1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.202

Published: 07 Sept. 2023

Harmonising, Harvesting, and Searching Metadata across a Repository Federation

Steffen Neumann[1][https://orcid.org/0000-0002-7899-7192], Felix Bach[2][https://orcid.org/0000-0002-5035-7978], Leyla Castro[3][https://orcid.org/0000-0003-3986-0510], Tillmann G. Fischer[1][https://orcid.org/0000-0003-4480-8661], Stefan Hofmann[2][https://orcid.org/0000-0003-0790-112X], Pei-Chi Huang[4][https://orcid.org/0000-0002-9976-4507], Nicole Jung[4][https://orcid.org/0000-0001-9513-2468], Bhavin Katabathuni[6][https://orcid.org/0009-0003-1198-9969], Fabian Mauz[1][https://orcid.org/0000-0003-4673-5494], Rene Meier[1][https://orcid.org/0000-0002-1501-1349], Venkata Chandrasekhar Nainala[5][https://orcid.org/0000-0002-2564-3243], Noura Rayya[5][https://orcid.org/0009-0001-5998-5030], Christoph Steinbeck[5][https://orcid.org/0000-0001-6966-0814], and Oliver Koepler[6][https://orcid.org/0000-0003-3385-4232]

[1] Leibniz Institute of Plant Biochemistry, Halle, Germany, https://ror.org/01mzk5576

[2] FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany, https://ror.org/04z92tv25

[3] ZB Med Information Centre for Life Sciences, Cologne, Germany, https://ror.org/0259fwx54

[4] Karlsruhe Institute of Technology, Karlsruhe, Germany, https://ror.org/04t3en479

[5] Friedrich-Schiller-University, Jena, Germany, https://ror.org/05qpz1x62

[6] TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany, https://ror.org/04aj4c181

Abstract. The collection of metadata for research data is an important aspect in the FAIR principles. The schema.org and Bioschemas initiatives created a vocabulary to embed markup for many different types, including BioChemEntity, ChemicalSubstance, Gene, MolecularEntity, Protein, and others relevant in the Natural and Life Sciences with immediate benefits for findability of data packages. To bridge the gap between the worlds of semantic-web-driven JSON+LD metadata on the one hand, and established but separately developed interface services in libraries, we have designed an architecture for harmonising, federating and harvesting metadata from several resources. Our approach is to serve JSON+LD embedded in an XML container through a central OAI-Provider. Several resources in NFDI4Chem provide such domain-specific metadata. The CKAN-based NFDI4Chem search service can harvest this metadata using an OAI-PMH harvester extension that can extract the XML-encapsulated JSON+LD metadata, and has search capabilities relevant in the chemistry domain. We invite the community to collaborate and reach a critical mass of providers and consumers in the NFDI.

Keywords: Metadata, Structured Markup, JSON+LD, schema.org, Bioschemas, OAI-PMH, Harvesting

1. Background

Research data is a critical component of scientific inquiry, and its value lies in its ability to support the reproducibility, transparency, and reuse of research findings. By sharing research data, researchers enable others to verify their findings, build upon them, and contribute to the development of new knowledge. Several generic and domain-specific data repositories have been introduced over the years to publish data. DataCite not only provides persistent identifiers

(DOIs) for research datasets and data publications, but also the DataCite Metadata Schema as a standard format for a generic and discipline-independent description of research data. This enables easy discovery, access, and reuse. By adopting the DataCite Metadata Schema, researchers can ensure their data is properly documented and discoverable, which increases the visibility and impact of their research. Nevertheless, DataCite Schema is somewhat limited when it comes to expressing domain-specific metadata like molecular entities, chemical structure information, analytical methods, or processes.

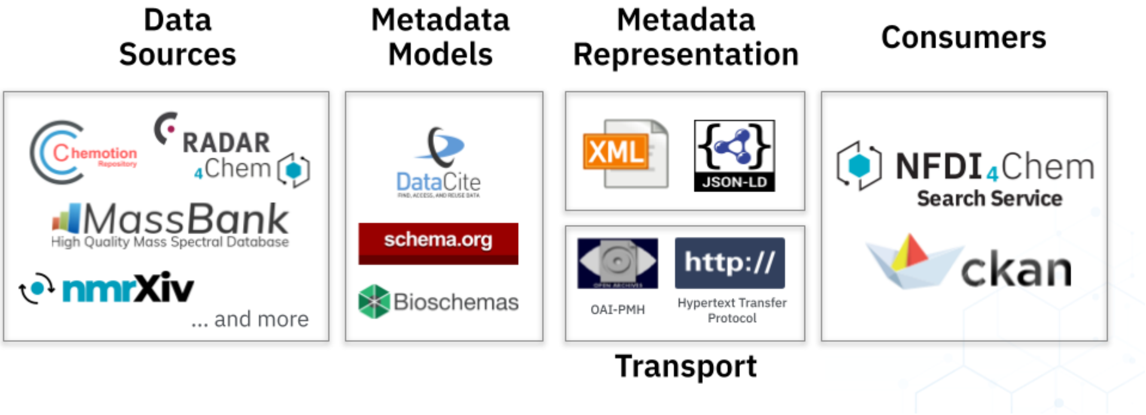
The schema.org initiative created a vocabulary that allows website owners to embed markup for many different types like people, events, organisations, or places using embedded JSON+LD, with the immediate benefit of better findability [1]. Bioschemas [2] is a community effort to improve the FAIRness of data of resources in the Life sciences by defining specific metadata schemas as JSON+LD and exposing that metadata from web based resources that have adopted it. To this end, it offers some tailored types that are readily applicable in many disciplines in the natural and life-sciences. Some of the types (e.g., BioChemEntity, ChemicalSubstance, Gene, MolecularEntity, Protein, and Taxon) have been picked up into schema.org. In addition to the types, Bioschemas also offers some validation and harvesting tools, making it easier to comply with specifications and consume the markup.

