GeoERA INFORMATION PLATFORM

Establishing the European Geological Surveys Research Area to deliver a Geological Service for Europe

Deliverable 4.3

GeoERA Project Vocabularies

Authors and affiliation:
M. Schiegl [GBA]
L. Sőrés [MBFSZ]
M. Pantaloni [ISPRA]
O. Johansson [SGU]
R. van Ede [TNO]

E-mail of lead author:
martin.schiegl@geologie.ac.at

Version: 12-11-2019

This report is part of a project that has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement number 731166.

Deliverable Data		
Deliverable number	D4.3	
Dissemination level	Public	
Deliverable name	GeoERA Project Vocabularies	
Work package	WP4, Semantic Harmonization Issues	
Lead WP/Deliverable beneficiary	GBA	
Deliverable status		
Submitted (Author(s))	12/11/2019	Martin Schiegl
Reviewed	28/11/2019	Martin Schiegl
Approved (Coordinator)	29/11/2019	Jørgen Tulstrup

D4.3 Report refers to the task

T4.2 Scope of vocabulary concepts, Guideline - Methods, models and workflows, Testing and support, URI Design, strategy –

T4.2 participants are GBA, IGME, ISPRA, SGU, TNO, CGS, GeoZS, MBFSZ, LfU, BGRM, GEUS, BGR

Abstract

Project vocabularies provide the opportunity to clarify expert knowledge and terminology in the form of project specific vocabulary concepts on a scientific level and to use them in datasets to code data. At the same time, parts of this vocabulary might be later included in international standards (e.g. INSPIRE or GeoSciML), if desired. By comparison, Project vocabularies are open collections of knowledge that, for example, may also contain deprecated, historical or only regionally relevant terms. In an ideal overall view, the sum of all Project vocabularies results in a knowledge database of bibliographically referenced terms that have been developed through scientific projects. Due to the consistent application of the data standards of Semantic Web and Linked Data nothing stands in the way of further use by modern technologies such as AI. The report explains what is meant by Project vocabularies in the context of GeoERA and examples of what problems, in semantics of data, can be solved by using them. In addition, project related methods and workflows around Linked Data, and SKOS in particular, are described.

Please note:

The present report aims more at a technical level for readers familiar with technologies and principles of Semantic Web and Linked Data. In addition, this paper will be finalized for report D4.4 and reconciled together with the deliverables D4.1. and D4.2 (Keyword Thesaurus) in June 2021. This refers especially to the chapters about “URI design” and the “documentation of vocabularies” actually created during the GeoERA project.

Acknowledgement:

thanks to Diepolder G., Feliachi A., Hernández Manchado J.R., Maul A. who were directly involved in this work

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

In geosciences, where nomenclature naturally has grown from regional approaches with limited cross-border harmonization, descriptive texts are often used for coding data whose meanings in the international context are not conclusively clarified. This leads to difficulties when cross border datasets are compiled. On one hand, this is caused by the national-language, regional and historical descriptions in geological map legends. On the other hand, it is related to the interdisciplinary orientation of the geosciences e.g. when concepts adopted from different areas have a different meaning. A consistent use and interpretation of data to international standards creates the potential for semantic interoperability. Datasets can then be integrated into international data infrastructures. But what if the interpretation to international standards is not possible, because there is none, or existing standards are not applicable? Then effort can be put into making the data machine-readable using knowledge representations based on Semantic Web and Linked Data principles. With a global and unique referenceability of concepts via web identifiers (URIs) and crosslinking them on the Web, Linked Data offers the necessary context for clarification of the meaning. This modern technology and approach ideally complements the mainstream GIS and relational database technologies in making data findable and semantic interoperable.

Semantic Harmonization of different data sets refers to the alignment of the meanings of the used data objects (entities, relations, attributes). The use of standardized data models and their descriptions already defines the semantics for entity types and relations. Standardized codes (e.g. GeoSciML, INSPIRE) and their descriptions should be used to align the meaning of textual attributes. Only in the case when a standardization is not practicable or feasible, we need to find another way to make datasets compatible. This is the case if no overarching standard is available e.g. the highly fragmented definition of geological formations as the core feature of all spatial representations of geological settings. Alternatively, such regional or historical names of geological formations can be processed as cross-linked concepts to get the context that specifies the meaning of these textual descriptive labels. Using Linked Data principles and SKOS references, a mapping to a global context on the (Semantic) Web can be defined. Therefore, WP4 will assist GeoERA projects in the generation of SKOS vocabularies in accordance with the Linked Data principles where standardized codes cannot replace textual attributes. To this purpose, WP4 is to provide or recommend an infrastructure for publishing vocabularies following the Linked Data principles.

2 GENERAL DESCRIPTION

Participants: MBFSZ, ISPRA, GBA, TNO, BGR, GeoZS, LfU It describes how to use project vocabularies suitable for semantic harmonization purposes. It evaluates Linked Data resources, the SKOS ontology, SKOS examples and best practices, define entity and relationship types, with restriction to a scientific use. Finally yet important, to find the differences between knowledge representations and standardized code lists like INSPIRE/GeoSciML.

GeoERA project vocabularies are collections of (linguistic labeled) scientific concepts, stored in named graphs and concept schemes according to different modeling approaches. GeoERA project vocabularies also could be understood as an initial part of a future EGDI[1] knowledge graph.
GeoERA project vocabularies are restricted to linguistic labeled concepts, described in SKOS/RDF plus metadata properties with focus on scientific reusability. For non-scientific vocabulary concepts, we recommend to use already published RDF data like Wikidata[2] or Geonames[3].
In order to extend a knowledge graph, GeoERA project vocabularies can be supplemented by RDF data files. Project data come along with project specific ontologies (probably GeoSciML compatible) uploaded and stored in a separate named graph. Project data support project related applications and functionality.

2.1 Content of Project Vocabularies

Project Vocabularies are collections of controlled dictionaries containing essential information about scientific concepts **created** and **used** by a project. The primary goal is to support a project and databases with linguistically labelled terms. Project Vocabularies provide stable and reusable links to concepts (units of thoughts) that can be referenced whenever unambiguity is important. Behind such links alternative names, translations, definitions synonyms and additional information about other related concepts are made available. In any situation when something must be unambiguously named a concept from a Project Vocabulary can be used. A Project Vocabulary can facilitate search and information access in a linked data environment, save and share knowledge gathered during a project. Project vocabularies are modeled only on the level of terms and words and thus do not compete with geological data objects stored in a SQL database (according to GeoSciML or INSPIRE). Self-contained, they may also contain smaller RDF Data Supplements (see 5.2.5) for better understanding, reflecting the project's view at the time of publication. An example of this would be an associative property of geological units with fault systems.

[1] European Geological Data Infrastructure (EGDI) http://www.europe-geology.eu/
[2] Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. It acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others. see https://www.wikidata.org
[3] GeoNames is a geographical database available and accessible through various web services, under a Creative Commons attribution license. see https://www.geonames.org/

The types of applicable vocabulary concepts are:

1. **Terms describing geoscientific feature types or properties (schema-level)**. An example here would be a concept "*Fault system*" describing a feature type, or "*kinematics*" describing a property. See another example below
2. **Terms named by classifications or prototype theory**[4]. An example for a widely agreed classification concept could be "*Sand*" classified by grain size. A scientific concept like "*Dachstein Limestone*" which was named by the first discovery and location at the Dachstein Mountain is an example for the prototype theory. Since the term is used to describe occurrences of this geologic unit in many other locations far from Dachstein too.
3. **Combination terms for geologic features (like map legend items)**. Combined terms often are used to describe groups of atomic (impartible) concepts. Which actually does not exclude that scientists later will find a single term to describe this kind of combination. A legend text like "*Sediments of the Molasse Zone and the intramontaneous basins*" would apply to this type.
4. **Located and named occurrences of geoscientific types or properties (instance-level)**. The concept "*Inntal Thrust*" as an occurrence of (or instance of) a feature type "*Fault*" applies to this concept type. See another example below

Example (ad 1): Two datasets (A and B) including the age of geological units are to be copied together into a harmonized database. Dataset A uses general and internationally standardized terms of ages such as *Neogene*[5]. Dataset B uses more detailed but regionally common terms, such as e.g. *Ottnangian*[6]. Geological units of a certain age are then to be queried from the common database. The simplest solution would be to agree on a common, but very general level, e.g. to encode *Ottnangian* from dataset B to *Neogene*. The term *Ottnangian* would thus only remain as a note in the dataset and would not be available for a joint request. Using project vocabularies we can describe the relation between *Neogene* and *Ottnangian* (via *Miocene* and *Burdigalian*) as a superordinate broader term. Semantically harmonized datasets, in addition to the very general standardized term *Neogene*, could also include an attribute *Ottnangian* - which would now allow for semantic queries with a certain degree of similarity.

Example (ad 4): "Fault system A" (displayed on one country geological maps) is published as a concept in a vocabulary - as is in another country "Fault zone B". Research results show that both structures can also be seen as a cross-border "Large Fault System C". In addition, the newly created structure C also defines the bounding of a tectonic superunit D and is linked to these concepts. Now, the newly created and independent "semantic" concepts (A, B, C, D) in any data publications (e.g. Excel, PDF, websites) by specifying the URIs (web addresses of the concepts) can be globally referenced and used. This new data is automatically cross-border harmonized by doing this conceptualization step - even if there are no updated maps or geodata yet.

[4] See https://en.wikipedia.org/wiki/Prototype_theory (the first stimulus to be associated with that category) which is very common in Geology e.g. when formations are named after a site location

[5] Example URI http://resource.geolba.ac.at/GeologicTimeScale/38

[6] Example URI http://resource.geolba.ac.at/GeologicTimeScale/188

2.2 Scope of application

In discussions and meetings during the first half of the GeoERA project, mainly **two use cases** were defined or specified. On the one hand, the use of semantic web technologies proved to be a viable solution for building up a collection of vocabulary concepts of geological features (e.g. "Fault systems" in HIKE project) - which should later enable differentiated **access to spatial data**. On the other hand, there is likely to be a need for future **extensions of INSPIRE code lists**. In this case, project vocabularies can provide the necessary information in advance in the appropriate SKOS technical format. The following describes the limitations on linguistic concepts and the use case in comparison to code lists and keywords.

2.2.1 Focus on (linguistically labeled) scientific concepts

GeoERA project vocabularies will extend or complete the harmonization efforts of INSPIRE and GeoSciML. The goal is to make two systems (SQL databases and Linked Data) working together and benefit from both systems. Big advantages from Linked Data (SKOS/RDF/triple stores) are e.g. easy reusable multiple-purpose data (website, application, PDF, etc.), multilingual, interdisciplinary, flexible data structures. Big advantages from SQL databases are e.g. a data model for geology (GeoSciML), variety of software, support (e.g. GIS).
GeoERA project vocabularies will handle only "linguistic labeled concepts" by using the SKOS ontology[7]. Scientific concepts in GeoERA project vocabularies are defined as a matter of "things you can talk about" (= semantic concept) with "proved existence" (= scientific citation). Beside "semantic relations" we can use metadata properties from well known "ontologies" to provide other information or links – preferably dcterms, foaf, dbpedia, etc. – because of their straightforwardly use without cascading object classes (even if they provide such).
To start defining a geological ontology in order to extend SKOS would run into complex editing and validating procedures similar to GeoSciML and INSPIRE. This part of modeling GeoSciML/INSPIRE is to be supported by GIP-P WP8 (BGS) with cookbooks, learning material and so on (prepared on https://github.com/GeoEra-GIP/WP8-Support). Reusing existing ontologies is recommended at first. A work is in progress for creating ontologies from standard domain models such as GeoSciML, HY_Features, GroundWaterML2, etc. Our final report D4.4 will come back on the publication of such ontologies on the OGC naming authority server.

Since scientific Concepts[8] are linguistic labelled concepts[9] GeoERA project vocabularies obligatory have SKOS preferred labels suitable to use in scientific text publications[10].
SKOS primarily is defined by so called “units of thought[11]”. In scientific context we need linguistic terms used in scientific documents. Pure data objects or entities like MappedFeatures rather have identifiers than such labels to use in language. But if they can be described by linguistic labels like Geologic formations, named groundwater basins, fault systems or even named wells or deposit sites – they would fit into a GeoERA project vocabulary. Further, scientific concepts can be meaningfully

[7] see https://www.w3.org/TR/skos-primer/
[8] see https://de.wikipedia.org/wiki/Konzept_(Kognitionswissenschaft) in German language, the term "Konzept" is used in a broader context - differently than in English where it is a synonym to a language term plus meaning
[9] See https://de.wikipedia.org/wiki/Begriff
[10] see https://en.wikipedia.org/wiki/Scientific_terminology
[11] Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval, Stella Dextre Clarke, Alan Gilchrist, Ron Davies and Leonard Will, Willpower Information. Available at http://www.willpowerinfo.co.uk/glossary.htm

classified as descriptive (directly observable e.g. xy formation), theoretical (e.g. geologic unit) or hypothetical (not in practice observable due to observational time frame, e.g. genesis) depending on the procedural knowledge structures (i.e., reasoning patterns) needed for concepts construction (Lawson et al 2000). Section 5.1.2 describes different kinds of scientific concepts in detail.

2.2.2 RDF data supplements

In principle project vocabularies can be extended by RDF Data supplements, used to exploit the efficiency of using the RDF technique in publishing geoscientific information (see also 5.2.5 Examples section). Please note that the RDF data supplements itself are not part of project vocabularies in proper sense. They are just delivered using the same "RDF envelope" together with the vocabulary.

2.2.3 Differences between code lists, keywords and Project Vocabularies

The knowledge about something is defined by the context provided. Since code lists exist in the context of a specific data model, they appear more likely as a set of attributes supporting standards than as independent concepts. By contrast, project vocabularies should become reusable collections of knowledge for various data models or texts. Here a simplified comparison of different storage of information, their central concepts and associated context in order to create, store and retrieve knowledge is given:

- Geologic map: legend item -> spatial geometries, explanation texts
- Scientific paper: scientific terms -> descriptions, bibliographic references
- Spatial dataset: spatial object, geologic feature -> data, code lists
- Semantic web: concept -> any other related web resources

Code lists manage mandatory standards (possibly versioned) and may be registered individually, depending on their validity and their responsible bodies. For Linked Data and project vocabularies, the registration of delivered, archived, and published RDF files seems sufficient and more important than registering all individual terms.

