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

Authors and affiliation:
Ch. Hörfarter [GBA]
M. Sanabria Pabón [IGME]
Román Hernández [IGME]
M. Novak [GEO-ZS]
L. Kondrová [CGS]

Deliverable 4.2

GeoERA Keyword Thesaurus

E-mail of lead author:
christine.hoerfarter@geologie.ac.at

Version: 13-11-2019

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	D4.2	
Dissemination level	Public	
Deliverable name	GeoERA Keyword Thesaurus	
Work package	WP4, Semantic Harmonization Issues	
Lead WP/Deliverable beneficiary	GBA	
Deliverable status		
Submitted (Author(s))	13/11/2019	Christine Hörfarter
Reviewed	13/11/2019	Martin Schiegl
Approved (Coordinator)	30/11/2019	Jørgen Tulstrup

Please note:
This paper will be finalized for the GeoERA report D4.4 and reconciled together with the deliverables D4.1. and D4.3 (GeoERA project vocabulary) in June 2021. This refers especially to the keyword thesaurus RDF file and the chapters about the “Governance Plan and Workflows” actually created during the GeoERA project.

For further information on references and terminology used in this document please have a look at the GeoERA GIP WP4 “D4.3 Project Vocabulary” report chapter 5.3 References and chapter 5.4 Glossary.

Acknowledgement:
Thanks the project members and all other supporting parties for the committed and constructive cooperation and the dedication for the benefit of this European wide program.

TABLE OF CONTENTS

1 INTRODUCTION 4
1.1 General Description 4
1.2 Scope of the keyword thesaurus 5
1.3 Current Version of the Keyword Thesaurus 5

2 EVALUATION OF EXISTING VOCABULARIES APPLICABLE FOR SUBJECT HEADINGS 6
2.1 Evaluation process 7
2.1.1 Survey among task partners 7
2.1.2 Search categories and terms extracted from GeoERA projects 10
2.1.3 Codelists and added vocabularies 13
2.1.4 Grouped vocabularies. “Possible search category Thesauri”. 14
2.1.5 Evaluation criteria 14
2.1.6 Database 15
2.2 Evaluation results 16
2.2.1 Vocabularies analyzed. 16
2.2.2 Search categories thesauri 20
2.3 Conclusions 30
2.3.1 Test keywords 30
2.3.2 Second evaluation 32
2.3.3 Final selection 34

3 COMPILATION OF THE GEOERA KEYWORD THESAURUS 40
3.1 Compilation process 40
3.1.1 Modelling and generating RDF file structure 40
3.1.2 First phase of compilation 43
3.1.3 Second phase of compilation 47
3.1.4 Translation 47
3.2 Integration and validation 47

4 GOVERNANCE PLAN AND WORKFLOWS AROUND THE KEYWORD THESAURUS 49
4.1 Governance Plan 49
4.2 Workflows 50
4.2.1 Management of changes, revision of keywords, translations, update and extension workflow during the project and after the project end 50
4.2.2 Backup processes 51
4.2.3 Management of the domain of the terms 51
4.2.4 Maintenance of the service 51
4.2.5 Contact point/support/information 51
4.2.6 Licensing 51
4.3 Use of the keyword thesaurus in the metadata catalogue 51
4.4 References 51

1 INTRODUCTION

1.1 General Description

Participants: GBA, IGME, ISPRA, SGU, TNO, CGS, GIU, GeoZS, MBFSZ, LfU, BGRM, GTK, GEUS, BGR, HGI-CGS, LNEG
This "T4.1 Multilingual semantic text search" report describes the establishment of a keyword thesaurus, which is created to support the semantic text search functionality for metadata concerning GeoERA project datasets. Hence, to improve the search capabilities within the whole EGDI metadata catalogue.

The GeoERA IP WP4 "Semantic Harmonization Issues" is to support two principal use cases for GeoERA projects. The supported use cases are "Multilingual Semantic Text Search" and "GeoERA project vocabularies" via Linked Open Data and SKOS/RDF. Both aim to ensure interoperability of GeoERA project results and make them searchable (e.g. by keywords). Thus, "Semantic Harmonization" stands for "making datasets and their attribute data consistent and compatible relating to the meaning in language and logic".

One reason why WP4 introduces Linked Data technology to GeoERA projects is to enable a Semantic Text Search. Search for data is the basic task for all data infrastructures. It needs to put all keywords used to tag datasets into a single hierarchy like a thesaurus. Data queries then can use this kind of a word net also to get search results for similar keywords within a "semantic radius".
For metadata descriptions, the clarification of the meaning of textual attributes applies mainly to keywords and the implementation of a semantic search within a metadata catalog.

Here, the Multilingual Semantic Text Search task of WP4 strives for a compilation (SKOS thesaurus) of keywords with URIs suitable for tagging metadata.
Hence, it is the aim of this task to establish a multilingual keyword thesaurus based on SKOS/RDF for a subject heading system which provides an ordered collection of terms used in the geoscientific area for indexing, storing and retrieving GoeERA datasets "

The multilingual semantic text search task T4.1 is subdivided into three subtasks:

- Evaluation of existing vocabularies applicable for subject headings (led by IGME)
    - includes a survey of vocabularies suitable for keywords and evaluates the covered geoscientific domains, the scope, granularity, and other criteria. It selects which existing terminology is suitable for a new GeoERA/EGDI subject heading system.

- Compilation of a keyword thesaurus (led by GeoZS)
    - modelling the subject heading system from selected and tested vocabularies, to create a new keyword thesaurus and also to complete and translate missing keywords.

- Governance plan, workflows around keyword thesaurus (led by CGS)
    - designs a governance plan for a keyword thesaurus including workflows for application, crosslinking to other Linked Data resources, and thesaurus maintenance - in order to establish a multilingual and semantic subject heading system for the GeoERA platform.

The progress and current results regarding these tasks are described in this report on the following pages.

1.2 Scope of the keyword thesaurus

The aim of T4.1 is to create a multilingual (GeoERA and EGDI) keyword thesaurus (SKOS/RDF file) to use as a subject heading system in a geoscientific context for tagging GeoERA project datasets. That means, to search, select and compile the thematic relevant keywords with preferred-, synonym- and hidden-labels in some EU languages with URIs from the source vocabularies used for the keyword compilation. Here, some examples of URIs as source for the keyword “quarry”, which occurs in several vocabularies:

- http://inspire.ec.europa.eu/codelist/MiningActivityTypeValue/quarry
  http://www.eionet.europa.eu/gemet/concept/6867
  http://resource.geosciml.org/classifier/cgi/mining-activity/quarrying
  http://data.uba.de/umt/_00023359, http://dbpedia.org/resource/Quarry
  https://www.wikidata.org/wiki/Q188040
  http://thes.bncf.firenze.sbn.it/termine.php?id=15550
  http://www.enciclopedia-aragonesa.com/voz.asp?voz_id=3070
  http://vocab.getty.edu/page/aat/300000402

  and further more.

In difference to a controlled vocabulary of GeoERA projects, these keyword thesaurus terms (concepts) are without description and scientific reference which are mandatory properties for a project vocabulary.

When the keyword thesaurus is fully implemented (e.g. used in MICKA Geonetwork and published on the EGDI triple store/Sparql endpoint) the only valid use case (by now) is to search datasets by metadata keywords – multilingual, with auto complete (e.g. 10 selected and proposed terms), and optional semantic radius (include all related concepts), and relevance sorted search results. Maybe WP4 additionally will provide “What you could be interested in too…” keywords.

In the end, the GeoERA keyword thesaurus is a SKOS/RDF file and shall be uploaded to the RDF triplestore and be available via SPARQL endpoint (= web service*). Via endpoint or directly imported into Geonetwork (latest edition) it is ready for use immediately when working with Geonetwork. Because this keyword thesaurus is also available via an endpoint, a HTML website (e.g. Javascript) can find out related terms and send them to query Geonetwork (= query with semantic radius) or support many languages for an autocomplete text search.

1.3 Current Version of the Keyword Thesaurus

Up to now, a collection of ~2500 geoscientific terms (keywords) in English, with unique web addresses (URIs), delivered in RDF format, with translations in different languages, with links to standardized codelists (INSPIRE, GeoSciML, GEMET) is provided in a first version. It has been extracted in form of a SKOS/RDF file and has been sent in the mid of August 2019 to the GIP Project leader. Up to now, the translation of keywords is not completed yet. To be most efficient it was agreed, that the translation will be done by WP4 task partners in a revised version of the keyword thesaurus in 2020. The revision refers mostly to quality improvements like hierarchy modelling, merging of concepts, detection of synonyms and literal errors.
A prototype of the functionality of the current possible multilingual semantic text search is provided by GBA and can be tested via https://schmar00.github.io/semantic-search/.

2 EVALUATION OF EXISTING VOCABULARIES APPLICABLE FOR SUBJECT HEADINGS

By Román Hernández, Margarita Gómez, Margarita Sanabria)

WP4 subtask 4.1.1. “Evaluation of existing vocabularies applicable for subject headings” is led by **IGME** and the following partners are involved: GBA, ISPRA, SGU, TNO, CGS, GIU, GeoZS, MBFSZ, LfU, BRGM, GTK, GEUS, BGR, HGI-CGS, and LNEG.

The aim of this first activity is to investigate existing vocabularies for geosciences applicable for subject headings (evaluation). This task includes a survey of vocabularies suitable for keywords and evaluates the covered geoscientific domains, the scope, granularity, and other criteria. It selects which existing terminology is suitable for a new GeoERA/EGDI subject heading system.

EXECUTIVE WP4 Subtask 4.1.1 REPORT SUMMARY

The following pages detail the process followed by the sub-task 4.1.1 on the evaluation of existing vocabularies applicable for a GeoERA subject heading system to analyze existing vocabularies that can be integrated with a semantic search system in the future – the Keyword Thesaurus (GeoERA subject heading system).

Initially, a questionnaire was designed and have been sent out to the subtask partners in order to collect available existing vocabularies at their institutions. At the same time some known interesting and important vocabularies were added as well, like INSPIRE, CGI codelists and some other we were aware of.

A database has been created to store all collected information and facilitate its evaluation.
For focusing the evaluation of the topics that frame GeoERA projects, abstracts and deliverables of the different GeoERA projects were analyzed. Relevant geoscientific terms were extracted, and grouped in - what has been called - “Search Categories“. These search categories reflect the relevant scientific topics to be covered. Simultaneously, test words were extracted for each of these search categories.
Vocabularies were analyzed concerning their extension (number of terms) and depth (number of hierarchical levels). The evaluation criteria were established following qualitative and quantitative principles, and may be in some cases selective.
The evaluation of the vocabularies will be reevaluated for the subtask partners and for the other GeoERA projects responsible.

