

Marius Michaelis¹ <https://orcid.org/0000-0002-6437-7152>, Günther Neher¹ <https://orcid.org/0000-0002-3548-3300>, Tobias Höhnow², Bernd Ritschel

¹ Faculty of Information Sciences, University of Applied Sciences Potsdam; ² GFZ German Research Centre for Geosciences, Potsdam

Corresponding Author: Günther Neher, g.neher@fh-potsdam.de

Abstract We present a proof of concept that follows the vision to make geoscientific research data easily findable. To achieve this, **metadata records of research data publications are integrated by means of Linked Data principles and semantic technologies**. In the course of this, not only the findability of the research data publications is improved, but also the interoperability of the associated metadata. By transforming metadata into the RDF format and integrating this data using semantic mappings, our proof of concept demonstrates what concrete steps can be taken to make research data publications **FAIR**, with a focus on **findability** and **interoperability**.

Proof of concept

Our proof of concept is called the *World Data System Vocabulary Broker*¹. This prototypical demonstrator connects the metadata vocabularies GCMD², SPASE³, ESPAS⁴, UAT⁵ and GEMET⁶. To do so, a for now relatively simple mapping algorithm identifies skos:closeMatch⁷ and skos:relatedMatch⁸ relationships between the terms of the different vocabularies. Data publications can thus be found not only via keywords with which they were originally indexed, but also via equivalent keywords from other vocabularies. In practice, this means improved **findability** of related research data. This is especially the case if they originate from different research projects and are therefore described using different vocabularies. The **accessibility** is ensured by the decentralized repositories that manage the research data publications.