In managing geosciences data, the definition of a code list probably will be taken from INSPIRE data specification – which says: In the case of an attribute type with coded values, an enumeration or code list should be used. If the set of allowed values is fixed, an enumeration should be used. If the set of allowed values may be extended by user communities or without a major revision of the data specification, a code list should be used. (INSPIRE 2014 p.45)
Code lists or enumerations in the sense of INSPIRE should be widely accepted and already used in the community as a standard. But this view also leads to "spontaneous standardization processes", to produce de facto standards[12] - prescribed for encoding geoscientific data. In contrast, conceptual collections like project vocabularies just specialize in the representation of the knowledge that was developed during the project - or was reused from the previous project. By focusing on a scientific description of the used concepts, a project vocabulary also represents the date of research and needs to be accepted only within the project team and project time line. Additionally it doesn't really

[12] A de facto standard is a custom or convention that has achieved a dominant position by public acceptance see https://en.m.wikipedia.org/wiki/De_facto_standard

make sense to create a separate registered[13] "project code list" with ambiguous terminology in order to drop or change when the following project starts. See also the chapter -> Support creating or extending INSPIRE code lists (Annex).

[13] INSPIRE Code List Register: Dictionary managed as a register describing the value domains for selected properties in an application schema, but which is managed separately from the application schema in its own dictionary. I.e. this establishes an extendable controlled vocabulary outside of the INSPIRE data specifications (INSPIRE 2014 p.98)

		INSPIRE or CGI Code lists	GeoERA Keyword Thesaurus	GeoERA Project Vocabularies
scientific scope	**short description**	largely	optional	mandatory
	bibliographic citation	to some extent	no	mandatory
	status of publication	yes	no	yes
modeling	**data structure**	relational data (SQL), XML/RDF, JSON, Atom export	graph data (SKOS/RDF)	graph data (SKOS/RDF, dcterms, foaf, ..)
	principle of modeling	relational, mono-hierarchical	poly-relational wordnet (what else could be searched for..)	poly-hierarchies, partitive or generic-specific
	poly-hierarchies	no	yes	yes
	Linked Data incl. mappings	no	INSPIRE, CGI, GEMET, GBA, ..	at least INSPIRE, CGI, (GBA, WIKIDATA, ..)
	multilingual	partly	yes, > 10 languages	desired
	type of vocabulary	European or international standard	subject heading system	knowledge base, representation
data access	**data governance**	EC, JRC, community	GeoERA, EGDI	GeoERA, EGDI projects
	extendable	yes, officially	yes, by EGDI	self-contained and completed, but extendable by other vocabularies
	web API	no	Sparql endpoint	Sparql endpoint
	archive download	re3gistry tables	yes, RDF	yes, RDF
use case	**search metadata (e.g. catalog)**	no	yes	no
	merge datasets on attribute level	yes	no	semantically
	select feature data semantically	partly	no	yes
	prepare code list extensions	yes	no	yes
	publish scientific terminology	possibly	only on metadata level	yes

Table 1: Code lists, keywords and project vocabularies

This comparison in figure 1 shows the difference in scope of applications for different kinds of vocabularies.

3 METHODS, WORKFLOWS

Participants: SGU, GBA, IGME, MBFSZ, BGR, ISPRA, GeoZS, CGS, LfU, BRGM To develop methods, models and workflows around project vocabularies - which is about knowledge modelling, generation of vocabularies, connection to metadata catalogue, linked data showcases. This task applies the principles of lexical semantic modelling. It also designs a workflow of data transformation and -integration for project vocabularies and ideas of using RDF data format and O&M standard to describe geological and geophysical features.

An important criterion in the publication of project vocabularies is to state a mentoring of vocabularies by a responsible organization behind. That answers the question of who is responsible for publishing a vocabulary concept - which is necessary to know for linking concepts from one published project vocabulary to the next project. In addition, the name of a responsible organization and the project acronym has to be a tagged part in the URLs of vocabulary concepts (URIs, web addresses). GeoERA URIs must therefore contain the organization GeoERA (or EGDI) and a project name as an acronym following EC guidelines for URI design (Archer et al. 2012).
The creation of a conceptual knowledge undertaken during the lifespan of a project ends when a vocabulary deliverable is submitted. This is another important principle which means that a vocabulary once published should always represent this archieved status regarding the knowledge created by the project. A permanently and dynamically updated knowledge base of vocabularies would require an organisation body behind, like a permanent editorial board of experts. This is not part of GIP-P project in order to create a technical infrastructure for Linked Open Data and project vocabularies.

3.1 Methods

3.1.1 SKOS ontology, a simple modeling approach

The modeling of project vocabularies in SKOS[14] needs at least two strict guidelines to ensure some functionality for future applications. Within a single SKOS concept schema - on the one hand, it is not allowed to mix "abstraction relation[15]" with "partitive relation". And on the other hand, within one concept schema only vocabulary concepts (see 1.1.) of the same entity type are to be modeled. Different project teams can model their project vocabularies until the projects results has to be delivered. The teams furthermore should be accompanied by expert advice from the GIP-P WP4.

Examples:
To create narrower concepts for a geologic structure concept "Fault system" like "Fault" would apply to partition, because a fault system has faults like parts of the whole system. This relation type is also known as "has" relation.
An example for an "abstraction relation" could be the relation between a "Granite" and its broader concept "Igneous material", because a granite is a kind of igneous material. This relation type is known as "is-a" relation.

[14] See SKOS Simple Knowledge Organization System Primer - W3C Working Group Note, https://www.w3.org/TR/skos-primer/
[15] Also known as "generic-specific relation"

3.1.2 Design of vocabularies

During the implementation phase in 2020, project leaders must decide whether to publish their elaborated scientific terms in a common cross-project vocabulary or in their own project-specific vocabulary. E.g. overarching classification terms probably will be favorably designed within a common vocabulary while on the other hand regional terms better will be goverened by different project leaders within their own project-specific vocabularies.

SKOS facilitates establishing **semantic relations between pre-existing concepts**. Such mappings (close match, exact match, broad or narrow match) are crucial for applications such as information retrieval tools.

On the other side it is also possible to **include already published concepts** in order to reuse and extend[16] them within a GeoERA project vocabulary by using skos:inScheme. The use of URIs on the Semantic Web allows resources to be shared and reused in a distributed fashion. As a result it is possible for a SKOS concept to participate in several concept schemes at the same time. For example, a SKOS publisher can choose to locally extend an existing concept scheme by declaring any new concepts that may be needed and simply linking to concepts that have already been defined in the existing scheme [SKOS primer]. This can be especially useful when different GeoERA projects want to achieve a better coverage of their specific fields, while following the principles that guided the design of the existing vocabularies, by re-using some of its concepts.

Note: The use of "Linked Open Data" means far more than converting data into RDF format and to publish it on the web. Rather, knowledge should be generated and made accessible by global cross-linking of all resources (global knowledge graph). For GeoERA projects this implies an integrative modeling of new knowledge together with published knowledge (like e.g. INSPIRE codes, Wikidata, Geonames, etc.), which is already available under different web domains. In practical terms, a central registration of vocabularies actually can only refer proper GeoERA concepts, but not to integrated knowledge from the "Semantic Web". The responsibility to keep vocabularies online long-term is in principle the responsibility of the owner of the webdomain under which the URIs were published. Of course, for reasons of performance it is advantageous to copy these ”external” (parts of) vocabularies into the same triple store and Sparql endpoint.

3.1.3 Support creating or extending INSPIRE code lists

The procedure to extend the INSPIRE codes could follow 2 ways:

1. The first is a simple extension of the list for lack of information; this must be proposed in the Thematic Cluster forum (https://themes.jrc.ec.europa.eu/ managed for Earth Science by BGS) which collects the requests, evaluates if they are feasible and proposes them to the MIG-T that must approve or reject them or ask for more information.
2. The second way is to prepare a code list (better if it is a hierarchical vocabulary) that then is published through a National or Community Resources Register (EGDI) and proposed either by a nation or by EGS/EGDI as a federated resource of the INSPIRE register. This second must then have the screening and the recognition of the Registry Board composed by all the National Registers Members appointed by the countries as referents and by the Control Body team. See also section 5.2.1 inspire code list extension of this document.

[16] see Isaac A., Summers E. (2009): SKOS Simple Knowledge Organization System Primer - W3C Working Group Note, https://www.w3.org/TR/skos-primer/#secextension

3.2 URI design

Participants: GBA, GEUS, TNO, IGME, BRGM, LfU A common agreed URI design and a strategy to assign for keywords and project vocabularies.

All scientfic concepts published in a ”GeoERA project vocabularies” framework are considered to be accessible and online available in the Semantic Web in a sustainable and long-term storage. In general a validation of Linked Open Data (LOD) is suggested to be done by resolving URIs and checking if this information is still online. In that way ”GeoERA Project Vocabularies” are considered to be a part of the Semantic Web and the LOD cloud.
For scientific concepts, terms or names used by projects, very often there is a need to clarify cross border terminology. Project leaders may deliver these concepts to GIP-P - together with synonymous or related terms, multilingual translations, short descriptions (2 lines), source references (e.g. bibliographic citations or DOIs), references to other websites or online resources (e.g. images, web-services, data stores, Wikipedia, YouTube, etc.). GIP-P will create an ID (URI, a resolvable HTTP web address) for each single term to index project data. It will be possible to integrate all this information (e.g. multilingual translations) live in a web application, project portal or a simple webpage via a web service (Sparql endpoint). Of course, this online information is also reusable for all other future applications or projects.
The biggest challenge will be to ensure the sustainable and long-term publication of vocabularies. Terms once published together with their global identifiers (URIs) cannot longer be deleted and must be permanently available and resolvable.
Relatively close to the deadline of this report, the GIP-P project team agreed on a domain name - https://data.geosience.earth (/ncl/) - to be used for URIs. The structure concerning subdomains and paths will be discussed further and decided in GIP-P WP5 - as a "URI naming policy" or recommendation, in order to unify the way vocabs and concepts are named (camelCase writing, number of levels in in the URI path, etc.). This solution will be applied not only for GeoERA Project Vocabularies, but also for the GeoERA Keyword Thesaurus editions.

Note: Behind URI’s domain name (and top-level domain[17]) is a need of a responsible organization in order to ensure a long-term online availability of Project Vocabularies concepts. This organization also ensures an access to a details page (website) to get human readable information (browsing facilities for Linked Data included) and machine readable information by creating an RDF suffix to the URL. It is the central contact point which has to agree when a project creates new concept URIs too.

[17] Management and responsibility for TLDs (top-level domain) is delegated to organizations by the Internet Corporation for Assigned Names and Numbers (ICANN) and the Internet Assigned Numbers Authority (IANA), which maintains the DNS root zone.

3.3 Workflows

Within the GIP-P project, WP 4 has documented a workflow for development of GeoERA project vocabularies. The workflow can be separated into four process steps from initiation by individual project needs, to archiving and publishing of the project vocabularies (Figure 1).

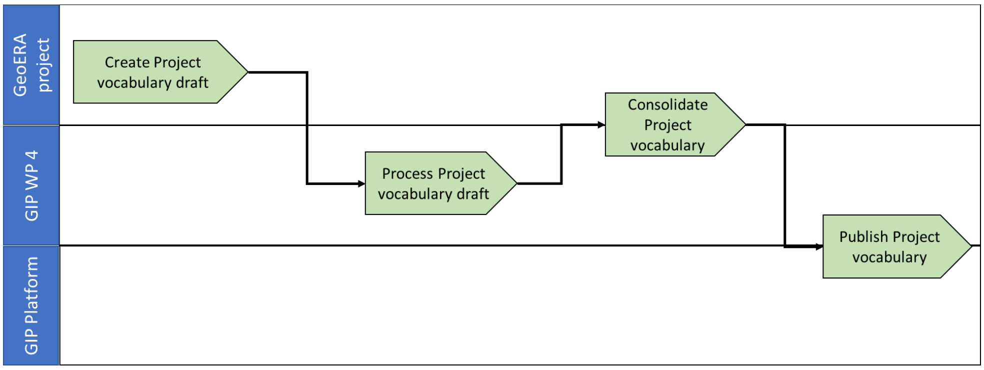


Figure 1. Four process steps to published Project vocabularies

Below is a brief description of the four process steps described in Figure 1, followed by more detailed descriptions in section 3.3.1.

1. Create Project vocabulary draft
   Based on each projects needs Project vocabulary drafts will be created.
   This step is done by the GeoERA Projects with the possibility of support from GIP.

2. Process Project vocabulary draft
   A mostly manual task done by WP 4 to structurally transform and visualize the Project vocabularies for further work.

3. Consolidate Project vocabulary
   This step is done by both the GeoERA projects and GIP WP 4. The aim is to finalize and harmonize Project vocabularies. The GeoERA project leaders will be considered "Publishers" of their respective Project vocabularies.

4. Publish Project vocabulary
   Storing/archiving the Project vocabularies as triplets (RDF) and making them available/accessible through standardized interfaces.

3.3.1 *Create Project vocabulary draft*

The first process step starts with initiation of identifying vocabulary needs within the GeoERA projects and establishing contact with GIP-P WP 4 regarding coming process steps within the workflow. The process step Create Project vocabulary draft can be divided into five sub process steps (Figure 2) that describes the process from initiation to delivery of a Project vocabulary draft. The work is supported by instructions and an MS Excel template provided by GIP-P WP 4.

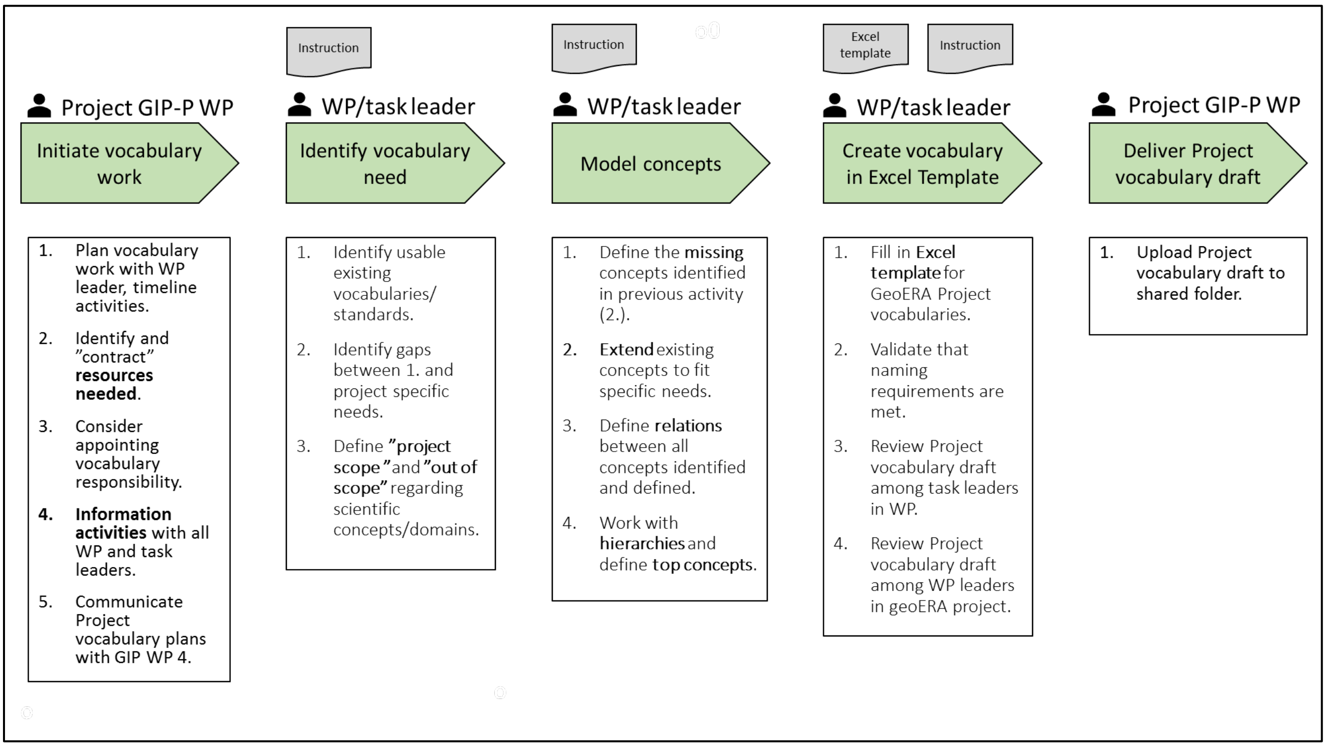


Figure 2. Create project vocabulary draft

3.3.2 *Process Project vocabulary draft*

The second process step (Figure 3) is completely done by GIP-P WP 4 and aims at transforming the structure from the MS Excel template to a triplet structure: When in the triplet structure, concepts will be analyzed and technically validated regarding mandatory elements and structure. A semantic validation will also be done with the aim to find "semantic conflicts" within the Project vocabulary but possible conflicts between different Project vocabularies must also be resolved. Before distributing the Project vocabularies back to the GeoERA projects as knowledge graphs URIs will be assigned to the concepts. The same URIs will also be used when publishing the Project vocabularies using the GBA website. URIs and published Project vocabularies using the GBA website are temporary during the work on GeoERA projects until a persistent solution is implemented.

