RDA DE e.V.; Helmholtz-Gemeinschaft (2019). RDA Deutschland Tagung 2019. GeoForschungsZentrum Potsdam, Germany, February 19-20, 2019.

Semantic Linking of Research Data Publications

Contributing to the FAIRification of Metadata Records – A Proof of Concept

Marius Michaelis[1] https://orcid.org/0000-0002-6437-7152, Günther Neher[1] https://orcid.org/0000-0002-3548-3300, Tobias Höhnow[2], 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 GCMD[2], SPASE[3], ESPAS[4], UAT[5] and GEMET[6]. To do so, a for now relatively simple mapping algorithm identifies skos:closeMatch[7] and skos:relatedMatch[8] 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.

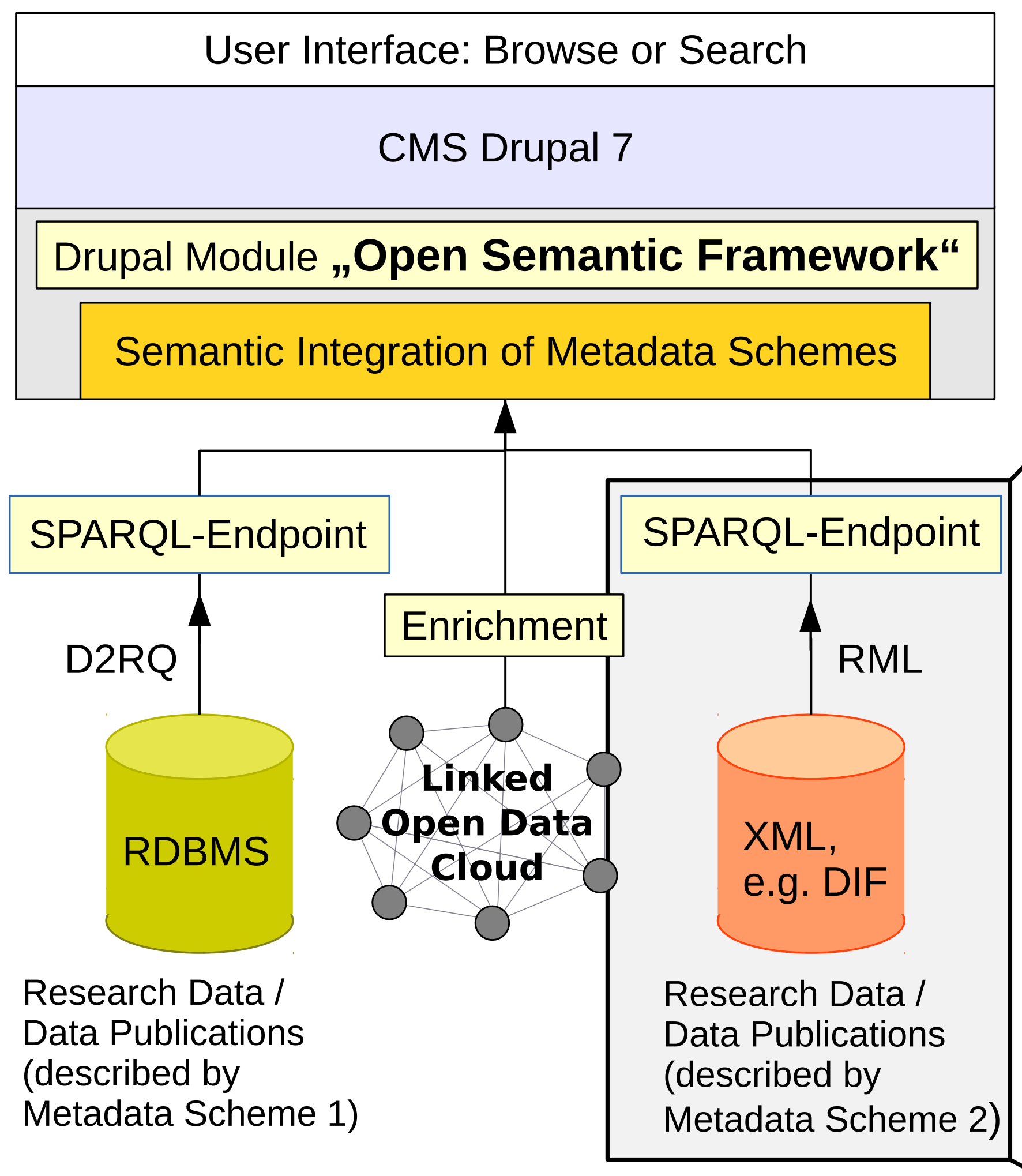
Our semantic search is implemented using the open source semantic content management system *Open Semantic Framework*[9]. It acts as a service provider and has access to the decentralized data providers via SPARQL[10] endpoints. The service provider system contains semantic mappings between different vocabularies as well as between different metadata models. Data providers which do not hold their data in RDF[11] format can be integrated without having to change their infrastructure. Access can be provided using a suitable middleware, e.g. D2RQ[12] or RML[13]. [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 DIF[17] from data providers using the OAI-PMH[18]. After the transformation from XML to RDF, the metadata is made available via a triple store.