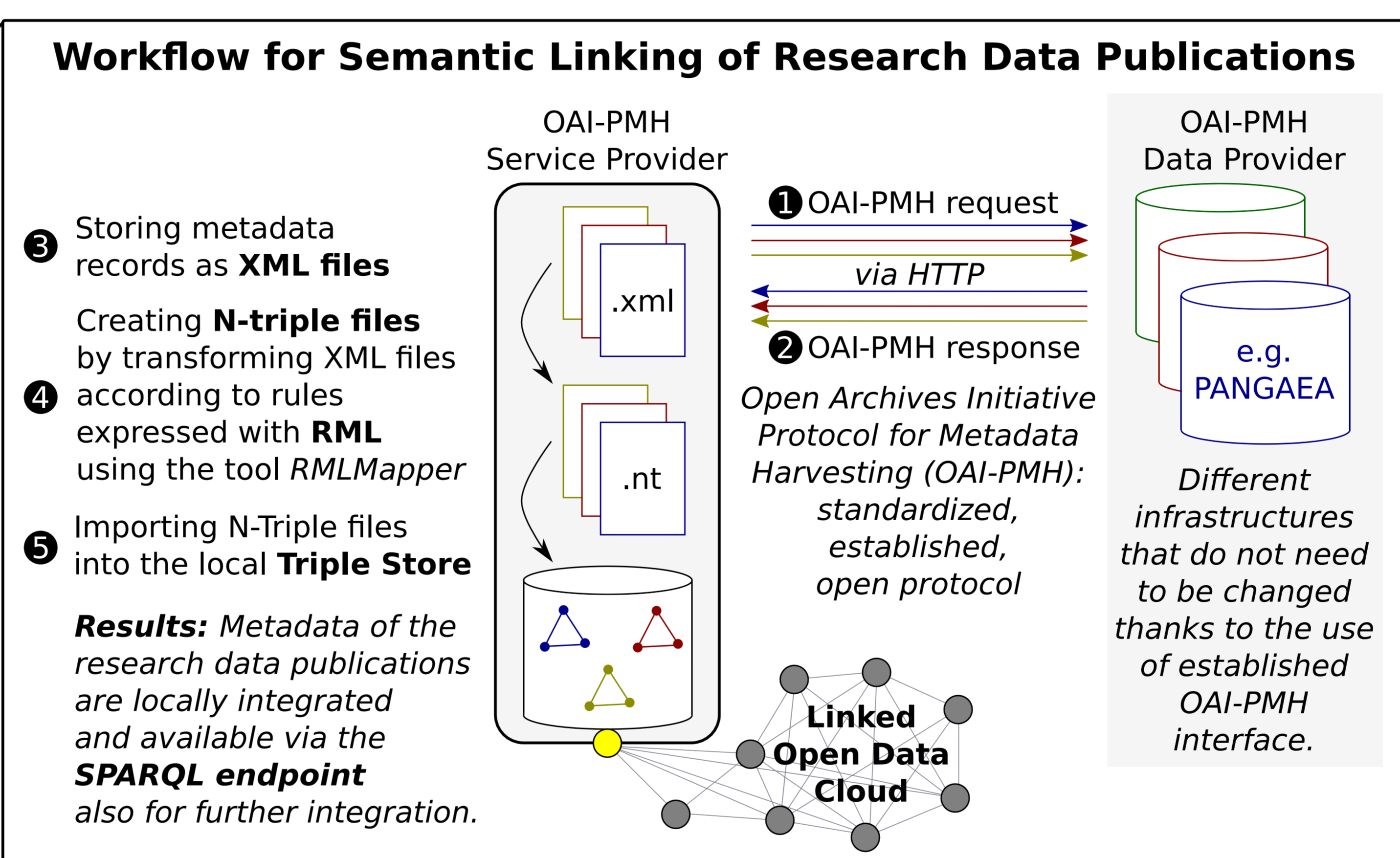
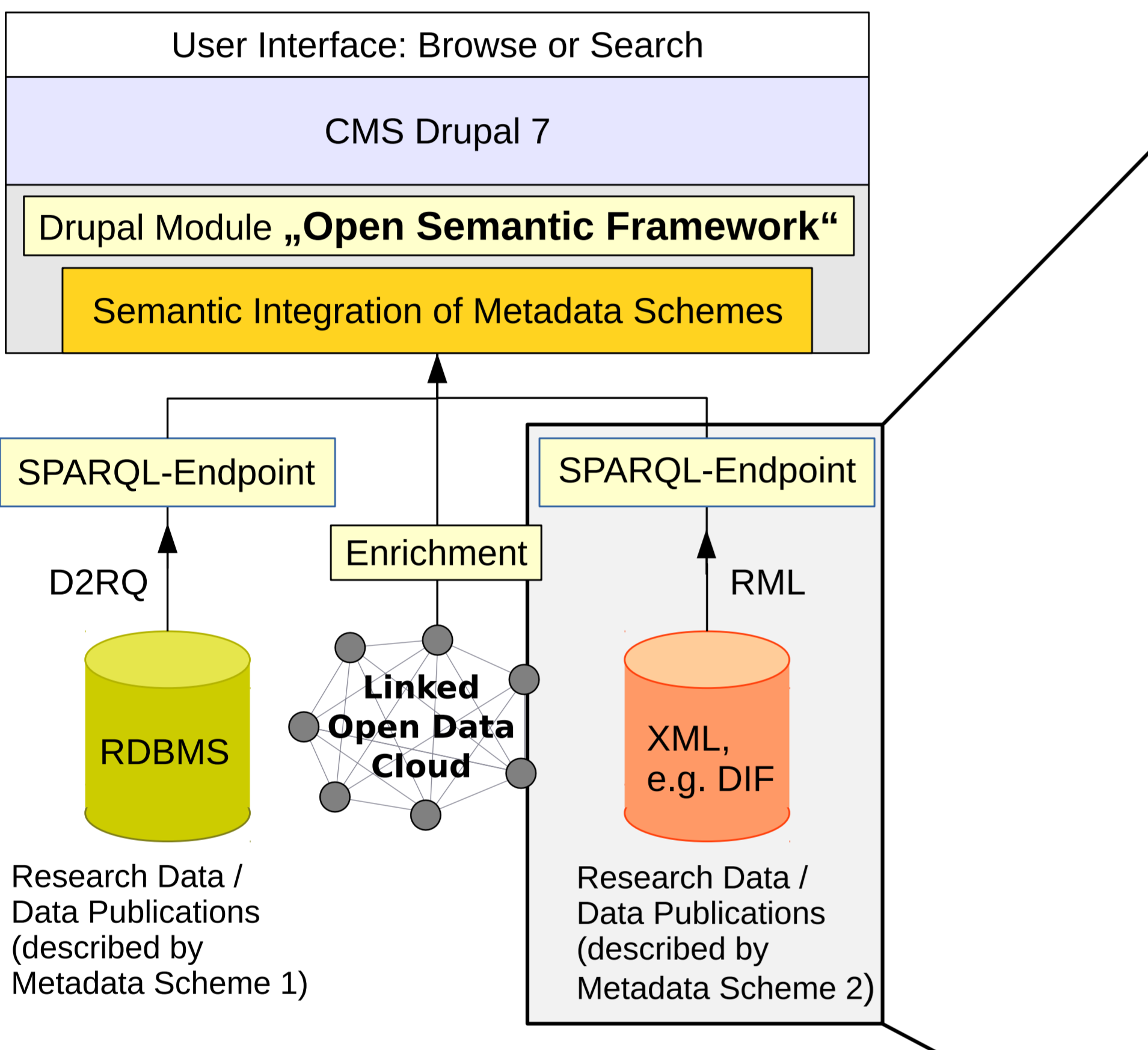
Our semantic search is implemented using the open source semantic content management system *Open Semantic Framework*⁹. It acts as a service provider and has access to the decentralized data providers via SPARQL¹⁰ endpoints. The service provider system contains semantic mappings between different vocabularies as well as between different metadata models. Data providers which do not hold their data in RDF¹¹ format can be integrated without having to change their infrastructure. Access can be provided using a suitable middleware, e.g. D2RQ¹² or RML¹³. [1]