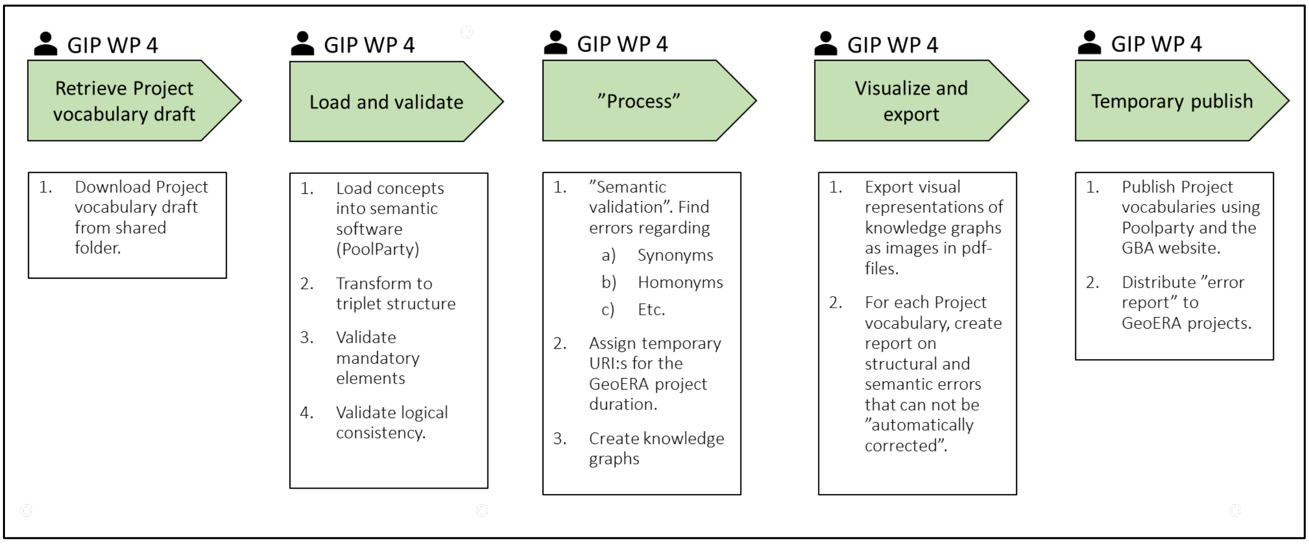


Figure 3. Process Project vocabulary draft.

3.3.3 Consolidate Project vocabulary

The third process step (Figure 4) is aimed at resolving issues within, and if needed between, Project vocabularies. The result of the process step is a confirmed Project vocabulary ready for “final” publishing. The work is mainly done by the GeoERA projects with support from GIP-P WP 4 when needed. Input is mainly from process step two, knowledge graphs and temporary published Project vocabularies. A basic instruction for communicating Project vocabularies is provided by GIP-P WP 4.

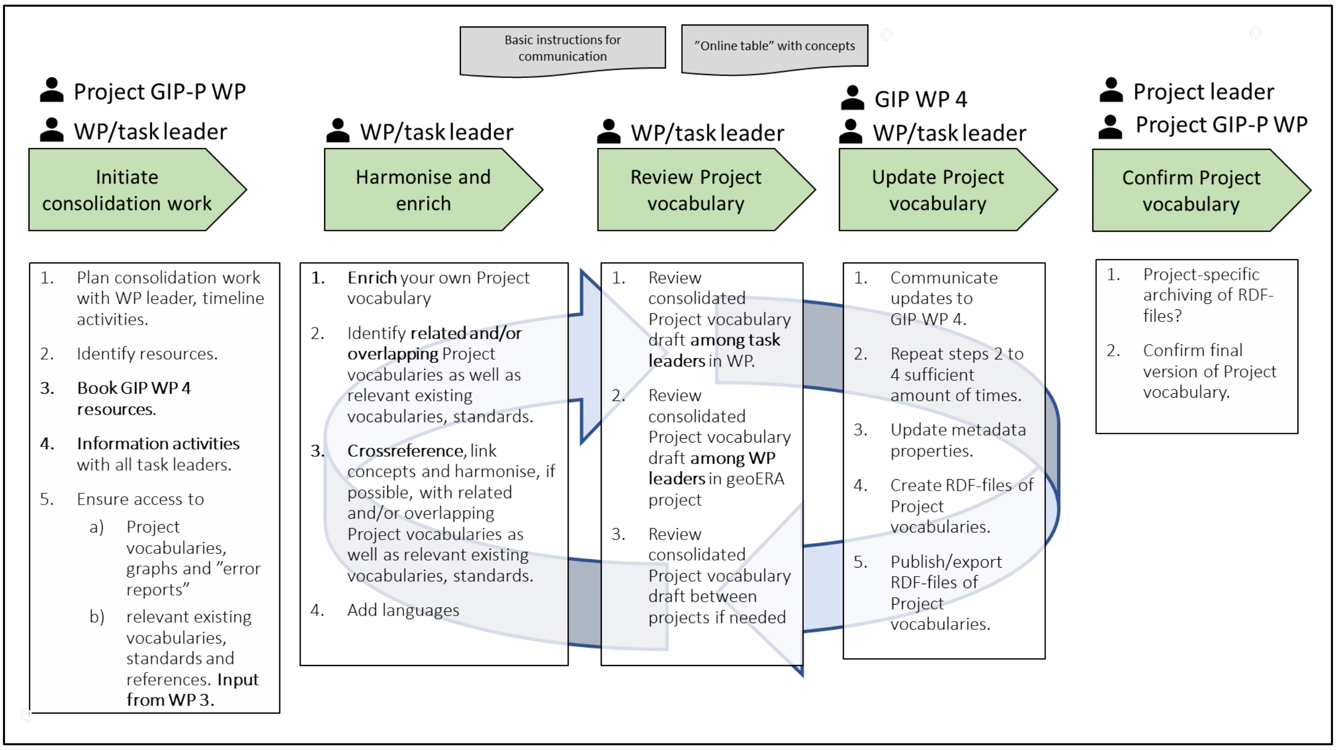


Figure 4. Consolidate Project vocabulary.

3.3.4 Publish Project vocabulary

Focus of the fourth and final process step (Figure 5) of the workflow is to assign persistent URIs to the Project vocabularies and their concepts followed by archiving and publishing the Project vocabularies. This process step still needs to be synchronized with the output of WP 3 - Standards and interoperability issues and WP 5 - Architecture of GIP-P. The sub-process steps will be relevant but more information, based on the technical solutions chosen in the GeoERA project and for future management of the GeoERA Project vocabularies, needs to be added.

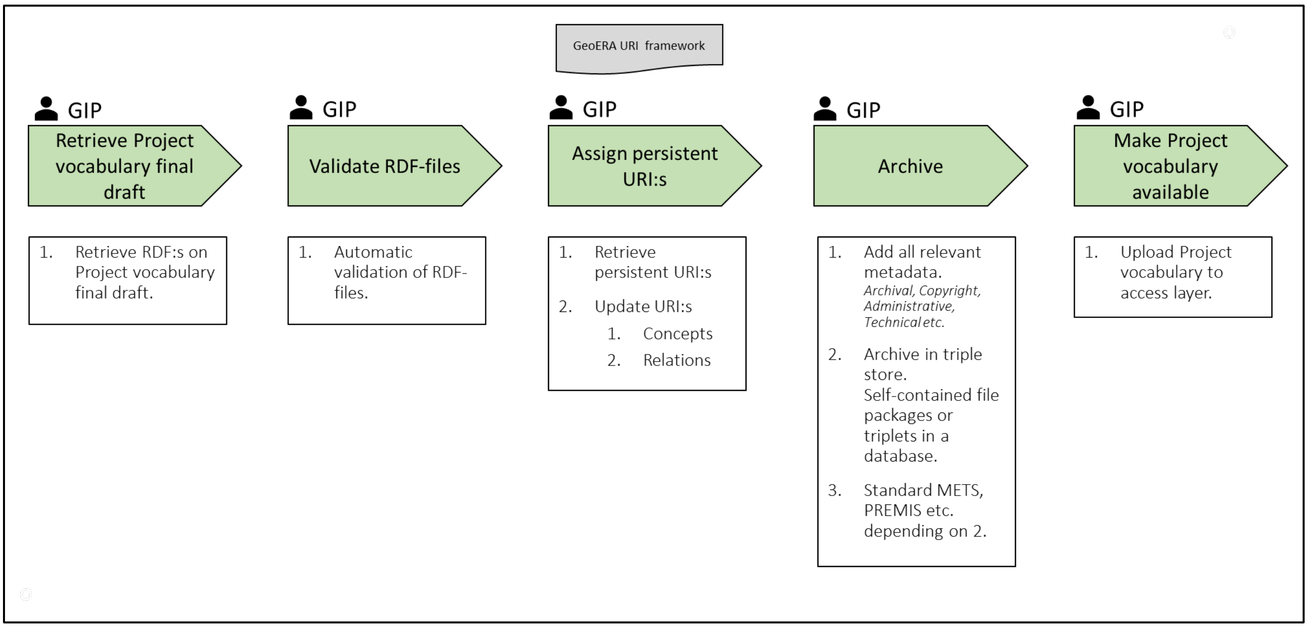


Figure 5. Publish Project vocabulary.

3.3.5 Open question – Bibliographic resources

In the case that *bibliographic resources* used as *bibliographic references* for Project vocabulary concepts are not published/available we might need an identifier for these resources.
A possible solution for this could be to create a separate concept scheme with only ”project vocabularies/concept schemes” as top concepts containing bibliographic resources instead of scientific concepts. A modification of the first process step (Create Project vocabulary draft) could then be used to describe that process step.
Note: this should only be used when referencing functionality does not exist elsewhere. Additionally, reference is made to the final report in 2021 for this issue.

3.4 Template and instruction

WP4 prepared a "quick start" manual for editing an Excel sheet in order to provide basic information for scientific concepts. This initial information is in general necessary to start a new GeoERA Project Vocabulary. This basic information will be supplemented later in the modeling phase, towards the end of the project, with further relations like mapping links, other links or properties. A detailed draft description is available under https://github.com/GeoEra-GIP/WP4-Semantics which further should be processed by WP8 cook books (see https://github.com/GeoEra-GIP/Project-Support-WP8).

Direct link to download the template: https://github.com/GeoEra-GIP/WP4-Semantics/raw/master/PV_template_v1.zip

Another helpful description how a Project Vocabulary is to be prepared in Excel – is publicly available under https://www.youtube.com/watch?v=-T5ccuvlU58 (provided by SemanticWebCompany)

4 PUBLISHING PROJECT VOCABULARIES - SUPPORT

Participants: TNO, GBA, LfU, ISPRA A support part is to advise and assist project teams in the building of project vocabularies and to interact with projects where the harmonization will be done. It is mainly about knowledge modelling, transformation, validation, and implementation of project vocabularies. The support team working on project vocabularies uses guidelines drafted under T4.2 and finalized under WP8 professional data provider support. In addition, a test of a project vocabulary use case (workflow, frontend, and implementation) is planned.

Please note that this chapter is just a preview to the phase of implementation in 2020, when GeoERA project vocabularies will be elaborated. A complete summary of delivered vocabularies will be part of our final report "experiences and status of the work (T4.1 and T4.2)" in month 34.

The projects in contact with regarding Project vocabularies are:

RESOURCE: Are currently not planning to create their own project vocabularies, but are using one INSPIRE code list: http://inspire.ec.europa.eu/codelist/LithologyValue
There is talk about this code list being too detailed, therefore they might want to define a higher-level vocabulary in the future. At this moment however, they are still on the path of using the INSPIRE code list. For other attributes that the RESOURCE project is going to include in their pan-European map, no use of vocabularies (project or external) is foreseen.

HIKE: Project vocabularies are being developed. Recently, a discussion was started amongst 4 surveys (GBA, LFU, GSB-RBINS, TNO) about the use of GeoERA-wide vocabularies vs country-vocabularies that can then be linked together. The possibility of having country-specific vocabularies seems to have come from a misunderstanding and is now regarded as adding too much complexity.
The decision about when data should be put into a database and when into a vocabulary is a decision that seems to be difficult to make: The project is currently struggling with this as the data model is being designed. There is a tendency to put things that are not expected to change into a vocabulary and things that are likely to change during the project (for example named Faults) into the database.