As a result of this activity 15 Search Categories were defined, 134 different vocabularies and codelists have been analyzed using the criteria of the 15 search categories, giving a total of 145 analyses. These vocabularies add up to more than 8,000 terms overall.
All the information collected and analyzed is decribed in detail further in this document, as well as the result of this analysis.
This document will be used to build a basis for the deliverable D4.1 Keyword Thesaurus (RDF file) and the subtask T4.1 Compilation of a keyword thesaurus.

2.1 Evaluation process

2.1.1 Survey among task partners

Throughout its working progress, each geological survey and each organization related to Geosciences have been developing their own vocabularies applied to the different fields of knowledge that are the object of their activities.
For this reason, it was considered from the beginning that it was necessary to start from the experience and knowledge, already encountered in this subject, of the geological surveys participating in GeoERA and the organizations related to earth sciences.
Sometimes these organizations present vocabularies that are broad in number of topics addressed, but elaborated very generally, while in others their vocabularies have few topics but very specific and very detailed.
Thus, in order to accomplish the first task of WP4 "4.1 Multilingual semantic text search", a questionnaire was designed and prepared by GBA and IGME.
The questionnaire was disseminated to all the participants of the subtask 4.1.1, consisting 25 questions in order to gather a complete information about the existence of these vocabularies, their domains, scope, formats and availability, granularity, etc.
The fundamental objective of this questionnaire was to be able to gather, for its later analysis and evaluation, the existing standard terminology, and thus be able to compile a new multilingual keyword thesaurus applicable for the complete GeoERA project.
Therefore, it was necessary to extract the maximum amount of information, and as exhaustively as possible, about available vocabularies and thesauri to facilitate the next phase of its review and analysis. Thus, to be able to select which of the existing terminologies are the most appropriate for a new subject heading system of topics in GeoERA / EGDI.

The questionnaire includes 25 questions (Table 1): the first 6 questions refer mainly to general aspects such as the title, organization, web representation, etc.; The Questions between 7 and 10 collect information regarding the general scope and specific scopes of the vocabulary, depending on the GeoERA domains. Questions 11 and 12 refer to language aspects and finally from question 13 to 25, the information requested cover several topics, for instance whether the vocabulary is hierarchically structured, if it is continuously updated, the extent, data format, terms of use, and further more.

Nº	Question	Explanation
1	**Title**	**Name of this vocabulary?**
2	**Point of contact**	What is the email from the responsible personf of this vocabulary list? (email address)
3	**Organization name. Thesaurus provider**	Name of organization – who provides this vocabulary? (free text)
4	**Short Description**	Please describe the vocabulary (free text)
5	**Web Representation**	Is there a web page to view/browse/download this vocabulary? (free text)
6	**Maintenance (if different from provider)**	Who is responsible for the editorial maintenance?
7	**General Scope**	Which geoscientific domains are covered by this vocabulary? one or more of the following (45 terms) or free text
8	**Specific Scope**	(Answer Yes or No)

№	Question	Explanation
	Geoscientific domains that cover to the GeoERA theme: GEO-ENERGY	
9	**Specific Scope** **Geoscientific domains that cover to the GeoERA theme: GROUNDWATER**	(Answer Yes or No)
10	**Specific Scope** **Geoscientific domains that cover to the GeoERA theme: RAW MATERIALS**	(Answer Yes or No)
11	**Multilingual?**	Is this vocabulary available in multiple languages? If yes, which languages? (free text, languages)
12	**Translation**	Is this vocabulary available in multiple languages? If yes, which languages? (free text, languages)
13	**Identifiers?**	Does this vocabulary have IDs for the concepts or just text-strings? If yes, what kind of (numbers, strings, URIs, web addresses)
14	**Hierarchically structured?**	(yes or no)
15	**Up to date?**	At what year the main part was created? Years of following updates? (years)
16	**Derivative (history)?**	Is this vocabulary extracted or rebuilt from one or more other older vocabularies? If yes, from which? (free text)
17	**Mapping Relation**	Does this vocabulary provide links to other published vocabularies? If yes, which? (yes or no and free text)
18	**Extent**	This vocabulary consists of how many geoscientific concepts/terms? How many of them are considered for the GeoERA domains (Geo-energy, Groundwater. Raw materials)? (number)
19	**Data Format**	Which data format is supported for this vocabulary (skos, owl, rdf, relational database, csv, xls, docx, others etc.) If other, describe
20	**Common in which region? (eg. Europe)**	Is this vocabulary well known (also when origin is from outside Europe)? If yes, within which European countries? (free text)
21	**Suitable?**	Do you think this vocabulary is suitable for metadata (keywords, subject headings) to describe EGDI datasets and webservices? Is it too detailed or too general? (free text)
22	**Use case**	Was/is this vocabulary already in use to encode or index geoscientific datasets? If yes, refer to one examplary dataset

Nº	Question	Explanation
23	**Purpose**	Please describe the original intent when the vocabulary was created? (e.g. library subject headings, knowledge base for terminology, official nomenclature, other..) (free text)
24	**Terms of use**	What is the license to reuse the vocabulary (royalty free use, creative commons, open science, etc.)? (free text)
25	**Comments on quality**	Please give a rating of the quality of concepts/terms and concept descriptions (if there are any) Subjects to consider: -references –definitions - relations…. (free text)

Table 1 Question of the survey which has been sent to the task partners

In mid-July 2018, the questionnaire was distributed among all partners participating in this sub-task. The deadline for replies was set at the beginning of September 2018. In order to encourage the completion of the questionnaire and to illustrate the type of reply that was expected for each question 7 examples where included (Table 2), mainly from GBA and IGME:

Title	Organism name.Thesaurus provider
GEMET	European Environment Agency (EEA)
GBA Thesaurus **Lithology**	Geological Survey of Austria
GeoSciML Simple **Lithology**	CGI Geoscience Concept Definitions Task Group (CDTG)
INSPIRE codelist **lithology**	JRC INSPIRE Registry Team
GBA Thesaurus **Geologic Timescale**	Geological Survey of Austria
CGI - **Geologic Time** Vocabulary - International Chronostratigraphic Chart - 2017	Research Vocabularies Australia - Linked Data API
Tesauro IGME de **ciencias de la tierra** IGME	Instituto Geológico y Minero de España, IGME

Table 2 Examples: Vocabularies completed included as example in the questionnaire sent

17 vocabularies were received (Table 3):

Title	Organism
GBA Thesaurus Rohstoffgeologie (**Raw Material**)	Geological Survey of Austria
INSPIRE Codelist **GeophProperty**NameValue	MBFSZ, Mining and Geological Survey of Hungary
INSPIRE Codelist **GeophProcess**NameValue	MBFSZ, Mining and Geological Survey of Hungary
INSPIRE Codelist **GeophProcess**ParameterNameValue	MBFSZ, Mining and Geological Survey of Hungary
INSPIRE Codelist **Resource**TypeValue	MBFSZ, Mining and Geological Survey of Hungary

Minerals4EU **Metadata Keywords**	Czech Geological Survey
OneGeology-Europe **keywords database**	Czech Geological Survey
Multilingual geological thesaurus (eWater, eEarth)	Czech Geological Survey
Geological codelists (Czech-English) - **chronostratigraphy, lithostratigraphy, regional geology, lithology**	Czech Geological Survey
English-Czech and Czech-English Professional Dictionary of **Hydrogeology** Hydrogeologický slovník (Aj, Čj)	Czech Geological Survey
Geologický slovník (Aj, Čj) **Dictionary of Geology** English-Czech and Czech-English	Czech Geological Survey
Catalogue of **geohazards**	Czech Geological Survey
Decorative stones	Czech Geological Survey
Applied geophysics - methods	Czech Geological Survey
Slope instabilities	Czech Geological Survey
Geological encyclopaedia on-line	Czech Geological Survey

Table 3 List of vocabularies collected among subtask partners

The resulting completed questionnaire with the 25 questions asked and the 24 vocabularies (reported ones and examples) collected are included in Annex I.

2.1.2 Search categories and terms extracted from GeoERA projects

The two main objectives of the GeoERA keyword thesaurus are to facilitate:

- The search of products within the GeoERA catalogue (metadata responsible).
- The assignment of keywords to each product produced within GeoERA (metadata responsible).

Therefore, before proceeding to evaluate existing thesauri for geosciences it was necessary to establish which are the generic geoscientific domains or topics that frame the products generated by the different GeoERA projects. Once these generic topics are established, thesauri can be classified referring to these topics and subsequently evaluated. This procedure ensures that the thesauri integrated in the GeoERA thesaurus are really focused on the needs of GeoERA projects in terms of product search and metadata tagging.

In order to establish these generic topics, the abstracts and deliverables of the 14 GeoERA projects were analysed. The objective of this analysis was to extract representative words (hereinafter referred to as "extracted terms") from each of the products from the GeoERA projects, and to group those words into general geoscientific categories. To establish the generic topics/domains, the Spanish List of Subject Headings for Public Libraries (http://www.bne.es/es/Micrositios/Publicaciones/MEMBNE/) was used. Then a description to each topic/domain was assigned. A first result in the form of an Excel file was presented in the first WP4 internal meeting held on 10 - 11 October 2018 in Vienna. In this first approach 14 generic topics/domains were established (Figure 2.1.1) and for each of the terms extracted from each GeoERA project, one or several generic topics were assigned.

