1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.232

Published: 07 Sept. 2023

Investigating the Landscape of Ontologies for Catalysis Research Data Management

Alexander S. Behr[1][https://orcid.org/0000-0003-4620-8248], Hendrik Borgelt[1][https://orcid.org/0000-0001-5886-7860], Taras Petrenko[2][https://orcid.org/0000-0002-0049-4835], Mark Dörr[3][https://orcid.org/0000-0003-3270-6895], Norbert Kockmann[1][https://orcid.org/0000-0002-8852-3812]

[1] Dept. of Biochemical and Chemical Engineering, TU Dortmund University, Germany

[2] High-Performance Computing Center Stuttgart (HLRS), University of Stuttgart, Germany

[3] Dept. of Biotechnology & Enzyme Catalysis, University of Greifswald, Germany

Abstract:

This work provides a survey of ontologies for catalysis research to improve the findability, accessibility, interoperability, and reusability (FAIRness) of research data. Applying tools that are commonly used by lab scientists, ontologies relevant to catalysis research are classified in a simple, well formatted spreadsheet template (Excel). This enables a scientist and domain expert without programming skills to evaluate a certain ontology. The entries of this template are then processed and visualized through automated creation of markdown files on GitHub using Python scripts. Furthermore, ontology mapping by searching for similar pairs of classes across different ontologies is performed, using the outcome of the ontology classification. This work contributes to the development of ontologies for catalysis research, facilitating better data integration and knowledge sharing while reusing existing semantic artefacts.

Keywords: Ontology Collection, Catalysis, Semantic Web, Classification

1. Introduction

As digitization of the scientific community advances, the need for FAIR (Findable, Accessible, Interoperable, Reusable) data rises to ensure machine-processability of data. Enabling a higher data FAIRness, ontologies represent knowledge explicitly in a machine-understandable way. [1] Furthermore, research data occurring in the field of catalysis research often is complex and diverse. Thus, the NFDI4Cat consortium focuses on ontology development for the catalysis research domain. [2]

To enhance semantic interoperability and compliance with existing ontologies, a collection of ontologies and semantic artefacts was created with importance to the data value chain of catalysis research. [3] Some of these ontologies are not easily reusable and do not provide proper documentation. This work presents a reiteration of the initial ontology landscape for catalysis research. The workflow and software is developed to be as reusable as possible, to enable other domains for such ontology classification.

2. Methods

To identify suitable ontologies, ontologies listed in the OLS [4] and BioPortal [5] are screened by look-up of different keywords. Additionally, the ontologies listed in [3] and [6] are considered.

The ontology survey is conducted with the help of a well formatted and intuitively designed spreadsheet template (Excel) to simplify access and handling of the ontology collection, capturing the relevant information on each ontology. For each ontology, such a template is filled in consisting of five topics and a comment section listed in Table 1 along the exact content included in each topic.

Table 1. Classification scheme of the ontologies. Information regarding the five topics is gathered for each ontology to classify the ontologies regarding the content of each topic.

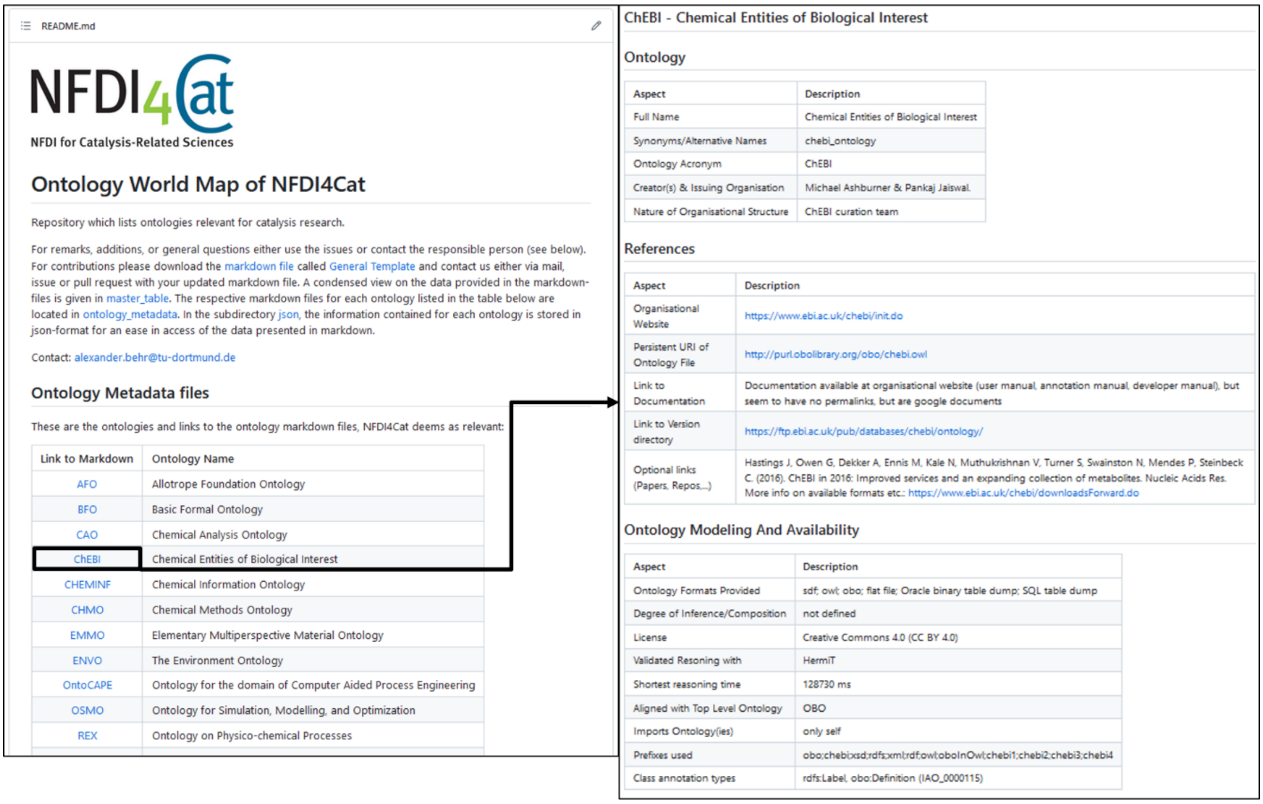
Topic	**Content**
General information on the ontology	Ontology name, alternative names, ontology acronym, creator(s) & issuing organization, kind of organizational structure
References	Organizational website, persistent URI of ontology file, link to documentation, link to version directory, additional links
Ontology modeling and availability	Provided ontology formats (ttl, owl, …), degree of inference and composition (inferred, non-inferred, compacted, …), license, working reasoners, shortest reasoning time, alignment with TLO, ontology imports, prefixes used, class annotation types
Classification of contained domains of interest	Biocatalysis, heterogenous catalysis, homogenous catalysis, chemical substance modeling, material modeling, process modeling, synthesis data, operando data, performance data, characterisation data, heat, transport and kinetic data, process design, energy and cost data, top level ontology
Ontology characteristics	Axioms, logical axiom count, declaration, class count, object property count, data property count, individual count, annotation property count
Comments	Any additional comments or remarks on topics not covered by the other topics