User Interface: Browse or Search

CMS Drupal 7

Drupal Module „Open Semantic Framework“

Semantic Integration of Metadata Schemes

SPARQL-Endpoint

Enrichment

SPARQL-Endpoint

D2RQ

RML

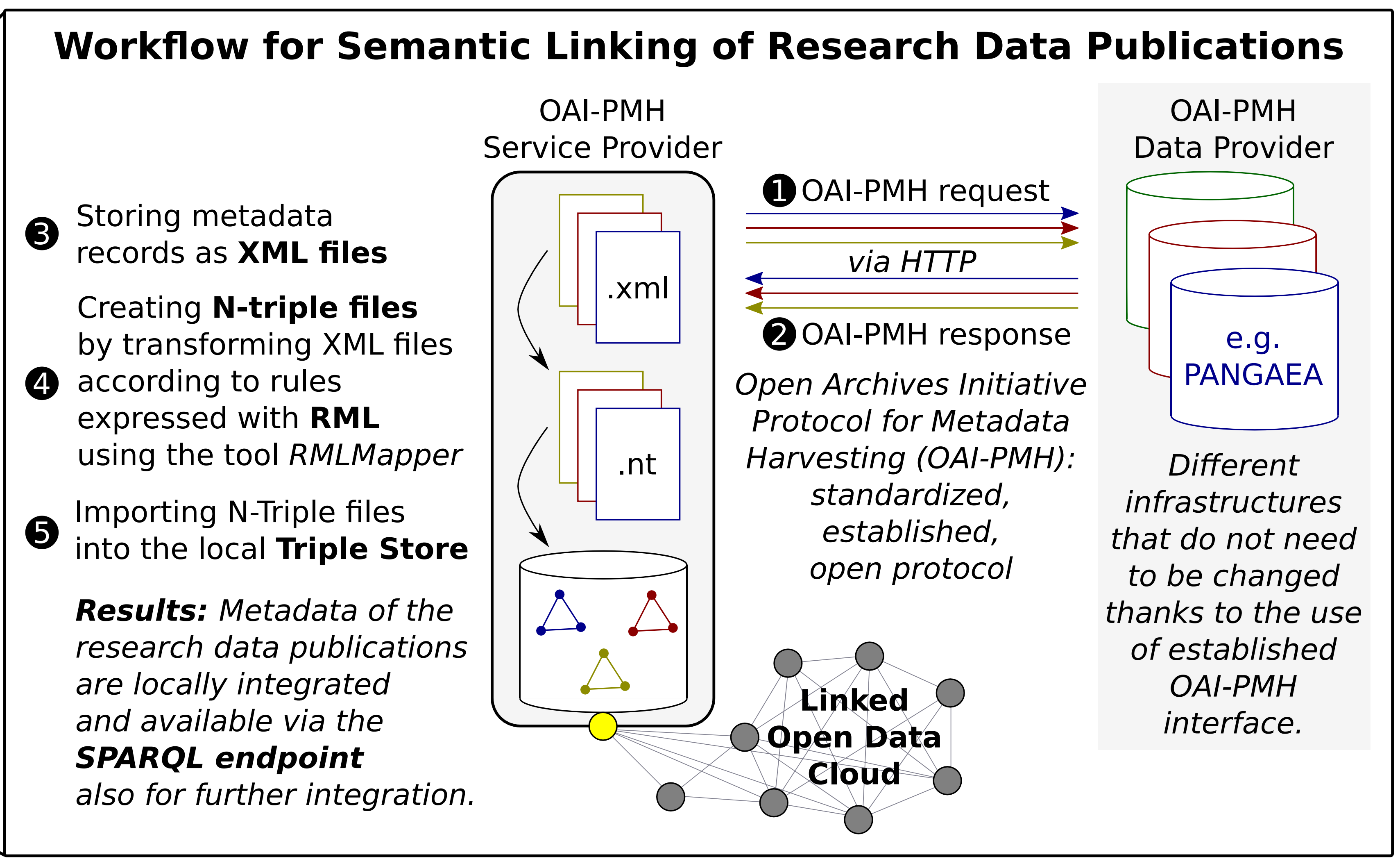
RDBMS

Linked Open Data Cloud

XML, e.g. DIF

Research Data / Data Publications (described by Metadata Scheme 1)

Research Data / Data Publications (described by Metadata Scheme 2)



Workflow for Semantic Linking of Research Data Publications

OAI-PMH Service Provider

OAI-PMH Data Provider

❶ OAI-PMH request

via HTTP

❷ OAI-PMH response

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): standardized, established, open protocol

❸ Storing metadata records as **XML files**

❹ Creating **N-triple files** by transforming XML files according to rules expressed with **RML** using the tool *RMLMapper*

❺ Importing N-Triple files into the local **Triple Store**

Results: *Metadata of the research data publications are locally integrated and available via the* ***SPARQL endpoint*** *also for further integration.*

Different infrastructures that do not need to be changed thanks to the use of established OAI-PMH interface.