HotLime: 's knowledge base (HotLime deliverable D5.2.3) will be based upon interlinked project vocabs resulting in a full thesaurus of the geological and technical concepts of HotLime and attached reports and publications. To this end 4 principal project vocabs will be prepared:
The HotLime Lithostratigraphic Units Vocab describing all lithostratigraphic units modelled resp. mapped in Hotlime's 3D and 2D spatial outputs and their hierarchical superstructure. All entries of this vocab, where applicable, shall be hyperlinked to existing more extensive national/regional SKOS lithostratigraphic schemes (e.g. the GBA thesaurus) or under preparation (e.g. the upcoming SKOS version of LithoLex https://litholex.bgr.de) via the URI.
The Hotlime Tectonic Units Vocab will describe the principal tectonic units modelled resp. mapped in Hotlime's 3D and 2D spatial outputs. Certain top concepts of this vocab will be also integral part of the "Structural Framework" vocab of GeoConnetec3d, i.e the concepts covering areas of GeoConn's WP "Transposing the Case", and should be non-redundantly (inScheme related?) part of both projects' vocabs.
The HotLime Tectonic Boundary Objects Vocab will feature the principal faults of Hotlime's 3D and 2D spatial outputs. Its descriptions will address i.a. the boundaries of tectonic unit concepts

described in the Tectonic Units Vocab, thus many terms of the description need a hyperlink to those concepts of other vocabs to avoid the frequent redundancy of elucidation of the semantic content. The Tectonic Units Vocab is well advanced and certain top concepts (each covering one of HotLime's case study areas) are ready for modelling and testing. This specifically as to include the various ways of interlinking concepts which are part of different vocabs of different GeoERA project. For instance, many concepts of the top concept "Tectonic Boundary Objects in the Central Molasse Basin Carbonate Reservoir" will be part of the European Fault Database and thus must be integral part of the HIKE vocab, and, likewise will be one of GeoConn's case studies, thus must be integral part of the GeoConnetec3d vocab.
The HotLime Geothermal Base Assessment Vocab will be kind of a glossary geared towards explaining all technical terms and physical values used in geothermal base assessment or otherwise helpful for the understanding of geothermal and related issues. Meaningfully, this vocab should be prepared jointly with other GeoERA projects dealing with similar topics and terminologies, e.g. MUSE.
Separately from the project vocabularies, a docs repository is needed to disseminate all HotLime reports, papers and other text documents or PDF sketch maps.

MUSE: The project had done some work in the direction of project vocabularies already but was not aware of the project vocabulary templates. The templates have been sent, and are being looked into. There has been discussions on how to proceed as apparently for many terms no links to references are currently avaliable.

EUROLITHOS: Discussions are ongoing and seem to converge towards starting from http://inspire.ec.europa.eu/codelist/CommodityCodeValue/dimensionStone and extend this with other "commodity types of dimension stones" like serpentinite – in order to elaborate a new project vocabulary for EuroLithos.

FRAME: Based on the CGI Commodity Code vocabulary (http://resource.geosciml.org/classifierscheme/cgi/2016.01/commodity-code), the project has made a start to fill out the Project Vocabulary template.

5 ANNEX

5.1 Background

5.1.1 Semantic web and Linked Data

By M.Schiegl

The objectives of the GeoERA project are fundamentally defined by the open handling of data. In addition, there is a variety of European, as well as national initiatives and guidelines, such as Open (Government) Data, Open Access, Open Science strategies, etc. Thus the project proposal for the "GeoERA Information Platform project – GIP-P" says "By creating an information platform that aligns and integrates with wider e-infrastructures across Europe and beyond (..) we will open up data from the European geological surveys to be integrated with a wider range of earth science data." Because of that the GIP-P project team has adopted the goal of integrating "Linked Data" and the principles of "Semantic Web" as a state-of-the-art technology in the data processing of GeoERA project results. The basics for this open and above all machine-readable access to data was drafted and initiated more than 10 years ago by Tim Berners-Lee - the "Inventor of the Web". This principle enables a better data access and data processing in the future for modern technologies like Artificial Intelligence. He as the initiator of Linked Data, suggested a 5 star deployment scheme for Open Data:

1. make your stuff available on the Web (whatever format) under an open license
2. make it available as structured data (e.g. Excel instead of an image scan of a table)
3. use non-proprietary formats (e.g. CSV instead of Excel)
4. use URIs to denote things, so that people can point at your stuff
5. link your data to other data to provide context (Berners-Lee 2006).

The 5-star system is cumulative. Each additional star requires that your data meet the criteria for the previous steps. Also the World Wide Web Consortium (W3C) defines standards for the Web (e.g. RDF or Sparql), including an open data model and various formats for this model (Wood et al. 2014). Linked Data is a set of techniques for publishing data on the Web using standard formats (like OWL, RDF, SKOS...) and interfaces (like Sparql endpoints). Data that comply with these techniques are called Linked Data. Another important constraint for Linked Data is that resources actually provide connections to other resources already published in the Web. This means that it is not enough to publish your own data just by using certain data formats or interfaces. This data must also link to data previously published in the Web. Only by cross-linking the data, "knowledge" is generated in the meaning of the Semantic Web. Traditional data management techniques have resulted in most of our data being split into silos or data lakes that cannot easily be recombined. Then we need to write applications to find, access, transform, and combine data from silos before we can proceed with a particular task. Linked Data makes this work much easier as it is easy to combine linked data from multiple sources (Wood et al. 2014).

Quite similar **W3C** defines 4 **principles of Linked Data** [18](2016): The term Linked Data refers to a set of best practices for publishing structured data on the Web.

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information.
4. Include links to other URIs so that they can discover more things.

[18] See https://www.w3.org/wiki/LinkedData

The idea behind these principles is on the one hand side, to use standards for the representation and the access to data on the Web. On the other hand, the principles propagate to set hyperlinks between data from different sources. These hyperlinks connect all Linked Data into a single global data graph, similar as the hyperlinks on the classic Web connect all HTML documents into a single global information space. Thus, LinkedData is to spreadsheets and databases what the Web of hypertext documents is to word processor files.

"Linked Open Data" focusses primarily on the processing of semi-structured or losely structured data especially connected to vocabularies, which are difficult to model in relational structures of GIS databases. In contrast to "simply" Linked Data, Linked Open Data also uses an "Open License". The so-called "Linked Open Data Cloud" (LOD Cloud) refers to freely available data sets published according to the principles of the Semantic Webs with cross-links to other data sets. (Auer et al. 2014).

The benefits of using Linked Open Data in publishing bibliographically referenced data are particularly important for the traceability of scientific research results. Here in the GIP-P project, the provision of a technical infrastructure for so-called "Project Vocabularies" should remedy the situation.

5.1.1.1 Change in data management

The use of vocabularies as collections of concepts enables a new method of managing geoscientific data. The possibilities to access to data via online-published vocabularies and concepts differs from access to data via map geometries or standardized code lists. Vocabularies published in the Semantic Web can bridge differently structured databases (e.g. of various European countries) and prepare them for a single-entry data access through web applications. This approach, which is relatively new in the management of geoscientific data, refers to a semantic alignment of relational database systems. Previous endeavors, such as the use of a common data model (INSPIRE, GeoSciML) or common code lists, represent the most important contribution to semantic harmonization. From a technical point of view, this concerns the use of common standardized web interfaces (e.g. WMS, WFS), referred to as technical interoperability. By means of an alignment of databases in terms of technology, then the use of common standards and code lists, and finally the use of project vocabularies, most of previously hidden knowledge now can be processed.

Thus, vocabularies if published by Linked Open Data can be used to connect datasets of different projects. A project result dataset or map like a compilation or interpretation of different source datasets is automatically cross-linked if a common project vocabulary was used to attribute data. This approach meets the standards of good scientific practice like the comprehensible handling of ideas, texts, data and other sources. It goes in the same direction like harmonization efforts for INSPIRE or GeoSciML by using common code lists.

5.1.1.2 Knowledge graph, another kind of knowledge base

Graphs treat the connections between information as equally important, a first-class citizen in the world of data. A knowledge graph not only gives you the data but also provides how the computer is to interpret and understand that data (McGinnis 2018).

A knowledge graph is a network (or significant part) of nodes and edges where relationships (edges) come first so that the links between data add huge value, as well as flexibility. It's semantic or self-

descriptive and has a natural language-like representation, making it easy to query and explore. It's smart too. Inference can uncover hidden knowledge that is too complex for human cognition, and being a graph means you can apply various graph-computing techniques and algorithms. Finally it's alive. You can use a knowledge graph to store information in a form that is easy to reuse (after Stichbury 2017).
In terms of data management: A knowledge graph is an interconnected set of information that is able to meaningfully bridge enterprise metadata silos - flexible, evolvable, semantic and intelligent.
In terms of Linked Data and RDF: Graphs (plus other standards) support descriptive, meaningful data. They provide globally unique identifiers. Using URIs as identifiers ensures that enterprise assets are truly unique and addressable. In addition, RDF graphs are connectable (Coyne 2018).
The set-up of GeoERA project vocabularies is to start building a geoscience knowledge graph, which is extended by each new projects knowledge.

5.1.2 Scientific concepts

By L. Sőrés

5.1.2.1 Abstract and Specific Concepts

A named concept can be both abstract or specific. An abstract concept is the archetype of all instances of the same kind of thing. For example, behind the term "tree" there is a conceptual model of all trees that grow in the forests. "Frodo's Mallorn" is a specific concept identifying a single instance from all trees. While RDF triple stores are natural places for abstract concepts, the multiplicity of specific concepts are much higher, so they are often managed in RDBMS. However, dictionaries may hold both abstract and specific concepts. As a matter of convention, GeoERA Project Vocabularies are mainly limited to abstract concepts, and should not be overloaded by large amount of entity or attribute data that is safely stored in RDBMS. Note, as all concepts need contextual modelling and scientific references. For example, a Project Vocabulary may contain the term "geothermal well", but it may not contain the names (identifiers) of all geothermal wells used in a project. At the same time, *located and named occurrences of geoscientific features* can be included. For example, the name "Inntal Thrust" is a commonly used term making it a good candidate for a Project Vocabulary concept.
In some cases, it can be useful to have borehole identifiers (specific concepts) in a triple store. Linking together tectonic lines and intersecting boreholes is a straightforward solution in linked data environment. If projects can provide such information, RDF Data supplements can be accompanied with the Project Vocabulary.

5.1.2.2 Simple and Complex Concepts

Concepts can be simple or complex. A simple concept refers to a phenomenon that is understandable in itself. *Terms describing geoscientific properties, or named by classifications,* such as age, lithology, environment are simple concepts. A complex concept refers to a phenomenon that can only be described with a combination of other concepts. *Terms describing geoscientific feature types,* such as fault, geologic unit, event are complex concepts. In traditional data models simple concepts are usually modelled as attributes that take values from code lists. Complex concepts can be modelled as classes or entities that are grouping a set of attributes.
What is considered 'simple' is also a matter of granularity. Going deeper in understanding a simple phenomenon often turns out to be the result of more elementary phenomena. A simple concept may become a complex one. A typical way of handling granularity of information is using *Combination*

terms for geologic features in map legend items. For example, a label saying "Sediments of the Molasse Zone and the intra mountainous basins" refers to a polygon of sediments on a map, suggesting that in higher resolution it is divided to Molasse Zone and intra mountainous basin sediments.

5.1.2.3 Scientific concept

A scientific concept represents a real world phenomenon in the context of a scientific domain. Geological concepts are scientific concepts in the geology domain.

Abstract scientific concepts:

- Terms describing geoscientific feature types
- Terms describing geoscientific properties
- Terms named by classifications
- Terms named by prototype theories
- Combination terms for geologic features

Specific scientific concepts:

- Located and named occurrences of geoscientific features

5.1.2.4 Knowledge

Knowledge is based on complex concepts that are associated to each other in many different, but well-defined ways. In a Project Vocabulary, knowledge is represented as a cross-linked set of controlled dictionaries encoded using the SKOS standard, which integrates projects knowledge in the Semantic Web (global knowledge graph). The type of concepts and their possible relations can be strictly defined by using ontologies. Ontologies are dictionaries of conceptual elements the can be used to build a knowledge base. For technical reasons in GeoEra Project Vocabularies only the most common ontologies (dcterms, foaf, dbpedia, etc.) are allowed to create associations. Currently lots of efforts are taken in setting up geoscientific ontologies based on OGC and INSPIRE standards. As domain specific ontologies become consolidated and safely available, Project Vocabularies may take more advantage of them. Meanwhile it is possible to use RDF Data supplements to extend the GeoERA knowledge base with domain specific information.

5.1.2.5 Topics

Scientific concepts related to different disciplines like geology, hydrogeology, mining, geophysics can fit in one Project Vocabulary. It is also possible that a Project Vocabulary contains only extensions to already existing dictionaries. In general, a Project Vocabulary main contain the following types of information:

- dictionaries of simple scientific concepts (age, lithology, environment, commodity)
- dictionaries of complex scientific concepts (geologic units, faults)
- dictionaries of abstract scientific concepts
- keyword dictionaries

Analyzing the foreseeable GeoERA products and functionalities listed in the D2.2 draft report from WP2, the following topics have been identified as possible Project Vocabulary content:

- catalogue of stone types (EuroLithos)
- commodity names (Mintell4EU, EuroLithos)
- fault names (HIKE, 3DGEO-EU)

- geological age (HIKE, 3DGEO-EU)
- lithology (HIKE, 3DGEO-EU)
- observed properties (all projects)
- observation process names (all projects)
- process parameter names (all projects)
- key concepts in descriptive scientific material (all projects)
- citations (all projects)
- links to photos (EuroLithos)
- names of (2D/3D) geological model components (HIKE, Geoconnect³d)
- names of elements building up workflows (VoGERA)

Content of RDF Data supplements:

- dictionaries of specific scientific concepts (entities in RDBMS)
- seismic dataset names and connections to model components (Geoconnect³d)
- borehole names and connections to model components (Geoconnect³d)

5.1.3 Controlled vocabularies

By G. Diepolder

Standardization of terminologies is an inevitable prerequisite for sound interoperability. In geosciences, due to the natural variability of the subject matter in space and time, many nomenclatures of factual scope have been set up. These standards with a limited areal validity evolved from regional approaches and only rarely have undergone cross-border harmonization. Descriptive texts are often used to caption those terminologies whose meanings are not standardized or are not conclusively clarified in the international context. This leads to distortion and ambiguousness when cross border datasets are compiled, not only caused by national languages but also due to regional peculiarities and the semantic changes in historical evolution of terms. Standardization of geological interpretation terminologies, however, is virtually impossible to gain as pluralism of terms is fact-based, well-established and has been used in geoscientific publications over decades. Furthermore, especially in geology, semantics also define the delimitation of units (e.g. depth of strata), thus, equalization of terms such as the lithostratigraphic subdivisions (e.g. geological formations) would require an extensive realignment/re-mapping of geological units.
State-of-the-art information technology using knowledge representations based on Semantic Web principles can make up for this deficiency of standardization: A Linked Data approach by cross-linking and interrelating globally and uniquely (URI) referenced terms (technically referred to as concepts) and glossaries (overriding concepts) enables to identify congruence, similarity or disparity of concepts and thus the generation of multi-lingual Controlled Vocabularies. This modern technology and approach ideally complements the mainstream GIS and relational database technologies in making data findable and interoperable.

5.2 Examples

Once vocabularies have been published on the Web, they can be used through hyperlinks in web applications, online portals, and websites or even in local documents or PDFs. An attractive option of the application will be the semantic extension of online queries. Thus, e.g. a click on an online map or web service not only provides information about the selected feature, but also matching context information to get the next feature.
This procedure allows for navigation between datasets or maps through query results. The visibility of datasets hosted in relational database systems generated by view services is thereby increased. GIP-P will provide a web frontend (website on portal) in order to search or browse all scientific concepts collected and published in GeoERA project vocabularies.