To facilitate documentation and access, the content of the Excel file is used to automatically generate markdown files, that contain simplified, text-based formatting instructions and can be rendered similarly to HTML. Rendering the markdown files in GitHub provides a comprehensive and interactive overview of each ontology, making it easier for researchers to assess the suitability of an ontology for their research needs. The data is then imported by Python to generate JSON files increasing machine-readability of the results, which eases further use of the data.

Overall, the Excel template and automated generation of markdown and JSON files enable efficient data collection, documentation, and access. This helps in automated detection of similar classes (mapping) between ontologies, too. For this, a Python script is used that detects similarities of ontology classes of two ontologies based on similar labels, prefLabels, altLabels, class names, and IRIs.

3. Results

The survey in this contribution classifies 28 ontologies containing the data as listed in Table 1. After conversion of the Excel-file to markdown files for the respective ontologies, the markdown files are visualized in the browser page of the GitHub repository. This also allows for a simple and clear presentation of the data, accessible without restrictions. The repository landing page contains a readme-file listing some general information about the ontology collection. Additionally, the acronym and the name of each ontology are listed in a table. As the markdown file allows for linking between different files, clicking on an acronym of an ontology directly redirects to the respective markdown file within the repository. The respective opened markdown file is visualized, too, and lists the information as given in Table 1. Figure 1 depicts the visualization of the repository readme file (left) and a resulting ontology information page (right) of the ChEBI ontology.



README.md

NFDI4Cat

NFDI for Catalysis-Related Sciences

Ontology World Map of NFDI4Cat

Repository which lists ontologies relevant for catalysis research.

For remarks, additions, or general questions either use the issues or contact the responsible person (see below). For contributions please download the markdown file called General Template and contact us either via mail, issue or pull request with your updated markdown file. A condensed view on the data provided in the markdown-files is given in master_table. The respective markdown files for each ontology listed in the table below are located in ontology_metadata. In the subdirectory json, the information contained for each ontology is stored in json-format for an ease in access of the data presented in markdown.

Contact: alexander.behr@tu-dortmund.de

Ontology Metadata files

These are the ontologies and links to the ontology markdown files, NFDI4Cat deems as relevant:

Link to Markdown	Ontology Name
AFO	Allotrope Foundation Ontology
BFO	Basic Formal Ontology
CAO	Chemical Analysis Ontology
ChEBI	Chemical Entities of Biological Interest
CHEMINF	Chemical Information Ontology
CHMO	Chemical Methods Ontology
EMMO	Elementary Multiperspective Material Ontology
ENVO	The Environment Ontology
OntoCAPE	Ontology for the domain of Computer Aided Process Engineering
OSMO	Ontology for Simulation, Modelling, and Optimization
REX	Ontology on Physico-chemical Processes

ChEBI - Chemical Entities of Biological Interest

Ontology

Aspect	Description
Full Name	Chemical Entities of Biological Interest
Synonyms/Alternative Names	chebi_ontology
Ontology Acronym	ChEBI
Creator(s) & Issuing Organisation	Michael Ashburner & Pankaj Jaiswal.
Nature of Organisational Structure	ChEBI curation team