XML excerpt

```
<Parameters>
     <Category>EARTH SCIENCE</Category>
     <Topic>LAND SURFACE</Topic>
     <Term>LAND USE/LAND COVER</Term>
     <Variable_Level_1>LAND COVER</Variable_Level_1>
</Parameters>
<Parameters>
     <Category>EARTH SCIENCE</Category>
     <Topic>AGRICULTURE</Topic>
     <Term>AGRICULTURAL PLANT SCIENCE</Term>
</Parameters>
```

RML excerpt

```
<http://www.example.com/Mapping#Parameters>
rml:logicalSource [
 rml:source "xml-to-be-processed/input.xml";
 rml:referenceFormulation ql:XPath;
 rml:iterator "/record/metadata/DIF/Parameters";
];
rr:subjectMap [
 rr:termType rr:BlankNode;
 rr:class dif:Parameters;
];
rr:predicateObjectMap [
 rr:predicate dif:Category;
 rr:objectMap [
  rml:reference "./Category";
  rr:datatype xsd:string
 ]
];
rr:predicateObjectMap [
 rr:predicate dif:Topic;
 rr:objectMap [
  rml:reference "./Topic";
  rr:datatype xsd:string
 ]
];
rr:predicateObjectMap [
 rr:predicate dif:Term;
 rr:objectMap [
  rml:reference "./Term";
  rr:datatype xsd:string
 ]
];
rr:predicateObjectMap [
 rr:predicate dif:Variable_Level_1;
 rr:objectMap [
  rml:reference "./Variable_Level_1";
  rr:datatype xsd:string
 ]
].
```