FAIR principles & semantic technologies

Geoscientific research data publications are currently distributed across different repositories and described using different vocabularies. This means that they are accessible, but not easily findable. Therefore, initiatives such as the *European Open Science Cloud* (EOSC)¹⁴ and *GO FAIR*¹⁵ promote FAIR data. FAIR refers to a set of four principles: data must be *findable*, *accessible*, *interoperable*, and *reusable* [2]. Both EOSC and GO FAIR follow the recommendations of the *European Commission expert group on FAIR data* [3][4]. This expert group recommends, among others, semantic technologies to achieve FAIR data [5]. Following this recommendation, we apply semantic technologies in the workflow presented here. While services like *re3data.org*¹⁶ enable the findability of data repositories as a whole, our prototype aims at findability on the level of individual data publications.

Objective of the workflow

Main subject of this poster is the workflow illustrated below. It integrates metadata records of research data publications in order to make them available via the *World Data System Vocabulary Broker*. The workflow's objective is therefore to achieve **interoperability** of metadata by transforming it into the RDF format. The metadata is queried in the metadata format DIF¹⁷ from data providers using the OAI-PMH¹⁸. After the transformation from XML to RDF, the metadata is made available via a triple store.



```

XML excerpt
Line 01 <Parameters>
02 <Category>EARTH_SCIENCE</Category>
03 <Topic>LAND_SURFACE</Topic>
04 <Term>LAND_USE/LAND_COVER</Term>
05 <VariableFormulation q1:XPath;
06 </Parameters>
07 <Parameters>
08 <Category>EARTH_SCIENCE</Category>
09 <Topic>AGRICULTURE</Topic>
10 <Term>AGRICULTURAL_PLANT_SCIENCE</Term>
11 </Parameters>
    
```

```

RML excerpt
Line 01 <http://www.example.com/Mapping#Parameters>
02 rml:logicalSource [
03   rml:source "xml-to-be-processed/input.xml";
04   rml:referenceFormulation q1:XPath;
05   rml:iterator "/record/metadata/Dif/Parameters";
06 ];
07 rr:subjectMap [
08   rr:termType rr:BlankNode;
09   rr:class dif:Parameters;
10 ];
11 rr:predicateObjectMap [
12   rr:predicate dif:Category;
13   rr:objectMap [
14     rml:reference ". /Category";
15     rr:datatype xsd:string;
16 ];
17 ];
18 rr:predicateObjectMap [
19   rr:predicate dif:Topic;
20   rr:objectMap [
21     rml:reference ". /Topic";
22     rr:datatype xsd:string;
23 ];
24 ];
25 rr:predicateObjectMap [
26   rr:predicate dif:Term;
27   rr:objectMap [
28     rml:reference ". /Term";
29     rr:datatype xsd:string;
30 ];
31 ];
32 rr:predicateObjectMap [
33   rr:predicate dif:Variable_Level_1;
34   rr:objectMap [
35     rml:reference ". /Variable_Level_1";
36     rr:datatype xsd:string;
37 ];
38 ];
    
```

Demonstrator: <http://wdcosf.fh-potsdam.de/>

Step 1: Search

Search results

SNOW **GEMET concept**

Description: The most common form of frozen precipitation, usually flakes or starlike crystals, matted ice needles, or combinations, and often rime-coated.

relatedMatch: Snow Pellets, Snow Storms, Snow Melt, Snow Cover, Snow Grains, Snow Depth, Snow Grains

inScheme: GEMET Vocabulary

closeMatch: Snow

broader: atmospheric precipitation

Step 2: View concept

SNOW **GEMET concept**

definition: The most common form of frozen precipitation, usually flakes or starlike crystals, matted ice needles, or combinations, and often rime-coated.

relatedMatch: Snow Pellets, Snow Storms, Snow Melt, Snow Cover, Snow Grains, Snow Depth

inScheme: GEMET Vocabulary

closeMatch: Snow

broader: atmospheric precipitation

Step 3: Access publication

datapublication(s)

• Supplement to: Monitoring snow depth by GNSS reflectometry in built-up areas: A case study for Wettzell, Germany (DOI: <http://dx.doi.org/10.5880/GFZ.1.1.2016.001>)

Explanation and implementation of the workflow

First, metadata is harvested via the OAI-PMH using a PHP¹⁹ script. The metadata records received are stored as XML²⁰ files. In the next step, the XML files are transformed into triples according to the RDF data model. To do this, the DIF metadata schema has been represented as an OWL²¹ ontology. Then the XML elements of the metadata records have been mapped to classes and properties of the unofficial DIF ontology. For the mapping of XML to RDF, the *RDF Mapping Language* (RML)¹³ has been used. The transformation rules expressed with RML are applied by the tool *RMLMapper*²² which generates the corresponding triples. Finally, the generated triples are imported into a local triple store, where they are available via a SPARQL endpoint. It is intended that the workflow is performed regularly, for example every day at midnight. In each case, the new metadata is queried, transformed and imported into the triple store. In other words, a bulk transformation takes place – not an on-the-fly transformation.