References

Aspect	Description
Organisational Website	https://www.ebi.ac.uk/chebi/init.do
Persistent URI of Ontology File	http://purl.obolibrary.org/obo/chebi.owl
Link to Documentation	Documentation available at organisational website (user manual, annotation manual, developer manual), but seem to have no permalinks, but are google documents
Link to Version directory	https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/
Optional links (Papers, Repos,...)	Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. More info on available formats etc.: https://www.ebi.ac.uk/chebi/downloadsForward.do

Ontology Modeling And Availability

Aspect	Description
Ontology Formats Provided	sdf; owl; obo; flat file; Oracle binary table dump; SQL table dump
Degree of Inference/Composition	not defined
License	Creative Commons 4.0 (CC BY 4.0)
Validated Resoning with	HermiT
Shortest reasoning time	128730 ms
Aligned with Top Level Ontology	OBO
Imports Ontology(ies)	only self
Prefixes used	obo;chebi;xsd;rdfs;xml;rdf;owl;oboInOwl;chebi1;chebi2;chebi3;chebi4
Class annotation types	rdfs:Label, obo:Definition (IAO_0000115)

Figure 1. Visualization of the ontology classification via markdown files on GitHub. The repository readme file (left) lists the ontologies and links to the markdown files describing the respective ontology (right) according to the classification listed in Table 1.

After classification of the ontologies, the search for class similarities is performed automatically for each pair of ontologies. This helps to identify close ontologies and get common classes to extend existing ontologies. The resulting list of common classes for each pair of ontologies is depicted in Figure 2 as heat map with low count of classes in red and high count of classes in green for 11 ontologies. The CHEBI ontology has, for example, 937 common classes with the ENVO ontology, which is the intersection of the two largest ontologies.

	AFO	BFO	CAO	CHEBI	CHMO	EMMO	ENVO	OSMO	REX	SBO	VIMMP
AFO	3028										
BFO	35	36									
CAO	27	2	446								
CHEBI	43	0	39	182375							
CHMO	232	12	53	22	3102						
EMMO	2	0	5	2	0	199					
ENVO	158	26	59	937	32	2	6997				
OSMO	6	0	0	0	0	0	0	173			
REX	8	0	2	0	18	0	6	0	553		
SBO	26	2	3	12	3	0	14	1	11	695	
VIMMP	40	2	12	2	3	25	16	7	0	8	1082

Figure 2. Heatmap of the 11 ontologies investigated. Green entries show a high absolute number of common classes, while red indicates the contrary.

Data availability statement

The data, code and markdown files presented in this abstract will be available at GitHub here: https://github.com/AleSteB/Ontology-Overview-of-NFDI4Cat

Author contributions

Conceptualization: A.S.B., H.B., T.P., M.D.; Methodology: A.S.B., H.B.; Software: A.S.B.; Validation: A.S.B.; Data Curation: A.S.B., H.B.; Writing – Original Draft: A.S.B., Writing – Review & Editing: N.K., M.D., T.P.; Visualization: A.S.B.; Supervision: A.S.B., N.K.

Competing interests

The authors declare that they have no competing interests.

Funding

The Deutsche Forschungsgemeinschaft (DFG) is acknowledged for funding this research as part of the Nationale Forschungsdateninfrastruktur (NFDI) initiative (grant No.: NFDI/2-1 - 2021).

Acknowledgement

The authors would like to thank all members of the NFDI4Cat Task Area 1.

A.S.B. thanks the networking program 'Sustainable Chemical Synthesis 2.0' (SusChemSys 2.0) for the support and fruitful discussions across disciplines.

References

1. T. R. Gruber, "A translation approach to portable ontology specifications," Knowl.Acquis., vol. 5, no. 2, pp. 199–220, 1993, doi: https://doi.org/10.1006/knac.1993.1008
2. C. Wulf et al., "A Unified Research Data Infrastructure for Catalysis Research – Challenges and Concepts," ChemCatChem 2021, 13, 3223 doi: https://doi.org/10.1002/cctc.202001974
3. M. Horsch et al., "Interoperability and Architecture Requirements Analysis and Metadata Standardization for a Research Data Infrastructure in Catalysis," In: Pozanenko, A., Stupnikov, S., Thalheim, B., Mendez, E., Kiselyova, N. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2021. Communications in Computer and Information Science, vol 1620. Springer, Cham., 2022, doi: https://doi.org/10.1007/978-3-031-12285-9_10
4. S. Jupp et al., "A new Ontology Lookup Service at EMBL-EBI," In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015, URL: https://www.ebi.ac.uk/ols/index
5. N.F. Noy et al., "BioPortal: ontologies and integrated data resources at the click of a mouse," Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W170-3. Epub 2009, URL: https://bioportal.bioontology.org/
6. P. Strömert et al., "Ontologies4Chem: the landscape of ontologies in chemistry," Pure and Applied Chemistry, vol. 94, no. 6, 2022, pp. 605-622. https://doi.org/10.1515/pac-2021-2007