NT excerpt

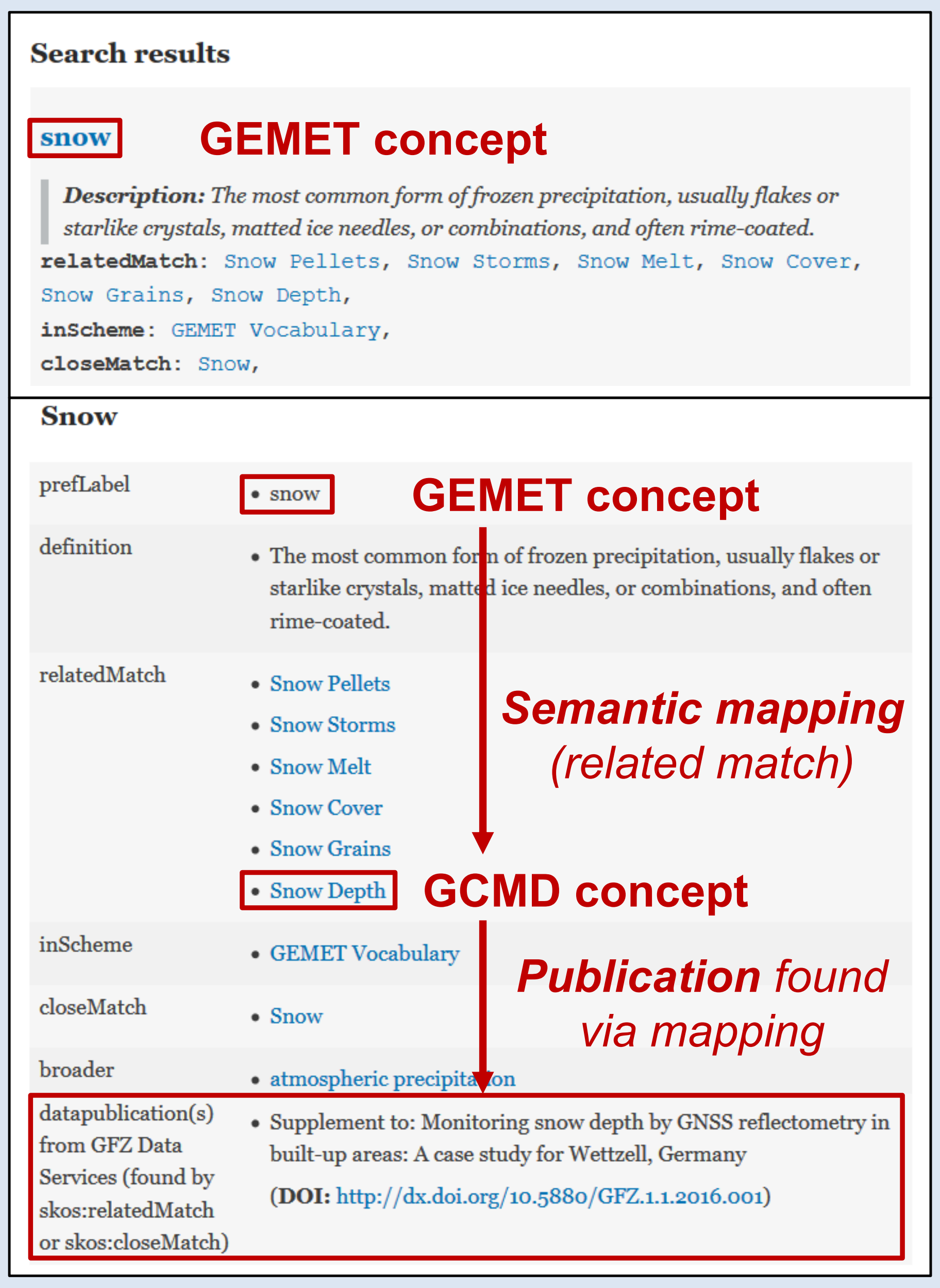
```
<http://www.example.com/dif-identifier-123>
    <http://www.example.com/Parameters> _:0 ,_:1
    .
_:0 <http://www.example.com/Category> "EARTH SCIENCE" ;
    <http://www.example.com/Topic> "LAND SURFACE" ;
    <http://www.example.com/Term> "LAND USE/LAND COVER" ;
    <http://www.example.com/Variable_Level_1> "LAND COVER"
    .
_:1 <http://www.example.com/Category> "EARTH SCIENCE" ;
    <http://www.example.com/Topic> "AGRICULTURE" ;
    <http://www.example.com/Term> "AGRICULTURAL PLANT
    SCIENCE" .
```