During the meeting, it was agreed to call the generic topics/domains "Search Categories" and the descriptions were collaboratively refined (Figure 1.1.2 and Annex II). It was also agreed to send the file to the project leaders and GIP contacts so they could:

- Check if the established "Search Categories" were adequate and really framed the products provided by GeoERA projects.
- Validate, correct or add "extracted terms" and their classification in several "Search Categories"

MAIN DOMAIN	DEFINITION
GEOCRONOLOGY	Highlights related to absolute ages and layering (stratification) of these ages for all sets of rocks, fossils and sediments. The following disciplines are considered: stratigraphy, geological history, geological time scale, Chronostratigraphy …
GEOTHERMAL RESOURCES	Highlights related to resources and reserves from geothermal energy.
GEOLOGICAL PROCESSES	Highlights related to geology and dynamic geological processes that act on land forms and surfaces. The following disciplines are considered: sedimentation, diagenesis, metamorphism, geomorphology ….
STRUCTURAL GEOLOGY	Highlights related to the study of the deformation of the lithosphere, trying to reconstruct the movements and processes that have originated its structure, the history of movements and deformations on a global and regional scale. The following disciplines are considered: Tectonic,..
APPLIED GEOPHYSIC	Highlights related to the use of geophysical techniques mainly for the exploration and exploitation of natural resources; but also applied to geological risks, environment and hydrogeology.
MODELING	Highlights related to the understanding of the geological environment and geological processes, using numerical, geostatistics and simulation techniques. The following disciplines are considered: 3D modelling, flow modelling, geochemical modeling etc.
MINERAL RESOURCES	Highlights related to the investigation and exploitation of the type of mineral resource that has economic interest as a raw material.
FOSSIL RESOURCES	Highlights related to the research and exploitation of fossils (Coal and Hydrocarbons)
INFORMATION SYSTEM	Highlights related to the methods and uses for managing of geoscience information systems
LITHOLOGY	Highlights related to the part of the geology that deals with the study of rocks and the physical and chemical characteristics of the rocks that appear constituting a certain geological formation.
GEOLOGICAL HAZARD	Higthligths related to the process, situation or natural or induced event which can cause causing damage or loss of
HYDROGEOLOGY	Highlights related to the part of geology that studies the movement and distributions of surface and groundwater, as well as its research, prospecting, catchment and protection.
SPATIAL PLANNING/ENVIRONMENT	Highlights related to an interdisciplinary and global approach, which analyzes, develops and manages the processes of planning and development of geographic spaces and territories, both Urban and Rural, at a local, regional or national scale, according to their environmental, economic and social possibilities.
GEOCHEMISTRY	Highlights related to the specialty of natural sciences that, based on geology and chemistry, studies the composition and dynamics of the chemical elements in the earth. The following disciplines are considered: Geochemistry, Hydrogeochemistry, Lithogeochemistry, Organic geochemistry

Figure 2.1.1: First Version of the Search Categories

Search categories:	List of generic categories established through the thematic grouping of most of the GeoEra project deliverables included in the proposals. They represent the main headlines of the future GeoERA subject heading system (keyword thesaurus). The main objective is to have a common group of terms by topic for its application in the selective search of the GeoEra product catalogue. This GeoERA subject heading system will also facilitate the labelling of products in the GeoEra metadata catalogue.

SEARCH CATEGORIES	DESCRIPTION
GEOCHRONOLOGY and CHRONOSTRATIGRAPHY	Related to the determination of relative and absolute ages of rocks, fossils, sediments and time sequences of events in the earth's history. The following disciplines are considered: stratigraphy, geological history, geological time scale, Chronostratigraphy …
GEOTHERMAL RESOURCES	Principal topics related to resources, exploitability and capacity of geothermal energy.
GEOLOGICAL PROCESSES	Related to dynamic geological processes that act on land forms and surfaces and within the Earth. The following disciplines are considered: sedimentation, diagenesis, metamorphism, geomorphology ....
STRUCTURAL GEOLOGY	The study of the three-dimensional distribution of rock units, trying to reconstruct the movements and processes that have originated its structure, the history of movements and deformations on a global and regional scale. The following disciplines are considered: Tectonics, geologic structures.
APPLIED GEOPHYSICS	The use of geophysical techniques mainly for the exploration and exploitation of natural resources; but also applied to geological risks, environment and hydrogeology.
MODELLING	Reconstructing the geological environment and geological processes, using numerical, geostatistical and simulation techniques. The following disciplines are considered: 3D modelling, flow modelling, geochemical modelling, etc.
MINERAL RESOURCES	It focuses on the investigation and exploitation of the type of mineral resource that has economic value as a raw material.
FOSSIL RESOURCES	Research and exploitation of fossil fuels (Coal and Hydrocarbons)
INFORMATION SYSTEM	Methods and uses for managing of geoscience information systems.
LITHOLOGY	Part of the geology that deals with the study of rocks and the physical and chemical characteristics of the rocks that appear constituting a certain geological unit.
NATURAL HAZARD	Process, situation or natural or induced event which can cause damage or loss of property and life. The hazards of the following types are considered: seismicity, ground movements (e.g. Landslide, Debris flow, subsidence, etc.), Climate change, pollution,….
HYDROGEOLOGY	Part of geology that studies the movement and distributions of surface and groundwater, as well as its research, prospecting, catchment and protection.
SPATIAL PLANNING/ENVIRONMENT	Geoscientific contributions related to an interdisciplinary -approach, which analyzes, develops and manages the processes of planning and development of geographic spaces and territories, both Urban and Rural, at a local, regional or national scale, according to their environmental, economic and societal situation.
GEOCHEMISTRY	Uses of tools and principles of chemistry to study the composition and dynamics of the chemical elements in the earth. The following disciplines are considered: Geochemistry, Hydrogeochemistry, Lithogeochemistry, Organic geochemistry, etc.

Figure 1.1.2: Second Version of the Search Categories

A great feedback was obtained during and after the WP2 Meeting held in Brussels at the end of October 2018 (9 from the 14 project answered our request). A new Search Category was added to the 14 Search Categories initially proposed ("SUBSURFACE ENERGY STORAGE"). This new category was suggested jointly by MUSE and GeoConnect3d projects. Three of the proposed categories (GEOTHERMAL RESOURCES, NATURAL HAZARD, SPATIAL PLANNING/ENVIRONMENT) were modified in their title and description after compiling project feedback. Figure 2.1.3 below presents the definitive Search Categories. These Search Categories are used to classify the different vocabularies analyzed.

On the other hand, "extracted terms" were revised and completed by synonyms in most projects (Annex II). These terms will be used as test words in the evaluation process to determine whether a vocabulary belongs to a particular "Search Category" and whether or not it should really be chosen to be part of the future GeoERA Thesaurus.

SEARCH CATEGORIES	DESCRIPTION
GEOCHRONOLOGY/STRATIGRAPHY	Related to the determination of relative and absolute ages of rocks, fossils, sediments and time sequences of events in the earth's history. The following disciplines are considered: stratigraphy, geological history, geological time scale, Chronostratigraphy...
GEOTHERMAL ENERGY	Principal topics related to resources, conflicts and management of geothermal energy.
GEOLOGICAL PROCESSES	Related to dynamic geological processes that act on land forms and surfaces and within the Earth. The following disciplines are considered: sedimentation, diagenesis, metamorphism, geomorphology ....
STRUCTURAL GEOLOGY	The study of the three-dimensional distribution of rock units, trying to reconstruct the movements and processes that have originated its structure, the history of movements and deformations on a global and regional scale. The following disciplines are considered: Tectonics, geologic structures.
APPLIED GEOPHYSICS	The use of geophysical techniques mainly for the exploration and exploitation of natural resources; but also applied to geological risks, environment and hydrogeology.
MODELLING	Reconstructing the geological environment and geological processes, using numerical, geostatistical and simulation techniques. The following disciplines are considered: 3D modelling, flow modelling, geochemical modelling, etc.
MINERAL RESOURCES	It focuses on the investigation and exploitation of the type of mineral resource that has economic value as a raw material.
FOSSIL RESOURCES	Research and exploitation of fossil fuels (Coal and Hydrocarbons)
INFORMATION SYSTEM	Methods and uses for managing of geoscience information systems.
LITHOLOGY	Part of the geology that deals with the study of rocks and the physical and chemical characteristics of the rocks that appear constituting a certain geological unit.
HAZARD, RISK AND IMPACT	Processes, events of natural or induced origin, including surface and subsurface activities, that can cause damage or loss of property and life in the surface and subsurface. The hazards of the following types are considered: seismicity, ground movements (e.g. surface deformation, Landslide, etc.), leakage and migration and facility hazards, climate change, pollution,....
HYDROGEOLOGY	Part of geology that studies the movement and distributions of surface and groundwater, as well as its research, prospecting, catchment and protection.
SUBSURFACE MANAGEMENT	Geoscientific contributions related to an interdisciplinary approach, which analyses, develops and manages the processes of planning and development of the subsurface, according to their environmental, economic and societal situation.
GEOCHEMISTRY	Uses of tools and principles of chemistry to study the composition and dynamics of the chemical elements in the earth. The following disciplines are considered: Geochemistry, Hydrogeochemistry, Lithogeochemistry, Organic geochemistry, etc.
SUBSURFACE ENERGY STORAGE	Temporary subsurface storage of energy (mechanical and thermal energy) for the purpose of a later reuse. This topic includes research fields dealing with exploration, testing, managing and monitoring of subsurface storage. The term subsurface storage includes geological storage (e.g. aquifer, hydrocarbon reservoir) as well as engineered subsurface storage (e.g. cavern storage, borehole thermal energy storage).

Figure 2.1.3: Final Version of the Search Categories

2.1.3 Codelists and added vocabularies

In a first quick evaluation of the vocabularies obtained through the questionnaire it became clear that some of the Search categories were not covered by these vocabularies. On the other hand, only 9 of the 24 vocabularies were accessible through the web and were in English.
So it was decided to include in the evaluation the registered codelists that support the GeoSciML and EarthResourceML standards (http://resource.geosciml.org/def/voc/) and part of the registred codelists in INSPIRE (http://inspire.ec.europa.eu/codelist). At the same time, suggested vocabularies where added:

- eENVplus – LusTRE (Linked Thesaurus framework for Environment - http://linkeddata.ge.imati.cnr.it/terminologies_new.jsp)
- UMTHES - An environmental thesaurus with a lot “geoscientifical” terms (https://sns.uba.de/umthes/de/concepts/_00017172.html)
- wikidata reference identifiers (see section „Identifiers”) https://www.wikidata.org/wiki/Q7946

2.1.4 Grouped vocabularies. “Possible search category Thesauri”.

After an initial glance at the vocabularies we had to analyse, we realized that many of them (mostly from INSPIRE and GeoSciML) had to be grouped together in order to be evaluated jointly. This is because these vocabularies dealt with different topics but related to one Search Category.
Being aware of this, we created a new concept called "Possible search category Thesauri" that allowed us to assign several vocabularies to a possible thesaurus that would be analysed together quantitatively in relation to a search category, as long as the vocabularies previously and individually passed the excluding qualitative assessments, which will be explained in the next section.

