol for web-based collaboration and Git support, and VocBench[13] for fine tuning as well as online RDF and SPARQL support.

Table 1 Current implementation status of the solution objectives and the technologies used.

Objective	Implementation
Research data management and data portals	Leibniz Data Manager
Collaborative network of repositories	Harvesting, DCAT, and Semantic mappings
Ontology development	VoCol, WebVOWL, and VocBench
Classification Schemes	Online repositories of ontologies
Promotion and reuse	Semantic annotation of data and metadata
Structure of typical data formats	RDA DTR[14] (data type registry)

Our plan (recently begun) is to develop a set of manageable hierarchical core ontologies to ground the foundation of semantic interoperability between RDM systems. Domain-specific experts will establish their cus-

[12] http://www.visualdataweb.org/
[13] http://vocbench.uniroma2.it/
[14] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

tomized fine-tuned ontologies on top of those cores. This allows for independent evolution of the core and satellite ontologies, while maintains the interoperability between them.

Currently, two expert groups in the context of NFDI for Chemistry and NFDI for Material Science have shown their interest to use the approach and recently started to work with it. They are not only preparing data to be ingested into the repositories but also develop/assemble light-weight ontologies to describe the datasets. NFDI for Chemistry has initiated a molecule-/spectrum-based search as well as a structural visualization of the query results.

Despite the recent initiation of the work, we have partially achieved some of the objectives. However, a formal evaluation is planned for near future work when we have enough datasets in the system. The evaluation will measure the quality of dataset discovery service on single node repositories, on federated systems of multiple repositories with central discovery services, and compares the precision and recall of the latter with semantic annotations on and off.

References

1. Ayris, P., Berthou, J.Y., Bruce, R., Lindstaedt, S., Monreale, A., Mons, B., Murayama, Y., Södergård, C., Tochtermann, K., Wilkinson, R.: Realising the European Open Science Cloud. European Union (2016). DOI 10.2777/940154. URL `https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf`
2. Bijsterbosch, M., Duca, D., Katerbow, M., Kupiainen, I., Dillo, I., Doorn, P., Enke, H., de Lucas, J.E.M.: Funding research data management and related infrastructures: Knowledge exchange and science europe briefing paper (2016)
3. Diepenbroek, M., Glckner, F.O., Grobe, P., Gntsch, A., Huber, R., Knig-Ries, B., Kostadinov, I., Nieschulze, J., Seeger, B., Tolksdorf, R., Triebel, D.: Towards an integrated biodiversity and ecological research data management and archiving platform: the german federation for the curation of biological data (gfbio). In: E. Pldereder, L. Grunske, E. Schneider, D. Ull (eds.) Informatik 2014, pp. 1711–1721. Gesellschaft fr Informatik e.V., Bonn (2014)
4. Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., Neufeld, D.: A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). Ecological Informatics **2**(1), 49 – 60 (2007). DOI https://doi.org/10.1016/j.ecoinf.2007.03.004. URL `http://www.sciencedirect.com/science/article/pii/S1574954107000106`
5. Gerlach, R., Blaa, D., Chamanara, J., Hohmuth, M., Navabpour, N., Thiel, S., König-Ries, B.: Bexis 2: A platform for managing heterogeneous biodiversity data and projects. In: TDWG 2015 ANNUAL CONFERENCE (2015)
6. Grunzke, R., Adolph, T., Biardzki, C., Bode, A., Borst, T., Bungartz, H.J., Busch, A., Frank, A., Grimm, C., Hasselbring, W., Kazakova, A., Latif, A., Limani, F., Neumann, M., de Sousa, N.T., Tendel, J., Thomsen, I., Tochtermann, K., Müller-Pfefferkorn, R., Nagel, W.E.: Challenges in Creating a Sustainable Generic Research Data Infrastructure. Softwaretechnik-Trends **37**(2), 74–77 (2017). URL `http://oceanrep.geomar.de/38756/`
7. Halilaj, L., Grangel-González, I., Coskun, G., Lohmann, S., Auer, S.: Git4Voc: Collaborative Vocabulary Development Based on Git. Int. J. Semantic Computing **10**(2), 167–192 (2016). DOI 10.1142/S1793351X16400067. URL `https://doi.org/10.1142/S1793351X16400067`
8. Halilaj, L., Petersen, N., Grangel-González, I., Lange, C., Auer, S., Coskun, G., Lohmann, S.: VoCol: An Integrated Environment to Support Version-Controlled Vocabulary Development. In: Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings, pp. 303–319 (2016). DOI 10.1007/978-3-319-49004-5_20. URL `https://doi.org/10.1007/978-3-319-49004-5\_20`
9. Karam, N., Lorenz, R.H., Müller-Birn, C.: The gfbio terminology service: enabling research data management beyond data heterogeneity. Tage 2017 p. 75
10. Martone, M.E.: FORCE11: Building the Future for Research Communications and e-Scholarship. BioScience **65**(7), 635–635 (2015). DOI 10.1093/biosci/biv095. URL `https://doi.org/10.1093/biosci/biv095`
11. Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., Vieglais, D.A.: Participatory design of DataONE-Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics **11**, 5 – 15 (2012). DOI https://doi.org/10.1016/j.ecoinf.2011.08.007. URL `http://www.sciencedirect.com/science/article/pii/S1574954111000768`
12. Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D.: Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use **37**(1), 49–56 (2017)
13. Romary, L., Chambers, S.: DARIAH: Advancing a digital revolution in the arts and humanities across Europe. e-data&research (2014). URL `https://hal.inria.fr/hal-00913691`
14. Tochtermann, K.: GO FAIR - Eine Initiative zum fairen Umgang mit Forschungsdaten. In: TK 7: lehren & unterstützen / Forschungsdatenmanagement International (14.06.2018, 11:00 - 12:30 Uhr, Estrel Saal) (2018)
15. Widmann, H., Thiemann, H.: EUDAT B2FIND : A Cross-Discipline Metadata Service and Discovery Portal. In: EGU General Assembly Conference Abstracts, *EGU General Assembly Conference Abstracts*, vol. 18, pp. EPSC2016–8562 (2016)
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data **3** (2016)
17. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today **17**(21), 1188 – 1198 (2012). DOI https://doi.org/10.1016/j.drudis.2012.05.016. URL `http://www.sciencedirect.com/science/article/pii/S1359644612001936`