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

Step 1: Search

Step 2: View concept

Step 3: Access publication



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,

inScheme: GEMET Vocabulary,

closeMatch: Snow,

Snow

prefLabel	• snow
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
datapublication(s) from GFZ Data Services (found by skos:relatedMatch or skos:closeMatch)	• 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)

GEMET concept
GEMET concept
Semantic mapping (related match)
GCMD concept
Publication found via mapping

Explanation and implementation of the workflow

First, metadata is harvested via the OAI-PMH using a PHP[19] script. The metadata records received are stored as XML[20] 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 OWL[21] 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 Services[23] and PANGAEA[24] 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

[1] Ritschel, Bernd; Borchert, Friederike; Kneitschel, Gregor; Neher, Günther; Schildbach, Susanne; Iyemori, Toshihiko; Koyama, Yukinobu; Yatagai, Akiyo; Hori, Tomoaki; Hapgood, 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-x.

[2] Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I. J. J.; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da 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, Alasdair J. G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; Hoen, Peter A. C. 't; Hooft, 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.

[3] 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.

[4] GO FAIR (2018) *Strategy* [Online]. Available at https://www.go-fair.org/go-fairinitiative/strategy/ (Accessed 2019-01-31).

[5] 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

[1] World Data System Vocabulary Broker – Proof of Concept: http://wdcosf.fh-potsdam.de/.
[2] GCMD: Global Change Master Directory
[3] SPASE: Space Physics Archive Search and Extract
[4] ESPAS: Near Earth Space Data Infrastructure for e-Science project (Voacbulary)
[5] UAT: Unified Astronomy Thesaurus
[6] GEMET: General Multilingual Environmental Thesaurus
[7] SKOS: Simple Knowledge Organization System, Property closeMatch, see http://www.w3.org/2004/02/skos/core#closeMatch.
[8] SKOS: Simple Knowledge Organization System, Property relatedMatch, see http://www.w3.org/2004/02/skos/core#relatedMatch.
[9] OpenSemanticFramework: http://opensemanticframework.org/.
[10] SPARQL: SPARQL Protocol and RDF Query Language, see https://www.w3.org/TR/sparql11-overview/.
[11] RDF: Resource Description Framework, see https://www.w3.org/TR/rdf11-primer/.
[12] D2RQ Platform: Accessing Relational Databases as Virtual RDF Graphs, see http://d2rq.org/.
[13] RML: RDF Mapping Language, see http://rml.io/.
[14] EOSC: European Open Science Cloud, see https://www.eosc-portal.eu/.
[15] GO (Global Open) FAIR initiative, see https://www.go-fair.org/.
[16] re3data.org: Registry of Research Data Repositories, see https://www.re3data.org/.
[17] DIF: Directory Interchange Format, see https://gcmd.nasa.gov/r/u/difguide/.
[18] OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting, see https://www.openarchives.org/OAI/openarchivesprotocol.html.
[19] PHP: Hypertext Preprocessor, see https://secure.php.net/.
[20] XML: Extensible Markup Language, see https://www.w3.org/XML/.
[21] OWL: Web Ontology Language, see https://www.w3.org/OWL/.
[22] RMLio/rmlmapper-java, GitHub, https://github.com/RMLio/rmlmapper-java.
[23] GFZ Research Data Repository: http://dataservices.gfz-potsdam.de/portal/.
[24] 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 1866, Mountain View, CA 94042, USA.

VISIT our **DEMO** and get further **INFO**