5.2.1 INSPIRE code list extension

An example for using the INSPIRE code http://inspire.ec.europa.eu/codelist/LithologyValue/sand within a GeoERA project vocabulary would be:

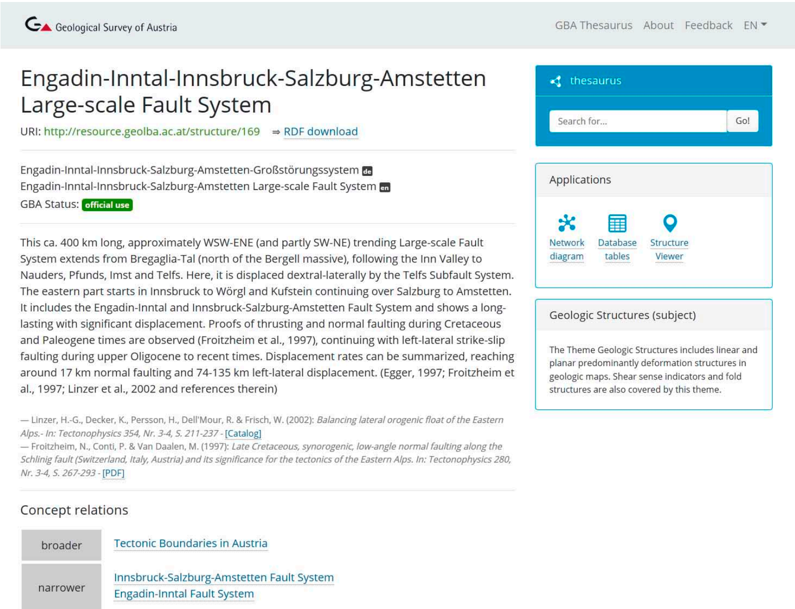
```
inspire: http://inspire.ec.europa.eu/codelist/LithologyValue/
project: https://data.geoscience.earth/ncl/project/

inspire: rdf:type skos:ConceptScheme;
   dct:title "Lithology values"@en.
inspire:sand rdf:type skos:Concept;
   skos:prefLabel "sand"@en;
   skos:inScheme inspire:.
```

The creator of projects concept scheme can freely include the reference concept (`inspire:sand`) in the new scheme (`project:projectScheme`), and then reference it as follows:

```
project:projectScheme rdf:type skos:ConceptScheme;
   dct:title "Sand vocabulary"@en.

inspire:sand skos:inScheme project:projectScheme.

project:fineSand rdf:type skos:Concept;
   skos:prefLabel "fine sand"@en;
   skos:broader inspire:sand;
   skos:inScheme project:projectScheme.

project:coarseSand rdf:type skos:Concept;
   skos:prefLabel "Coarse sand"@en;
   skos:broader inspire:sand;
   skos:inScheme project:projectScheme.
```

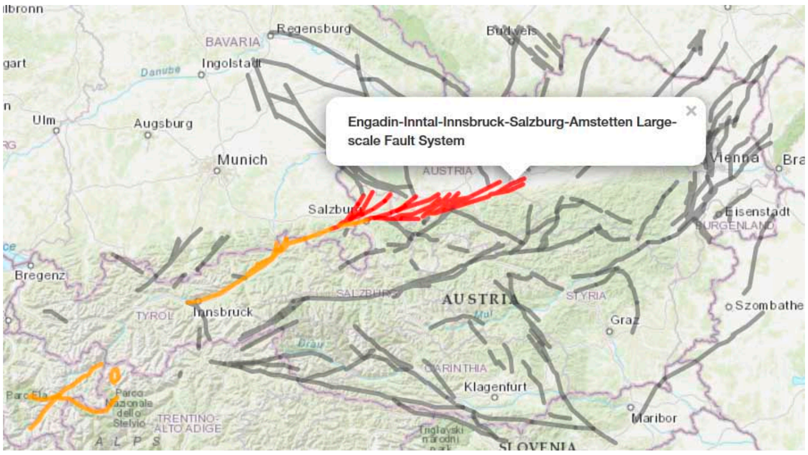
5.2.2 *Tectonic Boundaries in Austria*

"Tectonic Boundaries in Austria" is an example of a data management driven by vocabularies and Linked Data (see 1.1.2.) – prepared for the GeoERA project HIKE. In this case the GBA Thesaurus provides a web frontend to select a scientific concept[19], which is used to display related mapped features on the screen. To shape this functionality all imaginable semantic combinations on the vocabulary side are available to arrange such a map on the screen.



HTML representation of a scientific (vocabulary) concept of a geologic fault system, including different languages, short description, citation, bibliographic references and semantic relations.

Figure 6. Linked Data concept details page

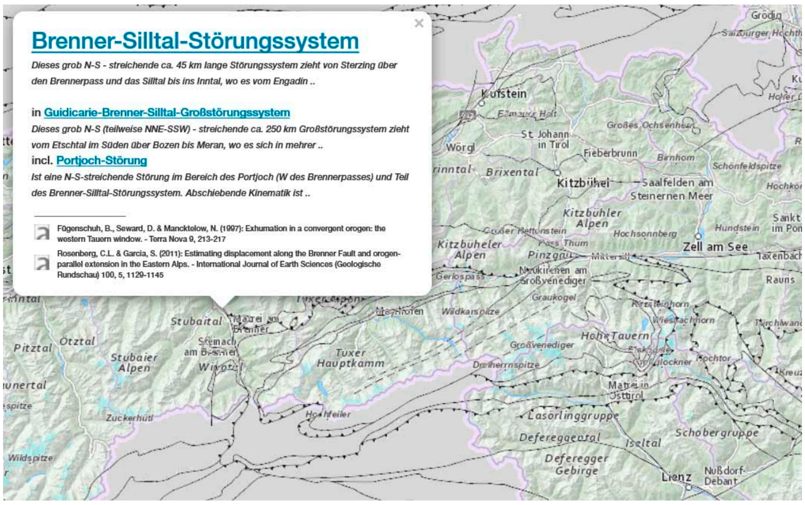


Web map displaying a geologic large fault system made up of geologic faults as part of the whole system (narrower concepts).

Figure 7. Web map displaying a "Large Fault System"

[19] See http://resource.geolba.ac.at/structure/169

5.2.3 WMS identify



WMS identify[20] pop up window generated by querying a SPARQL endpoint.

Figure 8. Web map (WMS) with Linked Data popup

5.2.4 Query a Sparql endpoint (using JavaScript, WHATWG[21], W3C, ES6[22])

A JavaScript “fetch” code example plus SPARQL query (in red) to get all narrower concepts:

```
let uri = 'http://resource.geolba.ac.at/mineral/284'; //example concept
let endpoint = 'https://resource.geolba.ac.at/PoolParty/sparql/mineral'; //example endpoint
//sparql query all narrower concepts with preferred labels in english
let query = encodeURIComponent(`PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
                                SELECT *
                                WHERE {
                                <${uri}> skos:narrower* ?s .
                                ?s skos:prefLabel ?Label . FILTER(lang(?Label)='en')
                                }`);
//JS for all modern browsers
fetch(`${endpoint}?query=${query}&format=application/json`)
    .then(res => res.json())
    .then(jsonData => {
        jsonData.results.bindings.forEach((i) => {
            console.log(i);
        });
    });
```

[20] See http://www.ce-gic.org/wms/GBA_structures.html

[21] Web Hypertext Application Technology Working Group (WHATWG, https://whatwg.org/)

[22] ECMAScript 6 is also known as ES6 and ECMAScript 2015 or maybe JavaScript 6 (http://es6-features.org/)

5.2.5 RDF Data supplements

By L. Sőrés

RDF Data supplements can be used to exploit the efficiency of using the RDF technique in publishing geoscientific information. The RDF elements are representations of the corresponding GeoSciML/INSPIRE features, ignoring deep database content. Links between them also try following the associations defined in the standards. In the lack of consolidated ontologies common RDF properties are used.

Nodes in RDF supplements should contain the least possible information about the corresponding database objects. The main reason is to capture the connections between the nodes and make them available for semantic search. Search results should come up with useful links to images, documents or WFS service calls that implement INSPIRE downloads. RDF data supplements are optional and are not part of Project Vocabularies.

The use cases below are simplified and presented for demonstration purposes.

5.2.5.1 Use case 1 - 3D Geological Model

A 3D Geological model produced in Geoconnected3D contains geological surfaces and a faults. Faults are also part of the HIKE dataset. The model has been created by using seismic and borehole data published through INSPIRE. The surfaces are stored in the central 3D model repository. Individual components of the model are 3D representations of named faults and geological contacts. These are also available in the project vocabulary as scientific concepts, and can be queried at the SPARQL endpoint. RDF descriptions also contain URLs that allow the user to view the surfaces or download them in standard formats. These model components may be used independently from the 3D model, so the knowledge can be shared with other projects. The RDF triple store contains the following items:

3DModel-1 represents the 3D model as a GeologicalCollection. It has title, description and references to the mapped geological features (volumes, and surfaces) that are members of the model. It may also have a link to the ISO metadata record of the same 3D model in the MICKA catalogue.

CNT-1 represents a mapped Contact with 3D surface geometry. Geometry is available in the central repository and can be accessed through srf_CNT-1. The relation to namedContact expresses that this 3D surface is an occurrence of the Contact geoc3d:namedContact.

namedContact: a known geological contact (eg. “top of Dachstein”), published in the Geoconnected3D Project Vocabulary. The object may also have type of Contact in a GeoSciML Ontology.

srf_CNT-1: Pointer to a downloadable instance of CNT-1 in some standard format, e.g. a gml GridSurface.

FLT-1 represents a mapped Fault with 3D surface geometry. Geometry is available in the central repository and can be accessed through srf_FLT-1. The relation to namedFault expresses that this 3D surface is an occurrence of the ShearDisplacementStructure hike:namedFault.

namedFault: a known Fault (eg. "Danube Fault"), published in the HIKE Project Vocabulary. The object may also have type of ShearDisplacementStructure in a GeoSciML Ontology.

srf_FLT-1: Pointer to a downloadable instance of FLT-1 in some standard format, e.g. a gml GridSurface.

SLN-1: identifies a seismic profile that was used to construct the 3D model components. This may be published in INSPIRE and available through a WFS service defined in DepthSection-1. SLN-1 is a SpatialSamplingFeature with several relatedObservations. DepthSection-1 is the result of one of them.

DepthSection-1: Pointer to a downloadable seismic profile. It can be an image, or some industry standard file.

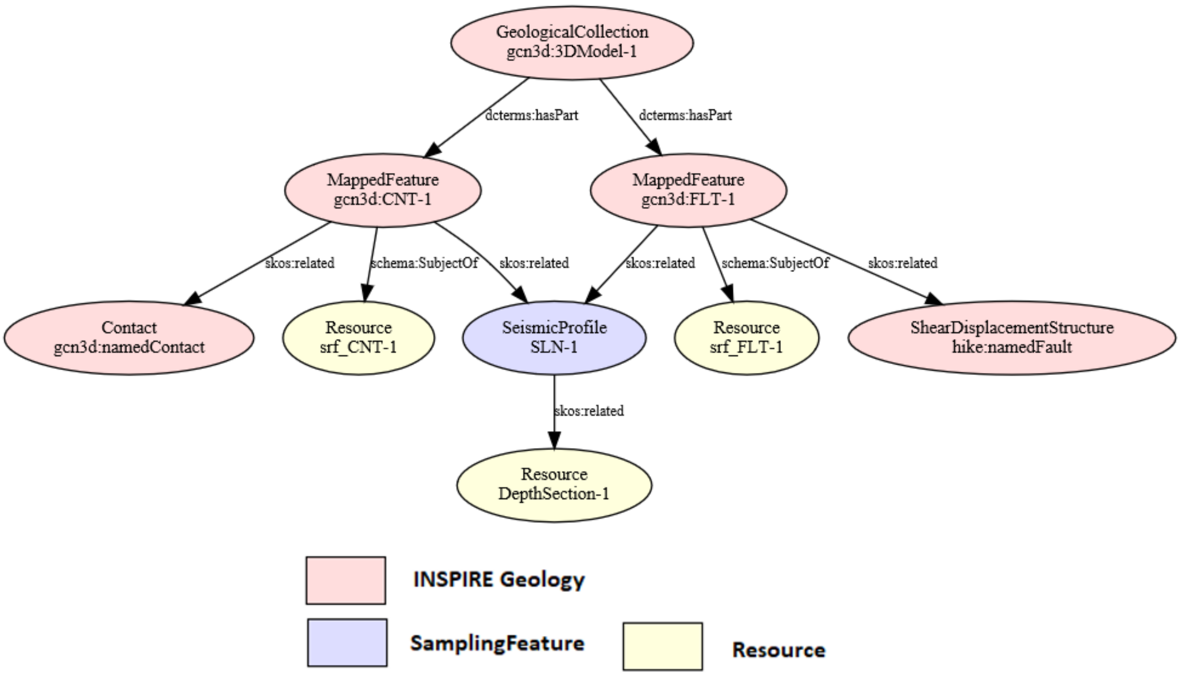


Figure 9. RDF Graph of the 3D Geological Model use case

5.2.5.2 Use case 2 - Particle size analysis of rock samples

In a hydrogeological project, rock samples from boreholes are analyzed to get particle size distribution of a named geological unit. The RDF triple store contains the following items:

BH-1: represents the Borehole from which the rock sample was taken.

PaksLoessFormation: A GeologicalUnit that belongs to the MappedInterval of BH-1 from where the rock sample originates. The scientific concept may be published in a project vocabulary.

SPC-1: In GeoSciML specimen is subclassed from SpatialSamplingFeature. The relation to BH-1 expresses a relatedSamplingFeature association. SPC-1 has references to the hosting geological unit and the observation resulting the particle size distribution. A photograph of the specimen may be found behind the schema:SubjectOf link.

particleSizeAnalysis: Observation that was carried out to determine particle size distribution. The result is available in histogram-1.

histogram-1: Pointer to a document containing the particle size distribution results.