2.1.5 Evaluation criteria

The evaluation criteria were designed to establish a vocabulary selection system so that we can objectively select thesauri for each search category defined to be included in the GeoERA keyword thesaurus.
To establish evaluation criteria and to carry out the evaluation is a complex objective, because the technicians who have to evaluate these vocabularies are not specialists in the GeoERA project topics. For this reason, an additional evaluation for project expert topics will be requested.

The evaluation consists of three parts:

- A qualitative evaluation that indicates the quality of the vocabulary. The result of this criterion will be a YES/NO
- A quantitative evaluation that indicates the extension and the depth of the possible search category thesauri;
- A qualitative evaluation, entirely subjective, that indicates the suitability of the vocabulary for a specific search category.

This last part should be done in two phases, an initial phase that relates the vocabulary to a search category and include it into a possible search category thesaurus. A later phase, in detail, that indicates if the possible search category thesauri are really adapted to the search category.

Evaluation steps should be carried out in a certain order, as there are steps that exclude the vocabulary, and there is no need for time-consuming subsequent evaluations.
Evaluation criteria with a brief description are listed below. It is also specified if a criterion is qualitative/quantitative and if it is exclusive.

1. **WEB (Qualitative,** exclusive**).**

Is there a website where we can check and evaluate the vocabulary?
If there is no web access, the vocabulary cannot be evaluated; in some cases, the vocabulary has been requested in RDF format or similar in order to be evaluated.

2. **References (Qualitative, not** exclusive**).**

Check whether the vocabulary keywords have external references that make them useful for linked data.

3. **Hierarchy (Qualitative, not** exclusive**).**

It indicates if the vocabulary has a hierarchy, mainly if there are broader and narrow terms.

4. **Multi-Languages (Qualitative,** exclusive**).**

Vocabulary is checked to determine whether it is available in more than one language. In order not to be excluded, it must be available in English at least; otherwise, the evaluation will not be possible.

5. **Accuracy of the Voc I (Qualitative,** exclusive**).**

A preliminary assessment is made to determine whether the vocabulary is suitable to be part of the GeoERA keyword thesaurus. If so, it will be assigned to one or more search categories and grouped

into a possible search category thesaurus, if needed. If it is not suitable, the vocabulary will be excluded from further evaluation.

6. **Number of terms (Quantitative, not** exclusive**).**

Number of terms. It is logical that the greater the number of terms, the more appropriate the vocabulary is.

7. **Number of levels (Quantitative, not** exclusive**).**

Number of levels. It is logical that the greater the number of levels, the more appropriate the vocabulary is.

8. **Is INSPIRE-CGI list (Qualitative, not** exclusive**).**

If a vocabulary is a codelist of INSPIRE or CGI it will be positively valued.

9. **Accuracy of the Voc II (phase 2 with test words - qualitative, not exclusive**).

This final phase of the evaluation is quite subjective and consists of assessing whether the search category thesauri that have passed the previous filters are adequate to the search category assigned to them. In order to carry out this evaluation, a list of keywords by search category was created. The degree of suitability will be higher if these representative keywords are included in the vocabulary.
This task is quite subjective, since the selection of representative keywords can vary depending on who made it, so a specialist in the specific topics should check the representative keywords selected for every search category.

The result of this assessment will give an ordered list of vocabularies according to the criteria. This ordered list allows us to define the GeoERA keyword thesaurus.

2.1.6 Database

Initially, we designed a spreadsheet as the easiest and simplest way to enter and review the information.
The evaluation of vocabularies was complicated because the suitability of a vocabulary was decided according to search categories and because we decided to create groups of vocabularies.
A vocabulary might be suitable for one search category and not for another or will have to be evaluated for more than one search category. In order to make the evaluation traceable, so that it can be reviewed, we decided to store all the results of every query and all the information in a structured way. In order to achieve this and facilitate the storage and editing of the information, we decided to create a database in Access. This database was filled in while the vocabularies were analysed.
The information we wanted to store in a structured way was:

- The relevant "extracted terms" from the deliverables of each GeoERA project grouped into the appropriate search categories
- The projects of GeoERA and their “extracted terms”
- All the vocabularies collected, those of the questionnaire and those that we decided to be added,
- Analyses performed that classifies each vocabulary in one or several search categories. These analyses include results such as number of keywords or levels suitable for each search category, as well as other queries.
- The results of the evaluation indicating the suitability of the vocabulary for the search category.
- Possible search category thesauri. We grouped some vocabularies into possible search category thesaurus, as we explained in section 1.4.

All these entities were included in a database with 8 tables:

✓	Domains (Search Categories)	15 records
✓	Domain- Terms	403 records
✓	Projects	14 records

✓	Projects- Terms	141 records
✓	Terms	153 records
✓	VocsCodelist	134 records
✓	VocsCodelist-Domains	335 records
✓	VocsCodelist-Domains_Evaluations	145 records

The name "Domains" was later changed to "Search Categories", although the access table keeps its original name.
The name "Terms" was later changed to "Extracted Terms", although the access table keeps its original name.

The data model is included in Annex III.

In order to exploit this information, we designed a series of queries to answer important questions in a simple and up-to-date way:

- Vocabularies to be analysed.
- Search categories without vocabularies. This is a problem because each search category should have keywords for tagging or searching.
- Vocabularies not assigned to a search category. They are vocabularies that, in our opinion, are not interesting for any search category.
- Number of vocabularies assigned to each search category. The more vocabularies assigned to a search category, the easier it is to find one, which is suitable for inclusion in the GeoERA keyword thesaurus.
- Number of keywords and Number of levels. These queries can be made by vocabulary or by search category.

The results of these queries either will be included as annexes or will be included in the following section of the evaluation result.

2.2 Evaluation results

2.2.1 Vocabularies analyzed.

93 Vocs/Codelist analyzed.
52 INSPIRE codelist.
14 GeoSciML codelist.
23 Vocabularies from questionnaire.
145 Vocabularies-Search categories analyzed.

According with the excluding criteria the vocabularies analyzed are:

Vocabularies that may be analyzed			
Title	Has WEB	Language	Will be analyzed
Decorative stones	NO	YES	NO
Applied geophysics - methods	NO	NO	NO
Slope instabilities	NO	YES	NO
Geological encyclopaedia on-line	YES	NO	NO
GEMET	YES	YES	YES
GBA Thesaurus Lithology	YES	YES	YES

GeoSciML Simple Lithology	YES	YES	YES
INSPIRE codelist lithology	YES	YES	YES
GBA Thesaurus Geologic Timescale	YES	YES	YES
GBA Thesaurus Rohstoffgeologie (Raw Material)	YES	YES	YES
CGI - Geologic Time Vocabulary - International Chronostratigraphic Chart - 2017	YES	NO	NO
Tesauro IGME de ciencias de la tierra IGME	YES	YES	YES
INSPIRE Codelist GeophPropertyNameValue	YES	YES	YES
INSPIRE Codelist GeophProcessNameValue	YES	YES	YES
INSPIRE Codelist GeophProcessParameterNameValue	YES	YES	YES
INSPIRE Codelist ResourceTypeValue	YES	NO	NO
Minerals4EU Metadata Keywords	YES	YES	YES
OneGeology-Europe keywords database	YES	YES	YES
Multilingual geological thesaurus (eWater, eEarth)	NO	YES	NO
Geological codelists (Czech-English) - chronostratigraphy, lithostratigraphy, regional geology, lithology	YES	YES	YES
Mineralogy - heavy minerals keywords (Czech)	NO	YES	NO
English-Czech and Czech-English Professional Dictionary of Hydrogeology Hydrogeologický slovník (Aj, Čj)	NO	YES	NO
Geologický slovník (Aj, Čj) Dictionary of Geology English-Czech and Czech-English	YES	YES	YES
Catalogue of geohazards	YES	NO	NO
EARTh.	YES	NO	NO
UMTHES Umweltthesaurus of the German Umweltbundesamt	YES	YES	YES
INSPIRE Codelist BoreholePurpose	YES	YES	YES
INSPIRE Codelist EventProcessValue	YES	YES	YES
INSPIRE Codelist EventEnvironmentValue	YES	YES	YES
INSPIRE Codelist FaultTypeValue	YES	YES	YES
INSPIRE Codelist ExplorationActivityTypeValue	YES	YES	YES
INSPIRE Codelist EndusePotentialValue	YES	YES	YES
INSPIRE Codelist ExplorationResultValue	YES	YES	YES
INSPIRE Codelist NaturalGeomorphologicFeatureTypeValue	YES	YES	YES
INSPIRE Codelist CurveModelTypeValue	YES	YES	YES
INSPIRE Codelist SwathTypeValue	YES	YES	YES
INSPIRE Codelist SurveyTypeValue	YES	YES	YES
INSPIRE Codelist StationTypeValue	YES	YES	YES
INSPIRE Codelist ProfileTypeValue	YES	YES	YES
INSPIRE Codelist PlatformTypeValue	YES	YES	YES
INSPIRE Codelist CampaignTypeValue	YES	YES	YES
INSPIRE Codelist WaterSalinityValue	YES	YES	YES
INSPIRE Codelist NaturalObjectTypeValue	YES	YES	YES
INSPIRE Codelist HydroGeochemicalRockTypeValue	YES	YES	YES
INSPIRE Codelist AquiferTypeValue	YES	YES	YES
INSPIRE Codelist AquiferMediaTypeValue	YES	YES	YES
INSPIRE Codelist ActiveWellTypeValue	YES	YES	YES
INSPIRE Codelist FoldProfileTypeValue	YES	YES	YES
INSPIRE Codelist ReserveCategoryValue	YES	YES	YES