Outlook

The for now simple workflow presented demonstrates that the principle of using established OAI-PMH interfaces to integrate metadata with the help of semantic technologies works. Remaining challenges are the maintenance of triples, e.g. in case of modifications of existing metadata, as well as the improvement, documentation and publication of the ontologies used for semantic annotation. Our prototype as a whole is already functional and can serve as a proof of concept. Planned improvements will include the optimization of the harvesting and RDF-transformation workflow as well as the integration of more data sources (currently only GFZ Data Services²³ and PANGAEA²⁴ are integrated). Besides, the semantic linking is currently solely based on the mapping of SKOS-concepts via skos:relatedMatch and skos:closeMatch. To increase findability, future work will try to establish a broader spectrum of semantic relations between data sources on the basis of other metadata elements.

References

- Ritschel, Bernd; Borchert, Friederike; Kneitschel, Gregor; Neher, Günther; Schildbach, Susanne; Iyemori, Toshihiko; Koyama, Yukinobu; Yatagai, Akiyo; Hori, Tomoaki; Hagood, Mike; Belehaki, Anna; Galkin, Ivan; King, Todd (2016) *Experiments using Semantic Web technologies to connect IUGONET, ESPAS and GFZ ISDC data portals*: Earth, Planets and Space, vol. 68, no. 1. DOI: [10.1186/s40623-016-0542-z](https://doi.org/10.1186/s40623-016-0542-z).
- Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I. J. J.; Appleton, Gabrielle; Axton, Myles; Baak, Aric; Blomberg, Niklas; Boiten, Jan-Willem; de Silva Santos, Luiz B.; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alastair J. G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; Horn, Peter A. C. T.; Hood, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend (2016) *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific data, vol. 3. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- Directorate-General for Research and Innovation (2018) *Prompting an EOSC in practice: Final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC)*. European Commission.
- GO FAIR (2018) *Strategy* [Online]. Available at <https://www.go-fair.org/go-fairinitiative/strategy/> (Accessed 2019-01-31).
- Directorate-General for Research and Innovation (2018) *Turning FAIR data into reality: Final report and action plan from the European Commission expert group on FAIR data*. European Union.

Annotations

- World Data System Vocabulary Broker – Proof of Concept: <http://wdcosf.fh-potsdam.de/>.
- GCMD: Global Change Master Directory.
- SPASE: Space Physics Archive Search and Extract.
- ESPAS: Near Earth Space Data Infrastructure for a Science project (Vocabulary).
- UAT: Unified Astronomy Thesaurus.
- GEMET: General Multilingual Environmental Thesaurus.
- SKOS: Simple Knowledge Organization System, Property relatedMatch, see <http://www.w3.org/2004/02/skos/core#relatedMatch>.
- SKOS: Simple Knowledge Organization System, Property relatedMatch, see <http://www.w3.org/2004/02/skos/core#relatedMatch>.
- OpenSemanticFramework: <http://opensemanticframework.org/>.
- SPARQL: SPARQL Protocol and RDF Query Language, see <https://www.w3.org/TR/sparql11-overview/>.
- RDF: Resource Description Framework, see <https://www.w3.org/TR/rdf11-primer/>.
- D2RQ Platform: Accessing Relational Databases as Virtual RDF Graphs, see <http://d2rq.org/>.
- RML: RDF Mapping Language, see <http://rml.io/>.
- EOSC: European Open Science Cloud, see <https://www.eosc-portal.eu/>.
- GO (Global Open) FAIR initiative, see <https://www.go-fair.org/>.
- re3data.org: Registry of Research Data Repositories, see <https://www.re3data.org/>.
- DIF: Directory Interchange Format, see <https://ncmi.nasa.gov/rdf/dif/>.
- OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting, see <https://www.openarchives.org/OAI/openarchivesprotocol.html>.
- PHP: Hypertext Preprocessor, see <https://secure.php.net/>.
- XML: Extensible Markup Language, see <https://www.w3.org/XML/>.
- OWL: Web Ontology Language, see <https://www.w3.org/OWL/>.
- RML: RMLMapper-java, GitHub: <https://github.com/RMLMapper/rmlmapper-java>.
- GFZ Research Data Repository: <http://dataservices.gfz-potsdam.de/portal/>.
- PANGAEA: <https://www.pangaea.de/>.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1886, Mountain View, CA 94042, USA.

VISIT OUR DEMO and get further INFO