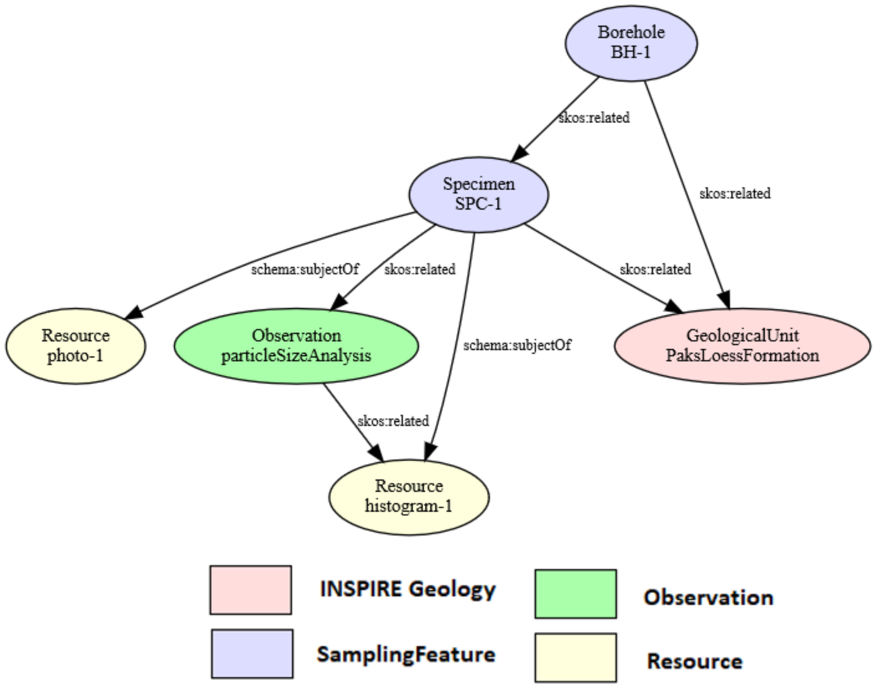


Figure 10. RDF Graph of the rock sample use case

5.2.6 GeoERA Project Vocabularies, delivered in RDF

5.2.6.1 Example – modeling with mappings (e.g. HIKE – Faultsystems)

incl short description, bibliographic citation and INSPIRE mapping.. formatted as RDF/trig

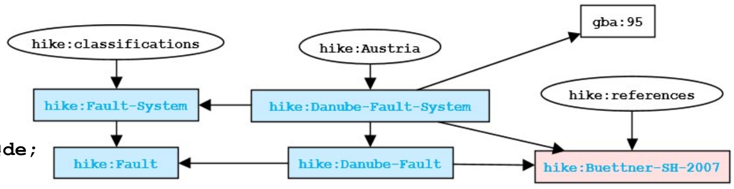


Figure 11. SKOS mapping Example

```
@prefix hike: <https://data.geoscience.earth/ncl/geoera_hike/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix gba: <http://resource.geolba.ac.at/tectonicunit/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <https://schema.org/> .

hike:classifications a skos:ConceptScheme;
dcterms:title "classifications"@en;
skos:hasTopConcept hike:Fault-System.

hike:Fault-System a skos:Concept;
skos:prefLabel "Fault System"@en,"Störungssystem"@de;
skos:topConceptOf hike:classifications;
skos:narrower hike:Fault.

hike:Fault a skos:Concept;
skos:prefLabel "Fault"@en,"Störung"@de;
skos:definition "The term fault is used .."@en;
skos:broader hike:Fault-System;
skos:exactMatch <http://inspire.ec.europa.eu/codelist/FaultTypeValue/fault>;
skos:closeMatch <http://de.dbpedia.org/resource/Störung_(Geologie)>;
foaf:depiction <http://www.geologie.ac.at/images/thesaurus/Stoerung.jpg>.

hike:Austria a skos:ConceptScheme;
dcterms:title "Austria"@en;
skos:hasTopConcept hike:Danube-Fault-System.

hike:Danube-Fault-System a skos:Concept,geo:Feature;
skos:prefLabel "Danube Fault System"@en,"Donau-Störungssystem"@de;
skos:definition "This NW-SE striking, steep to NNE dipping fault system runs from
Regensburg.."@en,"Dieses NW-SE streichende, steil nach NNE einfallende Störungssystem.."@de;
skos:notation "AT-01","BY-01";
skos:topConceptOf hike:Austria;
skos:narrower hike:Danube-Fault;
dcterms:references hike:Buettner-SH-2007;
geo:sfTouches gba:95;
schema:mainEntityOfPage <../structureViewer.html?uri=http://resource.geolba.ac.at/structure/128>.

gba:95 a skos:Concept,geo:Feature;
skos:inScheme gba:2;
skos:prefLabel "Moldanubian Superunit"@en,"Moldanubikum"@de .

hike:Danube-Fault a skos:Concept,geo:Feature;
skos:prefLabel "Danube Fault"@en,"Donau-Störung"@de;
skos:definition "Ist eine spröde NW-SE-streichende Störung.."@de,"Is a NW-SE trending fault and a
part of the Donau.."@en;
skos:notation "AT-02";
skos:broader hike:Danube-Fault-System;
dcterms:references hike:Buettner-SH-2007.

hike:references a skos:ConceptScheme;
dcterms:title "references"@en;
skos:hasTopConcept hike:Buettner-SH-2007.

hike:Buettner-SH-2007 a skos:Concept,dcterms:BibliographicResource;
skos:prefLabel "Buettner, S.H. (2007)"@en;
skos:topConceptOf hike:references;
dcterms:bibliographicCitation "Büttner, S.H. (2007): Late Variscan stress-field rotation initiating
escape tectonics in the south-western Bohemian Massif: A far field response to late-orogenic
extension.- In: Journal of Geosciences 52, Nr. 1-2, S. 29-43";
dcterms:source <http://www.jgeosci.org/content/jgeosci.004_2007_1-2_buettner.pdf> .
```

5.2.6.2 Example – inScheme modeling (e.g. EuroLithos)

incl short description, bibliographic citation and INSPIRE inScheme modeling,.. formatted as RDF/trig


```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix dbpo: <http://dbpedia.org/ontology/> .
@prefix inspire: <http://inspire.ec.europa.eu/codelist/CommodityCodeValue/> .
@prefix eurolithos: <https://data.geoscience.earth/ncl/geoera_eurolithos/> .

<http://inspire.ec.europa.eu/codelist/CommodityCodeValue> a skos:ConceptScheme;
dcterms:title "CommodityCodeValue"@en .

eurolithos:CommodityCodeValue a skos:ConceptScheme;
dcterms:title "CommodityCodeValue"@en;
skos:hasTopConcept inspire:dimensionStone .

inspire:dimensionStone a skos:Concept;
skos:inScheme <http://inspire.ec.europa.eu/codelist/CommodityCodeValue>,
eurolithos:CommodityCodeValue;
skos:prefLabel "Ornamental stone"@en;
skos:altLabel "Dimension stone"@en-US,"Natural stone"@en,"Building stone"@en-GB;
skos:definition "Dimension stone is natural stone.."@en;
skos:topConceptOf eurolithos:CommodityCodeValue;
skos:narrower inspire:marble,eurolithos:serpentinite .

inspire:marble a skos:Concept;
skos:inScheme
<http://inspire.ec.europa.eu/codelist/CommodityCodeValue>,eurolithos:CommodityCodeValue;
skos:prefLabel "marble"@en;
skos:definition "Commercial marble includes metamorphosed limestones and serpentine rocks.."@en;
dcterms:source <http://minerals.usgs.gov/minerals/pubs/commodity/stone_dimension/myb1-2007-
stond.pdf>;
skos:relatedMatch eurolithos:VerdeViana;
skos:broader inspire:dimensionStone .

eurolithos:serpentinite a skos:Concept;
skos:inScheme eurolithos:CommodityCodeValue;
skos:prefLabel "serpentinite"@en;
skos:broader inspire:dimensionStone;
dcterms:references eurolithos:Doe-2019 .

eurolithos:ornamental-stones a skos:ConceptScheme;
dcterms:title "ornamental stones"@en;
skos:hasTopConcept eurolithos:Portugal .

eurolithos:Portugal a skos:Concept;
skos:prefLabel "Portugal"@en;
skos:topConceptOf eurolithos:ornamental-stones;
skos:narrower eurolithos:VerdeViana .

eurolithos:VerdeViana a skos:Concept;
skos:prefLabel "Verde Viana Marble"@en,"Verde Viana mármore"@pt,"Verde Viana marmor"@no;
skos:altLabel "Viana Cristal Marble"@en,"Verde Viana Raminhado Marble"@en;
skos:broader eurolithos:Portugal;
skos:relatedMatch inspire:marble;
dcterms:references eurolithos:CEN-2016;
dbpo:colourName "green"@en;
geo:location <http://sws.geonames.org/8014856/>;
foaf:depiction <https://www.criteriofavorito.com/images/cache/catalogue/26/marmores-verde-viana3-
750x422.jpg>;
foaf:page <https://www.stonecontact.com/verde-viana-marble/s3646> .
```

```
eurolithos:references a skos:ConceptScheme;
dcterms:title "references"@en;
skos:hasTopConcept eurolithos:CEN-2016 .

eurolithos:CEN-2016 a skos:Concept;a dcterms:BibliographicResource;
skos:prefLabel "CEN (2016)"@en;
dcterms:bibliographicCitation "CEN, European Committee for Standardization (2016): Natural stone -
Denomination criteria; EN 12440:2017; Technical Committee CEN/TC 246, Avenue Marnix 17, B-1000
Brussels";
skos:topConceptOf eurolithos:references;
dcterms:source <https://dx.doi.org/10.31030/2773934> .

eurolithos:Doe-2019 a skos:Concept;a dcterms:BibliographicResource;
skos:prefLabel "Doe (2019)"@en;
dcterms:bibliographicCitation "Doe J. (2019): Description - Commodity types, ..";
skos:topConceptOf eurolithos:references .
```

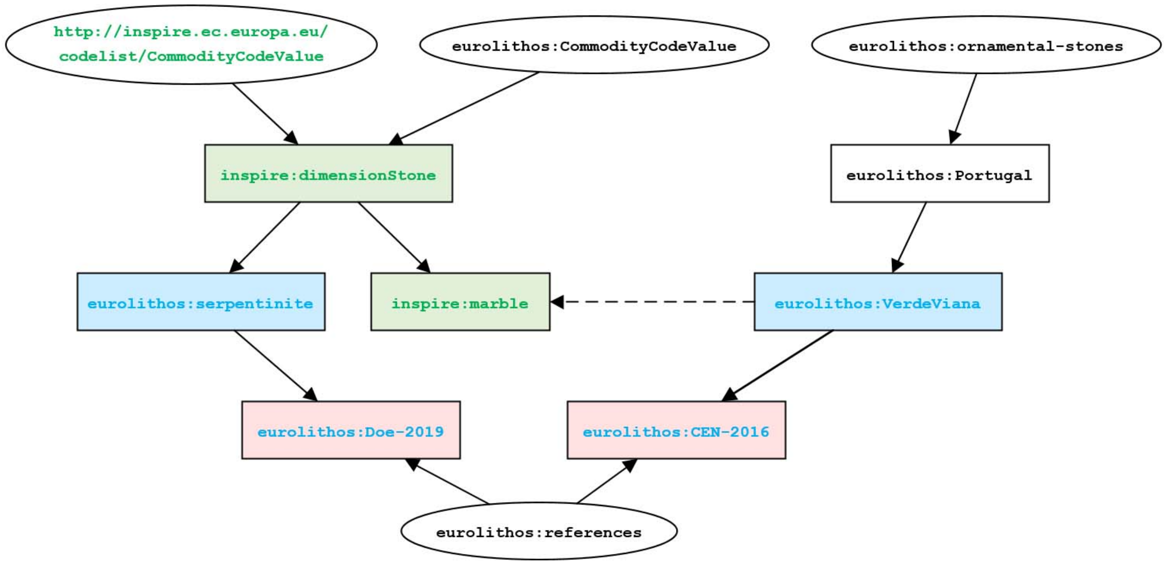


Figure 12. SKOS inScheme example

5.2.7 *Referencing Project Vocabulary entries from Standard dataset*

By L. Sőrés

5.2.7.1 Example – RESOURCE Grid Coverage Referencing Project Vocabulary Concepts

One important goal of Project Vocabularies is to provide links to domain specific concepts that can be referenced from external documents related to the same domain. In this example hydrogeological property names are referenced in the context of a hydrogeological grid coverage.

The RESOURCE project provides Pan European hydrogeological models delivered as a large set of shape files. Each shape file represents a hydrogeological layer containing 10x10 km rectangles that are coincident with the cells of the European INSPIRE grid and accompanied by a set of hydrogeological parameters. One way of encoding this dataset in INSPIRE standard format is using the ReferenceableGridCoverage model:

```
<gmlcov:ReferenceableGridCoverage gml:id="COV_1" xmlns:swe="http://www.opengis.net/swe/2.0"
xmlns:gml="http://www.opengis.net/gml/3.2" xmlns:gmlcov="http://www.opengis.net/gmlcov/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink"
xsi:schemaLocation="http://www.opengis.net/swe/2.0 http://schemas.opengis.net/sweCommon/2.0/swe.xsd
http://www.opengis.net/gmlcov/1.0 http://schemas.opengis.net/gmlcov/1.0/gmlcovAll.xsd">
   <gml:domainSet>
      <gml:MultiSurface>
         <gml:surfaceMember/>
         <gml:surfaceMember/>
         <gml:surfaceMember/>
      </gml:MultiSurface>
   </gml:domainSet>
   <gml:rangeSet>
      <gml:File>
         <gml:rangeParameters/>
         <gml:fileName>example.shp</gml:fileName>
         <gml:fileStructure codeSpace="">ESRI Shape</gml:fileStructure>
      </gml:File>
   </gml:rangeSet>
   <gmlcov:rangeType xlink:href="DREC_RESOURCE01.xml" xlink:title="RESOURCE Record structure"/>
</gmlcov:ReferenceableGridCoverage>
```