INSPIRE Codelist ProcessingActivityTypeValue	YES	YES	YES
INSPIRE Codelist MiningActivityTypeValue	YES	YES	YES
INSPIRE Codelist MineStatusValue	YES	YES	YES
INSPIRE Codelist MineralOccurrenceTypeValue	YES	YES	YES
INSPIRE Codelist MineralDepositTypeValue	YES	YES	YES
INSPIRE Codelist MineralDepositGroupValue	YES	YES	YES
INSPIRE Codelist ConditionOfGroundwaterValue	YES	YES	YES
INSPIRE Codelist StatusCodeTypeValue	YES	YES	YES
INSPIRE Codelist WaterPersistenceValue	YES	YES	YES
INSPIRE Codelist GeologicUnitTypeValue	YES	YES	YES
INSPIRE Codelist MediaValue	YES	YES	YES
INSPIRE Codelist Measurement Regime Value Definition: MeasurementRegimeValue	YES	YES	YES
INSPIRE Codelist HILUCSValue	YES	YES	YES
INSPIRE Codelist LevelOfSpatialPlanValue	YES	YES	YES
INSPIRE Codelist LayerTypeValue	YES	YES	YES
INSPIRE Codelist ProfileElementParameterNameValue	YES	YES	YES
INSPIRE Codelist OtherContaminatingActivityValue	YES	YES	YES
INSPIRE Codelist SupplementaryRegulationValue	YES	YES	YES
INSPIRE Codelist NaturalHazardCategoryValue	YES	YES	YES
INSPIRE Codelist ExposedElementCategoryValue	YES	YES	YES
INSPIRE Codelist RiskAssessmentStageValue	YES	YES	YES
INSPIRE Codelist RiskReceptorValue	YES	YES	YES
INSPIRE Codelist RiskTypeValue	YES	YES	YES
INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	YES	YES	YES
INSPIRE Codelist MappingFrameValue	YES	YES	YES
INSPIRE Codelist AnthropogenicGeomorphologicFeatureTypeValue	YES	YES	YES
INSPIRE Codelist ClassificationAndQuantificationFrameworkValue	YES	YES	YES
INSPIRE Codelist FossilFuelClassValue	YES	YES	YES
INSPIRE Codelist FossilFuelValue	YES	YES	YES
INSPIRE Codelist RenewableAndWasteValue	YES	YES	YES
INSPIRE Codelist SoilPlotTypeValue	YES	YES	YES
GeoSciML _ alteration_Type	YES	YES	YES
GeoSciML _ BoreholeDrillingMethod	YES	YES	YES
EarthReML_CommodityCode	YES	YES	YES
GeoSciML_CompositionCategory	YES	YES	YES
GeoSciML_Compound MatrialConstituentPart	YES	YES	YES
GeoSciML_ConsolidationDegree	YES	YES	YES
GeoSciML_ContactType	YES	YES	YES
GeoSciML_ConventionCodeStrikeDipMeasurements	YES	YES	YES
GeoSciML_DeformationStyle	YES	YES	YES
GeoSciML_DescriptionPurpose	YES	YES	YES
GeoSciML_EventEnviroment	YES	YES	YES
GeoSciML_EventProcess	YES	YES	YES
GeoSciML_FaultMovementSense	YES	YES	YES

GeoSciML_FaultMovementType	YES	YES	YES
GeoSciML_FaultType	YES	YES	YES
GeoSciML_FolationType	YES	YES	YES
GeoSciML_GeneticCategory	YES	YES	YES
GeoSciML_GeologicUnitMorphology	YES	YES	YES
GeoSciML_GeologicUnitPartRole	YES	YES	YES
GeoSciML_GeologicUnitType	YES	YES	YES
GeoSciML_LineationType	YES	YES	YES
GeoSciML_MappingFrame	YES	YES	YES
GeoSciML_MetamorphicFacies	YES	YES	YES
GeoSciML_MetamorphicGrade	YES	YES	YES
GeoSciML_ObservationMethodMappedFeature	YES	YES	YES
GeoSciML_OrientationDeterminationMethod	YES	YES	YES
GeoSciML_ParticleAspectRatio	YES	YES	YES
GeoSciML_ParticleShape	YES	YES	YES
GeoSciML_ParticleType	YES	YES	YES
GeoSciML_PlanarPolarityCode	YES	YES	YES
GeoSciML_ProportionTerm	YES	YES	YES
GeoSciML_SimpleLithology	YES	YES	YES
GeoSciML_StratigraphicRank	YES	YES	YES
GeoSciML_ValueQualifier	YES	YES	YES
EarthReML_EarthResourceExpression	YES	YES	YES
EarthReML_EarthResourceForm	YES	YES	YES
EarthReML_EarthResourceMaterialRole	YES	YES	YES
EarthReML_EarthResourceShape	YES	YES	YES
EarthReML_EndUsePotential	YES	YES	YES
EarthReML_EnvironmentalImpact	YES	YES	YES
EarthReML_ExplorationActivityType	YES	YES	YES
EarthReML_ExplorationResult	YES	YES	YES
EarthReML_MineStatus	YES	YES	YES
EarthReML_MineralOccurrenceType	YES	YES	YES
EarthReML_MiningActivity	YES	YES	YES
EarthReML_ProcessingActivity	YES	YES	YES
EarthReML_RawMaterialRole	YES	YES	YES
EarthReML_UNFCCode	YES	YES	YES
EarthReML_WasteStorage	YES	YES	YES
EarthReML_ReportingClassificationMethod	YES	YES	YES
EarthReML_ReserveAssessmentCategory	YES	YES	YES
EarthReML_ResourceAssessmentCategory	YES	YES	YES
INSPIRE CodelistCommodityValue	YES	YES	YES
INSPIRE Codelist LithologyValue	YES	YES	YES

The number of Vocabularies that should be analyzed is 134, from these only 93 are suitable with search categories.

2.2.2 *Search categories thesauri*

As indicated in section 1.4, for a better evaluation of the thesauri/codelists collected it has been decided to gather them together, in groups, within the different Search Categories.
In the following sections for each search category the couple search category-thesaurus evaluated are presented through a table and a bar graph:

- the table specifies which are the collected thesauri/codelists that belong to a Search category thesaurus couple
- the bar graph represents the numbers of levels and terms for each search category-thesaurus

There is no section devoted to MODELLING, INFORMATION SYSTEM and SUBSURFACE ENERGY STORAGE as no codelists/thesauri has been found that can be classified in these three categories

2.2.2.1 Applied Geophysics

The Applied Geophysics Category, brings together a total of 9 vocabularies, which have been collected in 4 groups as shown in the Figure 2.2.1 and in Figure 2.2.2.

From the analysis and review of both, in this case, it is easily deduced that the Group called INSPIRE_APPLIED_GEOPHYSIC, may be the most complete option in this matter, due to its greater number of terms and levels. Without ruling out that, ONEGE can be a second good option.

Group	CodeList	Codelist theme or model
EARTh_APPLIED GEOPHYSIC	EARTh.	Earth
GeoSciML_APPLIED_GEOPHYSIC	GeoSciML_GeologicUnitType	GeoSciML
INSPIRE_APPLIED_GEOPHYSIC	INSPIRE Codelist BoreholePurpose	GEOLOGY
	INSPIRE Codelist CurveModelTypeValue	GEOLOGY
	INSPIRE Codelist ProfileTypeValue	GEOLOGY
	INSPIRE Codelist StationTypeValue	GEOLOGY
	INSPIRE Codelist SurveyTypeValue	GEOLOGY
	INSPIRE Codelist SwathTypeValue	GEOLOGY
ONEGE_APPLIED_GEOPHYSIC	OneGeology-Europe keywords database	

Figure 2.2.1: Search Category-thesaurus for Applied Geophysics

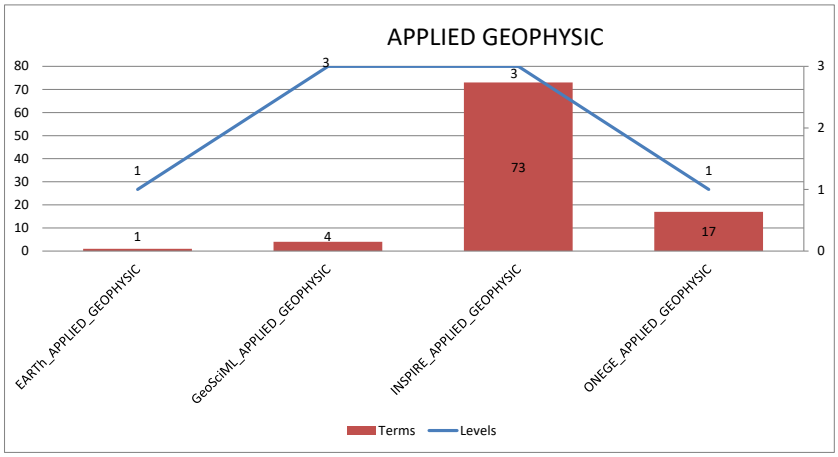

Figure 2.2.2: Number of levels and terms by search category-thesaurus for Applied Geophysics

2.2.2.2 Fossil Resources

Applied Fossil Category, brings together a total of 7 thesauri, which have been collected in 4 Groups as shown in Figure 3 and Figure 2.2.4.

In this category, it is necessary to evaluate the four search category-thesauri with the test keywords selected within the Annex II.

Group	CodeList	Codelist theme or model
EarthReML_FOSSIL_RESSOURCE	EarthReML_CommodityCode	
INSPIRE_FOSSIL_RESSOURCE_1	INSPIRE Codelist BoreholePurpose	GEOLOGY
INSPIRE_FOSSIL_RESSOURCE_2	INSPIRE Codelist ClassificationAndQuantificationFrameworkValue	ENERGY RESOURCE
	INSPIRE Codelist FossilFuelClassValue	ENERGY RESOURCE
	INSPIRE Codelist FossilFuelValue	ENERGY RESOURCE
INSPIRE_FOSSIL_RESSOURCE_3	INSPIRE Codelist OtherContaminatingActivityValue	SOIL
	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	SOIL

Figure 2.2.3: Search Category-thesaurus for Fossil Resources

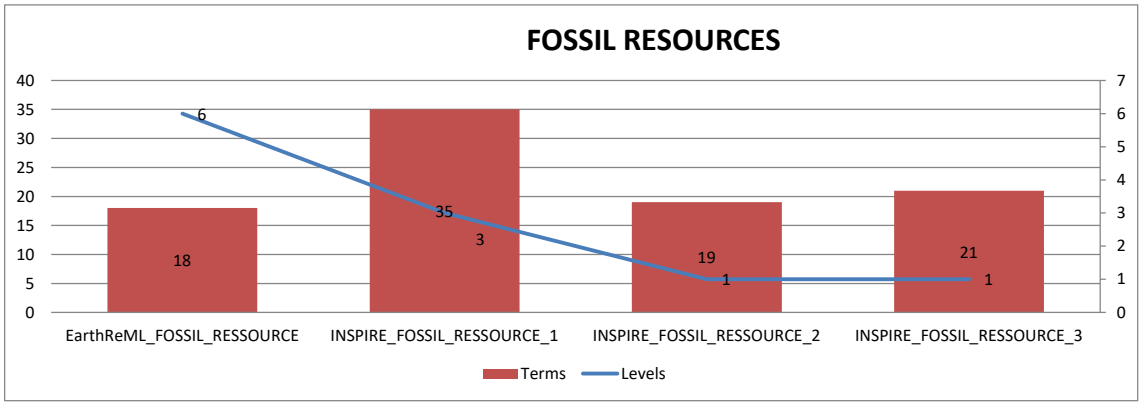

Figure 2.2.4: Number of levels and terms by search category-thesaurus for Fossil Resources

2.2.2.3 Geochemistry

Geochemistry Category, brings together a total of 6 vocabularies, which have been collected in 3 Groups as shown in Figure 2.2.5 and Figure 2.2.6.