Libraries have been sharing the metadata about their own content to improve the findability of a book or article across libraries. The de facto standard to support discovery, presentation, and analysis of data originating from compliant archives is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [3].

Within the NFDI, the NFDI4Chem consortium has set out to create an open and FAIR infrastructure for research data management in chemistry, initially focusing on data related to molecules and reactions including data for their experimental and theoretical characterisation [4]. To successfully achieve this goal, we need to merge the two aforementioned technologies and consolidate the metadata into a central location where it can be indexed by chemistry aware search services, and equally well by generic data discovery services.

2. Modeling of chemistry-specific metadata and Design of metadata workflows

The NFDI4Chem federation of data repositories has initially started with existing repositories that had already implemented metadata formats and interfaces. These repositories were the first ones considered for harvesting and aggregating metadata into a harmonised metadata store and NFDI4Chem search service.



To demonstrate the feasibility, but also challenges and requirements when designing an architecture and large-scale system to provide harmonised, federated, disparate metadata from different realms and allow to harvest, search, and integrate this information, we have connected several existing methods and systems for chemistry research data.

The NFDI4Chem has several resources which serve (or plan to serve) Schema.org, i.e., Bioschemas metadata. MassBank [5] is an open spectral database, and it has been supporting Bioschemas markup for the individual records as Dataset and MolecularEntity since 2018. The Chemotion repository is well integrated with the Chemotion Electronic Lab Notebook. Since the repository already has a powerful REST API, we were able to develop a light-weight conversion from the Chemotion data types to schema.org MolecularEntity and Dataset types.

Designed for NMR data, nmrXiv is an open repository for FAIR NMR spectroscopy data. The data model is closely following the ISA model [6]. Bioschemas markup was integrated early in the architecture of nmrXiv. RADAR4Chem is a generic repository that can capture domain-specific metadata. The following table summarises supported types.

	Chemotion Repository	nmrXiv	MassBank	RADAR4Chem
DataCatalog	✓	✓	✓	
Study	✓	✓		
CreativeWork	✓	✓		
Person	✓	✓		✓
LabProtocol	✓			
ChemicalSubstance	✓	✓	✓	
MolecularEntity	✓	✓		
Dataset	✓	✓	✓	✓
DataDownload	✓	✓		

To bridge the gap between the semantic-web driven JSON+LD metadata and the established workflows in libraries, we need to design an approach to serve JSON+LD via OAI-PMH. Since, by design, the OAI responses are XML documents, the JSON+LD markup from the individual resources needs to be encapsulated in an XML CDATA element. The FIZ OAI-Provider is an open-source software that has been developed for efficiently serving high-volumes of metadata from large resources via the OAI-PMH protocol. The architecture consists of a frontend, backend, the Cassandra document database, and Elasticsearch as an index. For demonstration purposes, we have imported metadata items from the NFDI4Chem resources into a test instance of the OAI-PMH server.

The NFDI4Chem search service is based on CKAN, an open-source data management system, and can harvest metadata through several mechanisms, including an OAI-PMH harvester, as well as JSON+LD harvesting from HTML pages. It has search capabilities relevant in the chemistry domain, including a search for measurement types and molecular structures.

3. Conclusion

We have integrated the capability to serve Bioschemas-compliant metadata into the development versions or local prototypes of several resources in the NFDI4Chem realm, demonstrating that it is possible to load the metadata as JSON+LD into an OAI-PMH server and subsequently harvest into the NFDI4Chem search service through a modified CKAN OAI-PMH harvesting module.

In the future, these developments need to be integrated into the production services in NFDI4Chem. We also invite users with similar requirements to contact us and collaborate to add the approach to the (de facto) standards in metadata management.

One of the main efforts here is the unification of the disparate metadata realms across these resources in chemistry. Obtaining the agreement among all participating resources will take several rounds of prototyping, refining, and implementations. It is not always possible to map the entire data model of the resources to existing schema types, and the agreement might be that, in the case of metadata, some compromises need to be made.

Competing interests

The authors declare that they have no competing interests.

Funding

The consortia are funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the NFDI Funding Programme. The NFDI4Chem project under grant no. 441958208, the NFDI4DataScience project under no. 460234259.

Acknowledgement

The authors acknowledge the BioHackathon Germany 2022 in Lutherstadt Wittenberg, which hosted the "*(Bio)Schemas4NFDI*" project where some of the ideas were discussed and prototyped.

References

[1] "Introducing schema.org: Search engines come together for a richer web," *Official Google Blog*. https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html (accessed Jan. 29, 2023).

[2] F. Michel and The Bioschemas Community, "Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites," *Biodiversity Information Science and Standards*, vol. 2. p. e25836, 2018. doi: 10.3897/biss.2.25836.

[3] C. Lagoze and H. Van de Sompel, "The making of the open archives initiative protocol for metadata harvesting," *Libr. Hi Tech*, vol. 21, no. 2, pp. 118–128, Jun. 2003, doi: 10.1108/07378830310479776.

[4] C. Steinbeck *et al.*, "NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany," *RIO journal*, vol. 6, p. e55852, Jun. 2020, doi: 10.3897/rio.6.e55852.

[5] H. Horai *et al.*, "MassBank: a public repository for sharing mass spectral data for life sciences," *J. Mass Spectrom.*, vol. 45, no. 7, pp. 703–714, Jul. 2010, doi: 10.1002/jms.1777.

[6] S.-A. Sansone *et al.*, "Toward interoperable bioscience data," *Nat. Genet.*, vol. 44, no. 2, pp. 121–126, Jan. 2012, doi: 10.1038/ng.1054.