The standard requires exact definitions for the range type. It is provided in DREC_RESOURCE01.xml as a swe:DataRecord element. Due to technical restrictions length of field name in ESRI shape files are limited, so the project uses abbreviations. Proper parameter names, descriptions, explanations of the abbreviations can be stored in Project Vocabularies. In this example a skos:ConceptScheme (https://data.geoscience.earth/ncl/resource/observedProperty) contains RESOURCE parameter names that are referenced in the swe:field elements.

```
<swe:DataRecord id="DREC_RESOURCE01" xmlns:swe="http://www.opengis.net/swe/2.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink"
xsi:schemaLocation="http://www.opengis.net/swe/2.0
http://schemas.opengis.net/sweCommon/2.0/swe.xsd">
   <swe:field name="inspireGridCellIdentifier">
      <swe:Text
definition="https://data.geoscience.earth/ncl/resource/observedProperty/inspireGridCellIdentifier">
      </swe:Text>
   </swe:field>
   <swe:field name="serialNumberOfLayer">
      <swe:Count
definition="https://data.geoscience.earth/ncl/resource/observedProperty/serialNumberOfLayer">
      </swe:Count>
   </swe:field>
   <swe:field name="topOfLayer">
      <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/topOfLayer">
         <swe:uom xlink:href="m"/>
```

```
        </swe:Quantity>
    </swe:field>
    <swe:field name="bottomOfLayer">
        <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/bottomOfLayer">
            <swe:uom xlink:href="m"/>
        </swe:Quantity>
    </swe:field>
    <swe:field name="isAquifer">
        <swe:Boolean
definition="https://data.geoscience.earth/ncl/resource/observedProperty/isAquifer">
        </swe:Boolean>
    </swe:field>
    <swe:field name="isAquitard">
        <swe:Boolean
definition="https://data.geoscience.earth/ncl/resource/observedProperty/isAquitard">
        </swe:Boolean>
    </swe:field>
    <swe:field name="lithology">
        <swe:Category
definition="https://data.geoscience.earth/ncl/resource/observedProperty/lithology">
            <swe:codeSpace xlink:href="http://resource.geolba.ac.at/geoera_keyword/"/>
        </swe:Category>
    </swe:field>
    <swe:field name="proportion">
        <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/proportion">
            <swe:uom xlink:href="percent"/>
        </swe:Quantity>
    </swe:field>
    <swe:field name="delineationMethod">
        <swe:Category
definition="https://data.geoscience.earth/ncl/resource/observedProperty/DelineationMethod">
            <swe:codeSpace xlink:href="http://resource.geolba.ac.at/geoera_keyword/"/>
        </swe:Category>
    </swe:field>
    <swe:field name="porosity">
        <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/porosity">
            <swe:uom xlink:href="percent"/>
        </swe:Quantity>
    </swe:field>
    <swe:field name="horizontalConductivity">
        <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/horizontalConductivity">
            <swe:uom xlink:href="m/d"/>
        </swe:Quantity>
    </swe:field>
    <swe:field name="verticalConductivity">
        <swe:Quantity
definition="https://data.geoscience.earth/ncl/resource/observedProperty/verticalConductivity">
            <swe:uom xlink:href="m/d"/>
        </swe:Quantity>
    </swe:field>
    <swe:field name="parameterEstimationMethod">
        <swe:Category
definition="https://data.geoscience.earth/ncl/resource/observedProperty/ParameterEstimationMethod">
            <swe:codeSpace xlink:href="http://resource.geolba.ac.at/geoera_keyword/"/>
        </swe:Category>
    </swe:field>
    <swe:field name="isPaleogenic">
        <swe:Boolean
definition="https://data.geoscience.earth/ncl/resource/observedProperty/isPaleogenic">
        </swe:Boolean>
    </swe:field>
    <swe:field name="isArtesian">
        <swe:Boolean
definition="https://data.geoscience.earth/ncl/resource/observedProperty/isArtesian">
        </swe:Boolean>
    </swe:field>
    <swe:field name="isGeothermal">
        <swe:Boolean
definition="https://data.geoscience.earth/ncl/resource/observedProperty/isGeothermal">
        </swe:Boolean>
    </swe:field>
</swe:DataRecord>
```

5.2.7.2 example – RDF, Observed Properties in RESOURCE

Human readable parameter names, descriptions, explanations of the RESOURCE property names can be stored in a skos:ConceptScheme. Shape file abbreviations can also be added as skos:altLabel. A fragment of the RDF is listed here:

```
@prefix rsc: <https://data.geoscience.earth/ncl/resource/> .
@prefix op: <https://data.geoscience.earth/ncl/resource/observedProperty/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

rsc:observedProperty a skos:ConceptScheme ;
dcterms:title "Observed Properties used in the GeoERA RESOURCE project"@en;
skos:hasTopConcept op:topOfLayer ;
skos:hasTopConcept op:bottomOfLayer ;
skos:hasTopConcept op:isAquifer ;
…
.

op:topOfLayer a skos:Concept ;
skos:topConceptOf rsc:observedProperty ;
skos:prefLabel "top of layer"@en ;
skos:altLabel "Top" ;
skos:definition "[m] below surface level"@en ;
.
op:bottomOfLayer a skos:Concept ;
skos:topConceptOf rsc:observedProperty ;
skos:prefLabel "bottom of layer"@en ;
skos:altLabel "Bottom" ;
skos:definition "[m] below surface level"@en ;
.
op:isAquifer a skos:Concept ;
skos:topConceptOf rsc:observedProperty ;
skos:prefLabel "is aquifer"@en ;
skos:altLabel "Aquifer" ;
skos:definition "'Y' if layer is aquifer"@en ;
.
…
```

5.3 References

5.3.1 Bibliographic references

- Auer S., Bryl V., Tramp S. (Eds.) (2014): Linked Open Data – Creating Knowledge Out of Interlinked Data, Results of the LOD2 Project, Springer
- Blumauer A. (2019): White Paper, Knowledge Graphs, Transforming Data into Knowledge, SWC Vienna, Austria [https://www.poolparty.biz/knowledge-graphs/]
- INSPIRE Drafting Team (2014): INSPIRE Generic Conceptual Model; D2.5: Generic Conceptual Model, Version 3.4; Version for Annex II/III data specifications v3.0 https://inspire.ec.europa.eu/documents/inspire-generic-conceptual-model
- Lawson A.E., Alkhoury S., Benford R., Clark B.R., Falconer K.A. (2000): What kinds of scientific concepts exist? Concept construction and intellectual development in college biology; https://doi.org/10.1002/1098-2736(200011)37:9%3C996::AID-TEA8%3E3.0.CO;2-J
- Loudon, T V. 2012. A scenario for systems geology: Suggestions concerning the emerging geosciences knowledge system and the future geological map. *British Geological Survey Research Report RR/11/05*. 373 pp. ISBN: 978 0 85272 710 2; http://www.bgs.ac.uk/systemsgeology/
- "SKOS: A Guide for Information Professionals", American Library Association, May 18, 2015. http://www.ala.org/alcts/resources/z687/skos Document ID: df364f20-e2ad-4b0e-bc8e-9b4eff9bc777
- STARDOG UNION (2018): Knowledge Graphs 101, How to overcome a major enterprise liability and unleash massive potential
- Wood D., Zaidman M., Ruth L., Hausenblas M. (2014): Linked Data - structured data on the web; Manning Publications Co. US
- van Harmelen F., Lifschitz V., Porter B. (2008): Handbook of Knowledge Representation; in Foundations of Artificial Intelligence, Elsevier B.V
- Ziébelin D., Genoud P., Natete M., Cassard D., Tertre F. (2018): A Web of Data Platform for Mineral Intelligence Capacity Analysis (MICA); https://link.springer.com/chapter/10.1007/978-3-319-90053-7_15

5.3.2 Web links

[viewed December 11, 2018]

- Archer P., Goedertier S., Loutas N. (2012): D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC, ISA Interoperability Solutions for European Public Administrations, https://joinup.ec.europa.eu/document/10-rules-persistent-uris
- Berners-Lee T. (2006): Linked Data; Date: 2006-07-27, last change: $Date: 2009/06/18 18:24:33 $ Status: personal view only. Editing status: imperfect but published. https://www.w3.org/DesignIssues/LinkedData.html
- BGS (2018): Geoscience vocabularies; https://www.bgs.ac.uk/services/ngdc/management/geology/vocabulary.html
- Cagle K. (2019): Why Knowledge Bases Are The Next Big Thing https://www.forbes.com/sites/cognitiveworld/2019/10/10/why-knowledge-bases-are-the-next-big-thing/

- Coyne R. (2018): Knowledge Graphs for integrated data governance; https://www.topquadrant.com/2018/07/12/knowledge-graphs-for-integrated-data-governance/
- GBA (2018): GBA Thesaurus – a controlled vocabulary for geosciences; http://resource.geolba.ac.at/ or test site http://www.geolba.net/thesaurus
- Isaac A., Summers E. (2009): SKOS Simple Knowledge Organization System Primer - W3C Working Group Note, https://www.w3.org/TR/skos-primer/
- McCrae J.P. (2019): The LOD Cloud https://lod-cloud.net/
- McGinnis J. (2018): The Knowledge Graph and the Internet's Memory Palace; https://www.ontotext.com/knowledge-graph-memory-palace/
- Sőrés L. (2018): Description of 3D Geological Models using INSPIRE and Semantic Web principles; https://themes.jrc.ec.europa.eu/pages/revision/836209
- Stichbury J. (2017): WTF is a knowledge graph?, Unpicking a tangle of terminology to conclude it's semantic, smart and alive; https://hackernoon.com/wtf-is-a-knowledge-graph-a16603a1a25f
- W3C (2018): Data on the web current status, https://www.w3.org/standards/techs/dataontheweb
- Webber J. (2017): The Motivation for Native Graph Databases; https://www.datanami.com/2017/05/17/motivation-native-graph-databases/
- Webber J. (2018): Why Knowledge Graphs Are Foundational to Artificial Intelligence; https://www.datanami.com/2018/03/20/why-knowledge-graphs-are-foundational-to-artificial-intelligence/

5.4 Glossary

A selection of terms definitions used by Linked Data and Controlled Vocabularies – mostly from W3C glossary page https://www.w3.org/TR/ld-glossary/ (2019)

5 Star Linked Open Data
5 Star Linked Open Data refers to an incremental framework for deploying data. Tim Berners-Lee, the inventor of the Web and initiator of the Linked Data project, suggested a 5 star deployment scheme for Linked Open Data. The 5 Star Linked Data system is cumulative. Each additional star presumes the data meets the criteria of the previous step(s). 5 Star Linked Open Data includes an Open License (expression of rights) and assumes publications on the public Web.
Organizations may elect to publish 5 Star Linked Data, without the word "open", implying that the data does not include an Open License (expression of rights) and does not imply publication on the public Web.

- Publish data on the Web in any format (e.g., PDF, JPEG) accompanied by an explicit Open License (expression of rights).
- Publish structured data on the Web in a machine-readable format (e.g., XML).
- Publish structured data on the Web in a documented, non-proprietary data format (e.g., CSV, KML).
- Publish structured data on the Web as RDF (eg Turtle, RDFa, JSON-LD, SPARQL)
- In your RDF, have the identifiers be links (URLs) to useful data sources.

(W3C)
Asset Description Metadata Schema (ADMS)
ADMS is a profile of DCAT, used to describe semantic assets (or just 'Assets'), defined as highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. code lists, taxonomies, dictionaries, vocabularies) that are used for eGovernment system development. https://www.w3.org/TR/vocab-adms/
Application Programming Interface (API)
An API is an abstraction implemented in software that defines how others should make use of a software package such as a library or other reusable program. APIs are used to provide developers access to data and functionality from a given system. (W3C)
CC-BY-SA License
A form of Creative Commons license for resources released online. Work available under a CC-BY-SA license means you can include it in any other work under the condition that you give proper attribution. If you create derivative works (such as modified or extended versions), then you must also license them as CC-BY-SA. (W3C)
Closed World

A concept from Artificial Intelligence and refers to a model of uncertainty that an agent assumes about the external world. In a closed world (vs Open World), the agent presumes that what is not known to be true must be false. This is a common assumption underlying relational databases, most forms of logical programming. (W3C)

Concept

Concepts (fundamental element of SKOS) are the *units of thought* [*WillpowerGlossary*]—ideas, meanings, or (categories of) objects and events—which underlie many knowledge organization systems [*SKOS-UCR*]. As such, concepts exist in the mind as abstract entities which are independent of the terms used to label them. (SKOS primer)

Content Negotiation

Also called "conneg", refers to a phase in establishing a network connection. In the HTTP Protocol, the use of a message header to indicate which response formats a client will accept. Content negotiation allows HTTP servers to provide different versions of a resource representation in response to any given URI request. (W3C)

Controlled Vocabulary

Carefully selected sets of terms that are used to describe units of information; used to create taxonomies, thesauri and ontologies. In traditional settings the terms in the controlled vocabularies are words or phrases, in a linked data setting then they are normally assigned unique identifiers (URIs) which in turn link to descriptive phrases. (W3C)

Comma Separated Values (CSV)

A tabular data format in which columns of information are separated by comma characters. CSV files are a non-proprietary format and are considered 3-star data on the 5-star scale. (W3C)

Creative Commons Licenses

Licenses that include legal statements by the owner of copyright in intellectual property specifically allowing people to use or redistribute the copyrighted work in accordance with conditions specified therein. See also About Creative Commons Licenses. (W3C)

Data Modeling

Data modeling is a process of organizing data and information describing it into a faithful representation of a specific domain of knowledge. Linked data modeling applies modeling techniques based on Linked Data Principles. (W3C)

Dataset, RDF

A collection of RDF data, comprising one or more RDF graphs that is published, maintained, or aggregated by a single provider. In SPARQL, an RDF Dataset represents a collection of RDF graphs over which a query may be performed. (W3C)

Data Warehouse

A data warehouse is one approach to data integration in which data from various operational data systems is extracted, cleaned, transformed and copied to a centralized repository. The centralized repository can then be used for data mining or answering analytical queries. By contrast, Linked Data assumes a distributed approach of data management using HTTP URIs to describe and access information resources. A Linked Data approach is seen as an valid alternative to the centralized data warehouse approach especially when integrating datasets available on the public Web. (W3C)

DBpedia

DBpedia is a community effort to extract structured information from Wikipedia and make it available on the Web. DBpedia is often depicted as a hub for the Data Cloud. An RDF representation of the metadata held in Wikipedia and made available for SPARQL query on the World Wide Web. (W3C)

Dereferenceable URIs

When an HTTP client can look up a URI using the HTTP protocol and retrieve a description of the resource, it is called a dereferenceable URI. Dereferenceable URIs applies to URIs that are used to identify classic HTML documents and URIs that are used in the Linked Data context (cool URIs) to identify real-world objects and abstract concepts. (W3C)

Description Logic (DL)

DL is a family of knowledge representation languages with varying and adjustable expressivity. DL is used in artificial intelligence for formal reasoning on the concepts of an application domain. The Web Ontology Language (OWL) provides a standards-based way to exchange ontologies and includes a Description Logic semantics as well as an RDF based semantics. Biomedical informatics applications often use DL for codification of healthcare and life sciences knowledge. (W3C)

Data Catalog Vocabulary (DCAT)

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use. https://www.w3.org/TR/vocab-dcat/

Dublin Core Metadata Element Set

Dublin Core Metadata Element Set refers to a vocabulary of fifteen properties for use in resource descriptions, such as may be found in a library card catalog (creator, publisher, etc). The Dublin Core Metadata Element Set, also known as "DC Elements", is the most commonly used vocabulary for Linked Data applications. See also Dublin Core Element Set (W3C)

Dublin Core Metadata Initiative (DCMI Metadata Terms)

The Dublin Core Schema is a small set of vocabulary terms that can be used to describe digital resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks.[1] The full set of Dublin Core metadata terms can be found on the Dublin Core Metadata Initiative (DCMI) website. http://dublincore.org/documents/dcmi-terms/

Dublin Core Metadata Terms

A vocabulary of bibliographic terms used to describe both physical publications and those on the Web. An extended set of terms beyond those basic terms found in the Dublin Core Metadata Element Set. See also Dublin Core Metadata Terms (W3C)

Entity

In the sense of an entity-attribute-value model, an entity is synonymous with the Subject of an RDF Triple. (W3C)

ETL

ETL is an abbreviation for extract, transform, and load. Linked Data developers routinely extract data from a relational database, transform data to RDF Triples, and load it into an RDF database. (W3C)

Friend Of A Friend (FOAF)

FOAF is a project devoted to linking people and information using the Web. Regardless of whether information is in people's heads, in physical or digital documents, or in the form of factual data, it can be linked. FOAF integrates three kinds of network: social networks of human collaboration, friendship and association; representational networks that describe a simplified view of a cartoon universe in factual terms, and information networks that use Web-based linking to share independently published descriptions of this inter-connected world. FOAF does not compete with socially-oriented Web sites; rather it provides an approach in which different sites can tell different parts of the larger story, and by which users can retain some control over their information in a non-proprietary format.
http://xmlns.com/foaf/spec/

Graph

A collection of objects (represented by "nodes") any of which may be connected by links between them. (W3C). A network of nodes and edges.