From the analysis and review of both, in this case it is easily deduced that the Group called INSPIRE_GEOCHEMISTRY1 may be the most complete option in this matter, due to its greater number of terms and levels. The other two groups are included in the final selection in order of a wider range of geochemical terms.

Group	CodeList	Codelist theme or model
INSPIRE_GEOCHEMISTRY_1	INSPIRE Codelist EventProcessValue	GEOLOGY
	INSPIRE Codelist HydroGeochemicalRockTypeValue	GEOLOGY
	INSPIRE Codelist NaturalGeomorphologicFeatureTypeValue	GEOLOGY
	INSPIRE Codelist WaterSalinityValue	GEOLOGY
INSPIRE_GEOCHEMISTRY_2	INSPIRE Codelist ProcessingActivityTypeValue	MINERAL RESOURCE
INSPIRE_GEOCHEMISTRY_3	INSPIRE Codelist ProfileElementParameterNameValue	SOIL

Figure 2.2.5: Search Category-thesaurus for Geochemisty

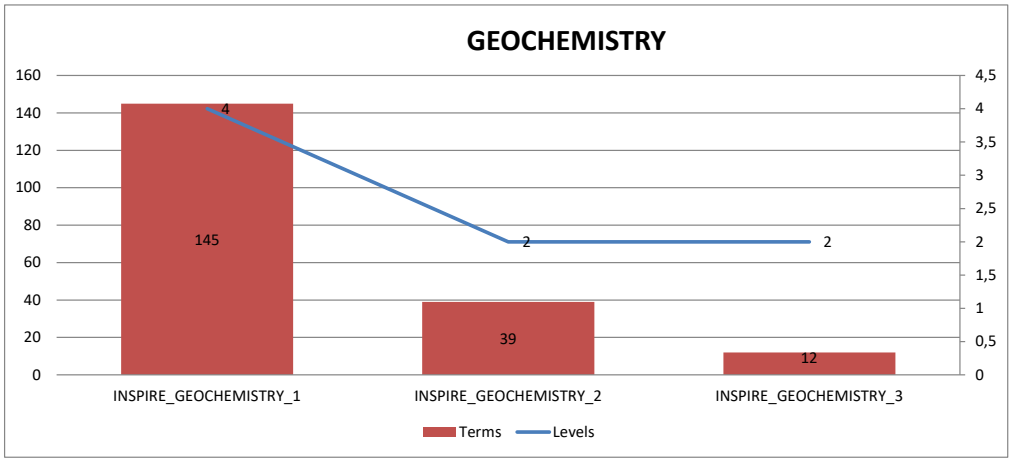

Figure 2.2.6: Number of levels and terms by search category-thesaurus for Geochemistry

2.2.2.4 Geochronology Stratigraphy

Geochronology Stratigraphy Category, collects 3 vocabularies, gathered in two groups (Figure 2.2.7 and Figure 2.2.8).

Although in both groups the number of levels is the same, the number of terms bring us to select GeoSciML_GEOCHRONO_STRATIGRAPHY_1, which exclusively contains the Codelist CGI - Geologic Time Vocabulary - International Chronostratigraphic Chart - 2017 (which is also a world reference).

Group	CodeList	Codelist theme or model
GeoSciML_GEOCHRONO_STRATIGRAPHY_1	CGI - Geologic Time Vocabulary - International Chronostratigraphic Chart - 2017	GeoSciML
GeoSciML_GEOCHRONO_STRATIGRAPHY_2	GeoSciML_GeologicUnitPartRole	GeoSciML
	GeoSciML_GeologicUnitType	GeoSciML

Figure 2.2.7: Search Category-thesaurus for Geochronology/Stratigraphy

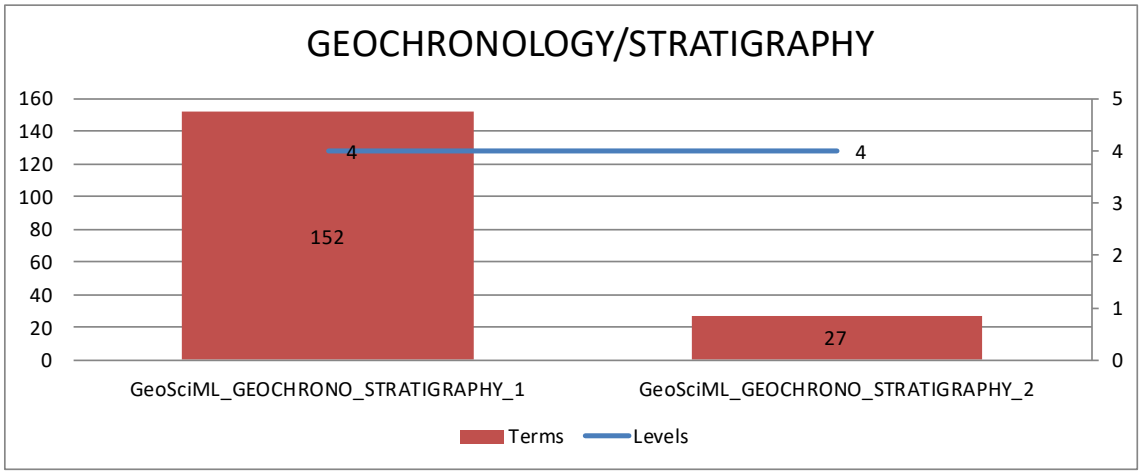

Figure 2.2.8: Number of levels and terms by search category-thesaurus for Geochronology/Stratigraphy

2.2.2.5 Geological Processes

Geological Processes Category gathers its 12 vocabularies, in three groups, as can be seen in Figure 2.2.9 and Figure 2.2.10.

On this case, both the thesaurus GeoSciML_GEOLOGICAL_PROCESSES and INSPIRE_GEOLOGICAL_PROCESSES must be considered within this search category.

Group	CodeList	Codelist theme or model
GeoSciML_GEOLOGICAL_PROCESSES	GeoSciML_DeformationStyle	GeoSciML
	GeoSciML_EventEnviroment	GeoSciML
	GeoSciML_EventProcess	GeoSciML
	GeoSciML_GeneticCategory	GeoSciML
	GeoSciML_GeologicUnitType	GeoSciML
	GeoSciML_MetamorphicGrade	GeoSciML
INSPIRE_GEOLOGICAL_PROCESSES	INSPIRE Codelist AnthropogenicGeomorphologicFeatureTypeValue	GEOLOGY
	INSPIRE Codelist EventEnvironmentValue	GEOLOGY
	INSPIRE Codelist EventProcessValue	GEOLOGY
	INSPIRE Codelist GeologicUnitTypeValue	GEOLOGY
	INSPIRE Codelist NaturalGeomorphologicFeatureTypeValue	GEOLOGY
ONEGE_GEOLOGICAL_PROCESSES	OneGeology-Europe keywords database	

Figure 2.2.9: Search Category-thesaurus for Geological Processes

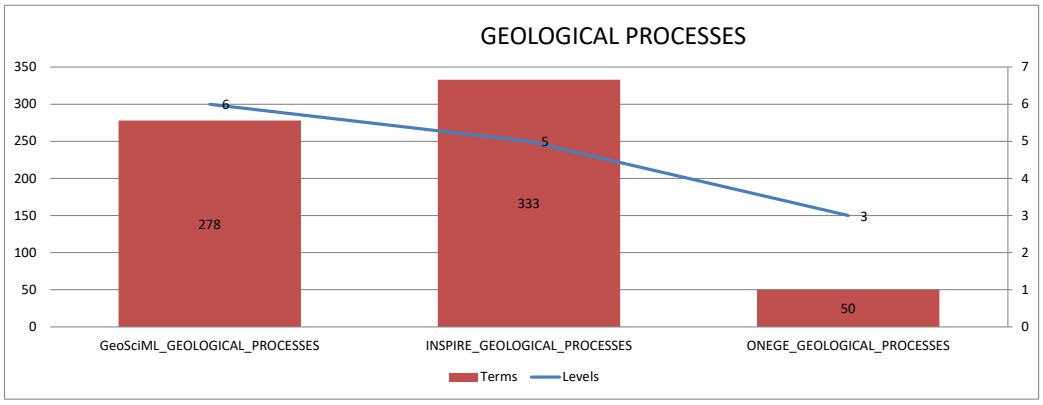

Figure 2.2.10: Number of levels and terms by search category-thesaurus for Geological Processes

2.2.2.6 Geothermal Energy

Geothermal Energy Category is only represented by the group INSPIRE_GEOTHERMAL_ENERGY, which collects the vocabulary INSPIRE Codelist ActiveWellTypeValue, since some of its terms refer to this topic.

Group	CodeList	Codelist theme or model
INSPIRE_GEOTHERMAL_ENERGY	INSPIRE Codelist ActiveWellTypeValue	GEOLOGY

Figure 2.2.11: Search Category-thesaurus for Geothermal Energy

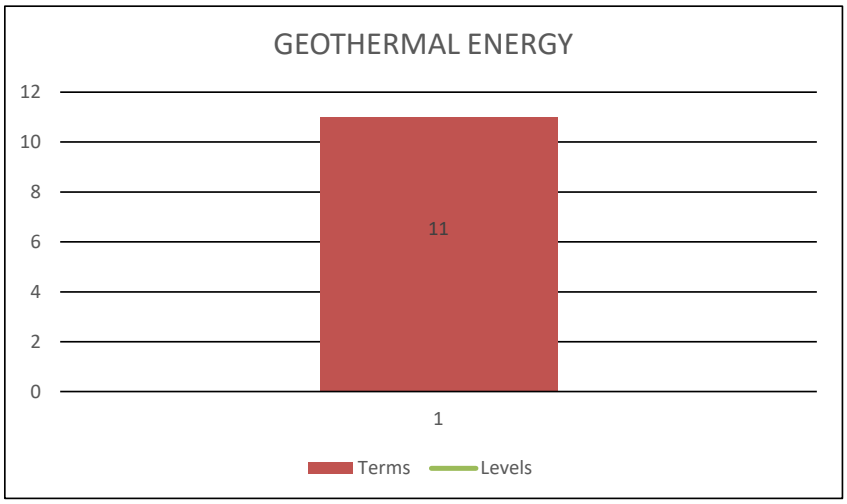

Figure 2.2.12: Number of levels and terms by search category-thesaurus for Geothermal Energy

2.2.2.7 Hazard, Risk and Impact

Hazard, Risk and Impact Category, is represented by 4 groups that collect 6 thesauri (see Figure 2.2.13 and Figure 2.2.14).