JSON

JavaScript Object Notation (JSON) is syntax for storing and exchanging text based information. JSON has proven to be a highly useful and popular object serialization and messaging format for the Web. (W3C)

JSON-LD

JavaScript Object Notation for Linking Data is a language-independent data format for representing Linked Data, based on JSON. JSON-LD is capable of serializing any RDF graph or dataset and most, but not all, JSON-LD documents can be directly transformed to RDF. JSON-LD Syntax is easy for humans to read and write as well as, easy for machines to parse and generate. JSON-LD is an appropriate Linked Data interchange language for JavaScript environments, Web service and NoSQL databases. (W3C)

Knowledge base

A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems, which were the first knowledge-based systems. (Wikipedia)

Knowledge graph

A knowledge graph is an interconnected set of information that is able to meaningfully bridge enterprise metadata silos. (Coyne 2018). A knowledge graph allows you to store information in a graph model and use graph queries to enable your users to easily navigate highly connected datasets. (AWS 2018)

INSPIRE spatial object

An abstract representation of a real-world phenomenon related to a specific location or geographical area - identified by an unique object identifier. Within an INSPIRE information system a user may .. dereference a URI identifying a spatial object and the INSPIRE service would return a data as an abstraction (a spatial object) representing a real world phenomenon.
INSPIRE defines a spatial thing as anything with spatial extent, i.e. size, shape, or position. Thematic identifiers are represented as attributes of spatial objects describing the real-world phenomenon. (INSPIRE 2014)

Linked Data

A pattern for hyperlinking machine-readable data sets to each other using Semantic Web techniques, especially via the use of RDF and URIs. Enables distributed SPARQL queries of the data sets and a browsing or discovery approach to finding information (as compared to a search strategy). Linked Data is intended for access by both humans and machines. Linked Data uses the RDF family of standards for data interchange (e.g., RDF/XML, RDFa, Turtle) and query (SPARQL). If Linked Data is published on the public Web, it is generally called Linked Open Data. See also [Linked Data Principles]. (W3C)

Linked Data client

A Web client that supports HTTP content negotiation for the retrieval of Linked Data from URLs and/or SPARQL endpoints. A Linked Data client understands standard REST API, for example the Linked Data REST API. Examples of Linked Data clients include: Tim Berners-Lee's early Tabulator browser, gFacet, and the Callimachus Shell (CaSH). (W3C)

Linked Data Principles

Provide a common API for data on the Web which is more convenient than many separately and differently designed APIs published by individual data suppliers. Tim Berners-Lee, the inventor of the Web and initiator of the Linked Data project, proposed the following principles upon which Linked Data is based:

- Use URIs to name things;
- Use HTTP URIs so that things can be referred to and looked up ("dereferenced") by people and user agents;
- When someone looks up a URI, provide useful information, using the open Web standards such as RDF, SPARQL;
- Include links to other related things using their URIs when publishing on the Web. (W3C)

Linked Open Data (LOD)

Linked Data published on the public Web and licensed under one of several open licenses permitting reuse. Publishing Linked Open Data enables distributed SPARQL queries of the data sets and a "browsing" or "discovery" approach to finding information, as compared to a search strategy. (W3C)

Linked Open Data Cloud (LOD Cloud)

A colloquial phrase for the total collection of Linked Data published on the Web. (W3C)

Machine Readable Data

Data formats that may be readily parsed by computer programs without access to proprietary libraries. For example, CSV, TSV and RDF formats are machine readable, but PDF and Microsoft Excel are not. Creating and publishing data following Linked Data principles helps search engines and humans to find, access and re-use data. Once information is found, computer programs can re-use data without the need for custom scripts to manipulate the content. (W3C)

Metadata

Information used to administer, describe, preserve, present, use or link other information held in resources, especially knowledge resources, be they physical or virtual. Metadata may be further subcategorized into several types (including general, access and structural metadata). Linked Data incorporates human and machine readable metadata along with it, making it self describing. (W3C)

Modeling Process

Modeling process in the context of RDF refers to the act by subject matter experts to work with developers to capture the context of data and define the relationships of the data. By doing so, high quality of Linked Data is obtained since capturing organizational knowledge about the meaning of the data within the RDF data model means the data is more likely to be reused correctly. Well defined context ensures better understanding, proper reuse, and is critical when establishing linkages to other data sets. (W3C)

Named Graph

Named graphs are a key concept of Semantic Web architecture in which a set of Resource Description Framework statements (a graph) are identified using a URI, allowing descriptions to be made of that set of statements such as context, provenance information or other such metadata. (Wikipedia)

Native triple store

Triple stores that have been built as database engines from scratch - while others have been built on top of existing commercial relational database engines (such as SQL-based) or NoSQL document-oriented database engines. It seems likely that native triple stores will have the advantage for performance over a longer period of time. (Wikipedia)

Notation3 (N3)

It has a readable RDF syntax used for expressing assertion and logic. N3 is a superset of RDF, extending the RDF model by adding formulae (literals which are graphs themselves), variables, logical implication, and functional predicates. (W3C)

N-Triples

A subset of Turtle that defines a line-based format to encode a single RDF graph. Used primarily as an exchange format for RDF data. See also N-triples. (W3C)

Object

In the context of RDF, the object is the final part of an RDF statement. See also [Subject] [Predicate] (W3C)

Ontology

A formal model that allows knowledge to be represented for a specific domain. An ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms). (W3C)

Open Government Data

Refers to content that is published on the public Web by government authorities in a variety of non-proprietary formats. (W3C)

Open-world assumption

In a formal system of logic used for knowledge representation, the open-world assumption is the assumption that the truth value of a statement may be true irrespective of whether or not it is known to be true. It is the opposite of the closed-world assumption, which holds that any statement that is true is also known to be true...(Wikipedia)

Persistent Identifier Scheme

A persistent identifier scheme is a mechanism for resolution of virtual resources. Persistent Uniform Resource Locator (PURLs) implement one form of persistent identifier for virtual resources. PURLs are valid URLs and their components must map to the URL specification. The scheme part tells a computer program, such as a Web browser, which protocol to use when resolving the address. The scheme used for PURLs is generally HTTP. Other persistent identifier schemes include Digital Object Identifiers (DOIs), Life Sciences Identifiers (LSIDs) and INFO URIs. All persistent identification schemes provide unique identifiers for (possibly changing) virtual resources, but not all schemes provide curation opportunities. (W3C)

Persistent Uniform Resource Locator

URLs that act as permanent identifiers in the face of a dynamic and changing Web infrastructure. Persistent Uniform Resource Locators (PURLs) redirect to the current location of or proxy specific Web content. A user of a PURL always uses the same Web address, even though the resource in question may have moved or changed ownership. (W3C)

Predicate

The middle term (the linkage, or "verb") in an RDF statement. For example, in the statement "Alice knows Bob" then "knows" is the predicate which connects "Alice" (the subject of the statement) to "Bob" (the object of the statement). (W3C)

Provenance

Data related to where, when and how information was acquired. (W3C)

Quad Store

A colloquial phrase for an RDF database that stores RDF triples plus an additional element of information, often used to collect statements into groups. (W3C)

Query

Programmatic retrieval of resources and their relationships. Using the SPARQL language, developers issue queries based on (triple) patterns. (W3C)

R2RML

R2RML (RDB to RDF Mapping Language) is a language for expressing customized mappings from relational databases to RDF datasets. Such mappings provide the ability to view existing relational data in the RDF data model, expressed in a structure and target vocabulary of the mapping author's choice. (W3C)

RDF database

A type of database designed specifically to store and retrieve RDF information. May be implemented as a triple store, quad store or other type. (W3C)

RDF Schema

The simplest RDF vocabulary description language. It provides much less descriptive capability than the Simple Knowledge Organization System (SKOS) or the Web Ontology Language (OWL). A standard of the W3C (W3C)

RDF/XML

An RDF syntax encoded in XML. A standard of the W3C. (W3C)

Representational State Transfer (REST)

An architectural style for information systems used on the Web. It explains some of the Web's key features, such as extreme scalability and robustness to change. REST is the foundation of the World Wide Web and the dominant Web service design model. The term "Representational State Transfer" was introduced and defined in 2000 by Roy Thomas Fielding in his doctoral dissertation. See also "Architectural Styles and the Design of Network-based Software Architectures" by Roy Thomas Fielding. (W3C)

Resource

In an RDF context, a resource can be anything that an RDF graph describes. A resource can be addressed by a Unified Resource Identifier (URI). (W3C)

Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats. It is also used in knowledge management applications. https://www.w3.org/RDF/

REST API

An application programming interface (API) implemented using HTTP and the principles of REST to allow actions on Web resources. The most common actions are to create, retrieve, update and delete resources. See also Representational State Transfer. (W3C)

Schema

Schema refers to a data model that represents the relationships between a set of concepts. Some types of schemas include relational database schemas (which define how data is stored and retrieved), taxonomies and ontologies. (W3C)

Semantic Technologies

The broad set of technologies that related to the extraction, representation, storage, retrieval and analysis of machine-readable information. (W3C)

Semantic Web

An evolution or part of the World Wide Web that consists of machine-readable data in RDF and an ability to query that information in standard ways (e.g. via SPARQL) (W3C)

Semantic Web Search Engine

A search engine capable of making use of semantic technologies to model its knowledge base and to deliver content. (W3C)

Semantic Web Standards

Standards of the World Wide Web Consortium relating to the Semantic Web, including RDF, RDFa, SKOS, OWL and SPARQL. (W3C)

Simple Knowledge Organization System (SKOS)

Many knowledge organization systems, such as thesauri, taxonomies, classification schemes and subject heading systems, share a similar structure, and are used in similar applications. SKOS captures much of this similarity and makes it explicit, to enable data and technology sharing across diverse applications. The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive language for developing and sharing new knowledge organization systems. It may be used on its own, or in combination with formal knowledge representation languages such as the Web Ontology language (OWL). https://www.w3.org/TR/skos-reference/

SPARQL Query Language for RDF (SPARQL)

RDF is a directed, labeled graph data format for representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs. https://www.w3.org/TR/rdf-sparql-query/

SPARQL endpoint

A service that accepts SPARQL queries and returns answers to them as SPARQL result sets. It is a best practice for datasets providers to give the URL of their SPARQL endpoint to allow access to their data programmatically or through a Web interface. A list of some endpoints status is available at http://labs.mondeca.com/sparqlEndpointsStatus/ (W3C)

Subject

The subject is the first part of an RDF statement. A subject in the context of a triple <?s ?p ?o> refers to who or what the RDF statement is about. (W3C)

Triple

An RDF statement, consisting of two things (a "subject" and an "object") and a relationship between them (a verb, or "predicate"). This subject-predicate-object triple forms the smallest possible RDF graph (although most RDF graphs consist of many such statements). (W3C)

Triple store

A colloquial phrase for an RDF database that stores RDF triples. (W3C)

Turtle

An RDF serialization format designed to be easier to read than others such as RDF/XML. The term "Turtle" was derived from Terse RDF Triple Language. Turtle allows an RDF graph to be written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. Turtle provides levels of compatibility with the existing N-Triples format, as well as the triple pattern syntax of the SPARQL W3C Recommendation. (W3C)

Uniform Resource Identifier (URI)

A global identifier standardized by joint action of the World Wide Web Consortium and Internet Engineering Task Force. A Uniform Resource Identifier (URI) may or may not be resolvable on the Web. URIs play a key role in enabling Linked Data. URIs can be used to uniquely identify virtually anything including a physical building or more abstract concepts such as colors. See also Internationalized Resource Identifier (IRI) and Uniform Resource Locator (URL). See also Uniform Resource Identifier (URI): http://www.w3.org/DesignIssues/Architecture.html.

URIs have been known by many names: Web addresses, Universal Document Identifiers, Universal Resource Identifiers. If you are interested in the history of the many names, read Tim Berners-Lee's design document Web Architecture from 50,000 feet. (W3C)

Uniform Resource Locator (URL)
A global identifier for Web resources standardized by joint action of the World Wide Web Consortium and Internet Engineering Task Force. A URL is resolvable on the Web and is commonly called a "Web address". All HTTP URLs are URIs however, not all URIs are URLs. See also Internationalized Resource Identifier and Uniform Resource Identifier. (W3C)
Validation Service
The W3C offers an RDF validation service to check and validate RDF files. It is considered a best practice to validate RDF files prior to publishing them on the Web. See http://www.w3.org/RDF/Validator/. See also http://www.w3.org/People/Barstow/#online_parsers . (W3C)
Vocabulary
A collection of "terms" for a particular purpose. Vocabularies can range from simple such as the widely used RDF Schema, FOAF and Dublin Core Metadata Element Set to complex vocabularies with thousands of terms, such as those used in healthcare to describe symptoms, diseases and treatments. Vocabularies play a very important role in Linked Data, specifically to help with data integration. The use of this term overlaps with Ontology. (W3C)
Vocabulary Alignment
The process of analyzing multiple vocabularies to determine terms that are common across them and to record those relationships. (W3C)
Vocabulary of Interlinked Datasets (VoID)
VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets and a standard of the World Wide Web Consortium. VoID is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloging and archiving of datasets. VoID can be used to express general metadata based on Dublin Core, access metadata, structural metadata, and links between datasets. (W3C)
Web of Data
A subset of the World Wide Web which contains machine readable data represented as Linked Data. (W3C)
Web of Documents
The original, or traditional, World Wide Web in which published resources were nearly always documents as opposed to machine readable data. (W3C)
Web Ontology Language (OWL)
OWL is a family of knowledge representation and vocabulary description languages for authoring ontologies, based on RDF and standardized by the W3C. (W3C)
Web Resource
A web page addressed by a URL. Examples include: an HTML web page, an image offered by a web server, or a dataset accessible by a URL. A Web Resource may have different representations. For example, an RDF database might be accessed at a single URL using multiple syntaxes, such as RDFa, JSON-LD, and Turtle. (W3C)
World Wide Web Consortium (W3C)
An international community that develops and promotes protocols and guidelines for the long-term growth for the Web. W3C's standards define key parts of the World Wide Web, including Web Design, Web Architecture and the Semantic Web. (W3C)