In this case, initially, the three groups with the highest number of terms EarthReML_HAZARD_RI, INSPIRE_HAZARD_RI_1 and INSPIRE_HAZARD_RI_2 are recommended.

Group	CodeList	Codelist theme or model
EarthReML_HAZARD_RI	EarthReML_EnvironmentalImpact	EarthResourceML
INSPIRE_HAZARD_RI_1	INSPIRE Codelist AnthropogenicGeomorphologicFeatureTypeValue	GEOLOGY
	INSPIRE Codelist NaturalGeomorphologicFeatureTypeValue	GEOLOGY
INSPIRE_HAZARD_RI_2	INSPIRE Codelist ExposedElementCategoryValue	NATURAL RISK ZONE
	INSPIRE Codelist NaturalHazardCategoryValue	NATURAL RISK ZONE
ONEGE_HAZARD_RI	OneGeology-Europe keywords database	

Figure 2.2.13: Search Category-thesaurus for Hazard, Risk and Impact

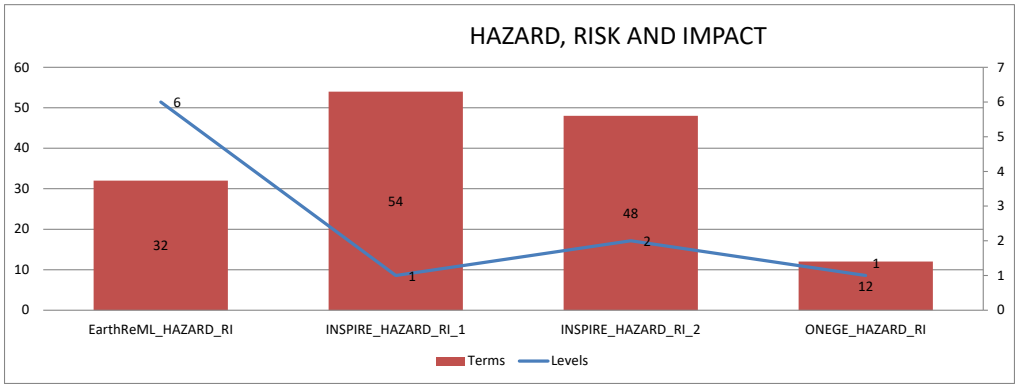

Figure 2.2.14: Number of levels and terms by search category-thesaurus for Hazard, Risk and Impact

2.2.2.8 Hydrogeology

Hydrogeology Category, is represented by 14 thesauri gathered in 5 groups.

Although the group INSPIRE_HYDROGEOLOGY_1 has only one level, it is the group that has the largest number of terms, and belongs to the Hydrology schema included in the INSPIRE Geology theme.

On the other hand, at least the groups INSPIRE_HYDROGEOLOGY_2 and INSPIRE_HYDROGEOLOGY_3 must be also considered, due to the relationship of a large part of their terms with hydrogeology: hydrogeological purpose of soundings and risk of groundwater pollution respectively.

Group	CodeList	Codelist theme or model
INSPIRE_HYDROGEOLOGY_1	INSPIRE Codelist ActiveWellTypeValue	GEOLOGY
	INSPIRE Codelist AquiferMediaTypeValue	GEOLOGY
	INSPIRE Codelist AquiferTypeValue	GEOLOGY
	INSPIRE Codelist ConditionOfGroundwaterValue	GEOLOGY
	INSPIRE Codelist HydroGeochemicalRockTypeValue	GEOLOGY
	INSPIRE Codelist NaturalObjectTypeValue	GEOLOGY
	INSPIRE Codelist StatusCodeTypeValue	GEOLOGY
	INSPIRE Codelist WaterPersistenceValue	GEOLOGY
	INSPIRE Codelist WaterSalinityValue	GEOLOGY
INSPIRE_HYDROGEOLOGY_2	INSPIRE Codelist BoreholePurpose	GEOLOGY
INSPIRE_HYDROGEOLOGY_3	INSPIRE Codelist OtherContaminatingActivityValue	SOIL
	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	SOIL
INSPIRE_HYDROGEOLOGY_4	INSPIRE Codelist RiskTypeValue	NATURAL RISK ZONE
ONEGE_HYDROGEOLOGY	OneGeology-Europe keywords database	

Figure 2.2.14: Search Category-thesaurus for Hydrology

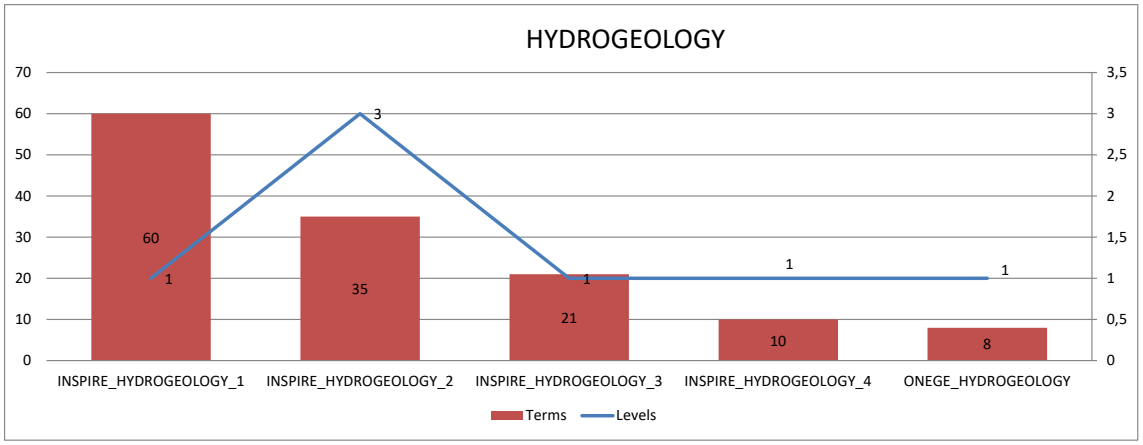

Figure 2.2.15: Number of levels and terms by search category-thesaurus for Hydrology

2.2.2.9 Lithology

Lithology Category, brings together a total of 10 thesauri, which have been collected in 6 Groups as shown in Figure 2.2.16 and Figure 2.2.17.

In this case, all the groups seem to have considerable relevance in this Search Category, in general with numerous levels and terms in each group.

GBA Thesaurus Lithology is outstanding by number of terms, followed by GeoSciML_LITHOLOGY_1 and INSPIRE_LITHOLOGY_1 that have the same numbers of terms. In third position will be EARTH_LITHOLOGY, even if it has slightly more terms.

Group	CodeList	Codelist theme or model
EARTH_LITHOLOGY	EARTh.	EARTh.
GBA_LITHOLOGY	GBA Thesaurus Lithology	GBA
GeoSciML_LITHOLOGY_1	GeoSciML_SimpleLithology	GeoSciML
GeoSciML_LITHOLOGY_2	GeoSciML_CompositionCategory	GeoSciML
	GeoSciML_GeologicUnitType	GeoSciML
	GeoSciML_MetamorphicFacies	GeoSciML
INSPIRE_LITHOLOGY_1	INSPIRE Codelist LithologyValue	GEOLOGY
INSPIRE_LITHOLOGY_2	INSPIRE Codelist EventEnvironmentValue	GEOLOGY
	INSPIRE Codelist EventProcessValue	GEOLOGY
	INSPIRE Codelist HydroGeochemicalRockTypeValue	GEOLOGY

Figure 2.2.16: Search Category-thesaurus for Lithology

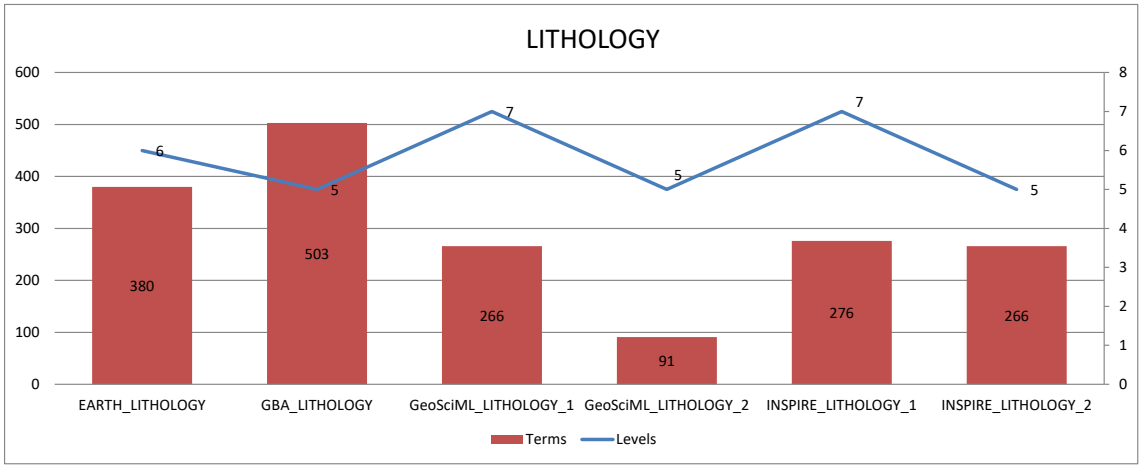

Figure 2.2.17: Number of levels and terms by search category-thesaurus for Lithology

2.2.2.10 Mineral Resources

Mineral Resources Category is the search category that gathers more vocabularies. A total of 29 vocabularies are collected in 6 groups (**Fehler! Verweisquelle konnte nicht gefunden werden.**.18 and **Fehler! Verweisquelle konnte nicht gefunden werden.**).

The group GBA_MINERAL_RESOURCE, which includes the vocabulary GBA Thesaurus Rohstoffgeologie (Raw Material), stands out above all others.

It is advisable to include, INSPIRE_MINERAL_RESOURCE_1 and EarthReML_MINERAL_RESOURCE group due to the number of terms and level they have. Minerals4EU_MINERAL_RESOURCE group, that includes M4EU Metadata Keywords, should also be included as it is one the result of M4EU project having continuity in the GeoERA project Mintel4EU.

Group	CodeList	Codelist theme or model
EARTH_MINERAL_RESOURCE	EARTh.	EARTh.
EarthReML_MINERAL_RESOURCE	EarthReML_EarthResourceMaterialRole	EarthResourceML
	EarthReML_EndUsePotential	EarthResourceML
	EarthReML_ExplorationResult	EarthResourceML
	EarthReML_MineralOccurrenceType	EarthResourceML
	EarthReML_MineStatus	EarthResourceML
	EarthReML_MiningActivity	EarthResourceML
	EarthReML_ProcessingActivity	EarthResourceML
	EarthReML_RawMaterialRole	EarthResourceML
	EarthReML_ReportingClassificationMethod	EarthResourceML
	EarthReML_ReserveAssessmentCategory	EarthResourceML
	EarthReML_CommodityCode	EarthResourceML
INSPIRE_MINERAL_RESOURCE_1	INSPIRE Codelist EndusePotentialValue	MINERAL RESOURCE
	INSPIRE Codelist ExplorationActivityTypeValue	MINERAL RESOURCE
	INSPIRE Codelist ExplorationResultValue	MINERAL RESOURCE
	INSPIRE Codelist MineralDepositGroupValue	MINERAL RESOURCE
	INSPIRE Codelist MineralDepositTypeValue	MINERAL RESOURCE
	INSPIRE Codelist MineralOccurrenceTypeValue	MINERAL RESOURCE
	INSPIRE Codelist MineStatusValue	MINERAL RESOURCE
	INSPIRE Codelist MiningActivityTypeValue	MINERAL RESOURCE
	INSPIRE Codelist ProcessingActivityTypeValue	MINERAL RESOURCE
	INSPIRE CodelistCommodityValue	MINERAL RESOURCE
	INSPIRE Codelist ReserveCategoryValue	MINERAL RESOURCE
INSPIRE_MINERAL_RESOURCE_2	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	SOIL
	INSPIRE Codelist OtherContaminatingActivityValue	SOIL
	INSPIRE Codelist BoreholePurpose	GEOLOGY
GEMET_MINERAL_RESOURCE	GEMET	GEMET
GBA_MINERAL_RESOURCE	GBA Thesaurus Rohstoffgeologie (Raw Material)	GBA
Minerals4EU_MINERAL_RESOURCE	Minerals4EU Metadata Keywords	Minerals4EU

Figure 2.2.18: Search Category-thesaurus for Mineral Resources

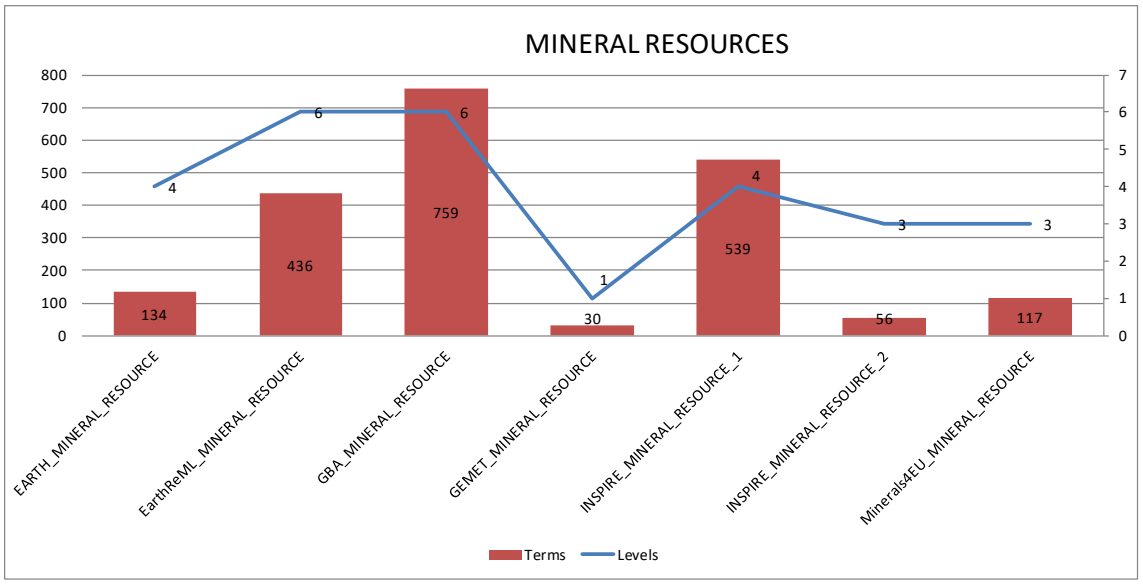

Figure 2.2.19: Number of levels and terms by search category-thesaurus for Mineral Resources

2.2.2.11 Structural Geology

Structural Geology Category gathers 10 thesauri in 3 groups, as can be seen in **Fehler! Verweisquelle konnte nicht gefunden werden.** and **Fehler! Verweisquelle konnte nicht gefunden werden.**.

The INSPIRE_STRUCTURAL group is the one with the most levels and terms. However, the EARTH_STRUCTURAL and GeoSciML_STRUCTURAL group should also be included.

Group	CodeList	Codelist theme or model
EARTH_STRUCTURAL	EARTh.	EARTh.
GeoSciML_STRUCTURAL	GeoSciML_ContactType	GeoSciML
	GeoSciML_DeformationStyle	GeoSciML
	GeoSciML_FaultMovementSense	GeoSciML
	GeoSciML_FaultType	GeoSciML
	GeoSciML_GeologicUnitType	GeoSciML
INSPIRE_STRUCTURAL	INSPIRE Codelist EventEnvironmentValue	GEOLOGY
	INSPIRE Codelist EventProcessValue	GEOLOGY
	INSPIRE Codelist FaultTypeValue	GEOLOGY
	INSPIRE Codelist FoldProfileTypeValue	GEOLOGY

Figure 2.2.20: Search Category-thesaurus for Structural Geology

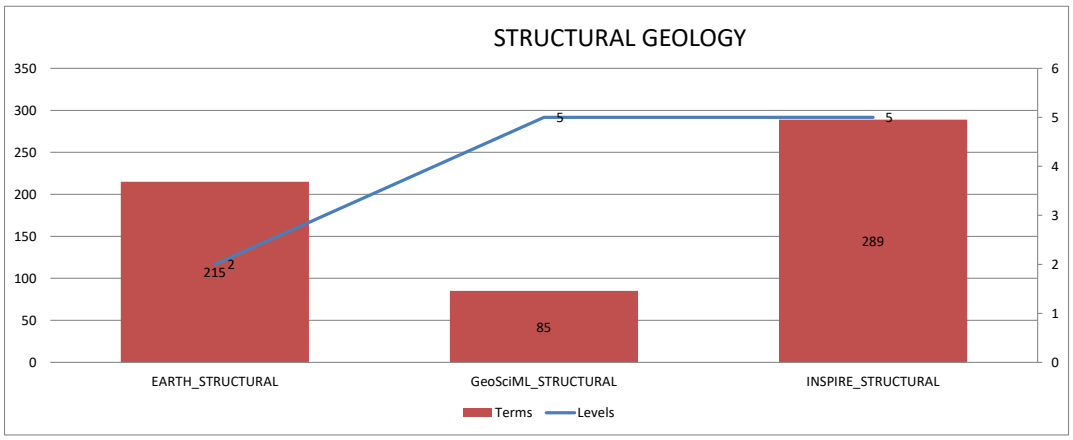

Figure 2.2.21: Number of levels and terms by search category-thesaurus for Structural Geology

2.2.2.12 Subsurface Management

Subsurface Management Category brings together 17 vocabularies in 8 groups, (Figure 2.2.22 and Figure 2.2.23).

In principle, the INSPIRE_SUBSURFACE_MANAGEMENT_2 group stands out from the rest because of its levels and terms. However, the following groups should be evaluated through the test words in a second phase to discriminate if they have to be included:

- EarthReML_SUBSURFACE_MANAGEMENT
- INSPIRE_SUBSURFACE_MANAGEMENT_5
- ONEGE_SUBSURFACE_MANAGEMENT

The CATALOG_GEOHAZARD_CZECH_SUBSURFACE_MANAGEMENT group is discarded as this thesaurus is only in Czech.

Group	CodeList	Codelist theme or model
EarthReML_SUBSURFACE_MANAGEMENT	EarthReML_ExplorationActivityType	EarthResourceML
	EarthReML_ReserveAssessmentCategory	EarthResourceML
INSPIRE_SUBSURFACE_MANAGEMENT_1	INSPIRE Codelist AnthropogenicGeomorphologicFeatureTypeValue	GEOLOGY
INSPIRE_SUBSURFACE_MANAGEMENT_2	INSPIRE Codelist HILUCSValue	LAND USE
	INSPIRE Codelist SupplementaryRegulationValue	LAND USE
	INSPIRE Codelist LevelOfSpatialPlanValue	LAND USE
INSPIRE_SUBSURFACE_MANAGEMENT_3	INSPIRE Codelist Measurement Regime Value	ENVIRONMENTAL MONITORING FACILITIES
	INSPIRE Codelist MediaValue	ENVIRONMENTAL MONITORING FACILITIES
INSPIRE_SUBSURFACE_MANAGEMENT_4	INSPIRE Codelist ReserveCategoryValue	MINERAL RESOURCES
INSPIRE_SUBSURFACE_MANAGEMENT_5	INSPIRE Codelist RiskAssessmentStageValue	SOIL
	INSPIRE Codelist RiskReceptorValue	SOIL
	INSPIRE Codelist RiskTypeValue	SOIL
	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	SOIL
	INSPIRE Codelist LayerTypeValue	SOIL
	INSPIRE Codelist OtherContaminatingActivityValue	SOIL
CATALOG_GEOHAZARD_CZECH_SUBSURFACE_MANAGEME	Catalogue of geohazards	
ONEGE_SUBSURFACE_MANAGEMENT	OneGeology-Europe keywords database	

Figure 2.2.22: Search Category-thesaurus for Subsurface Management

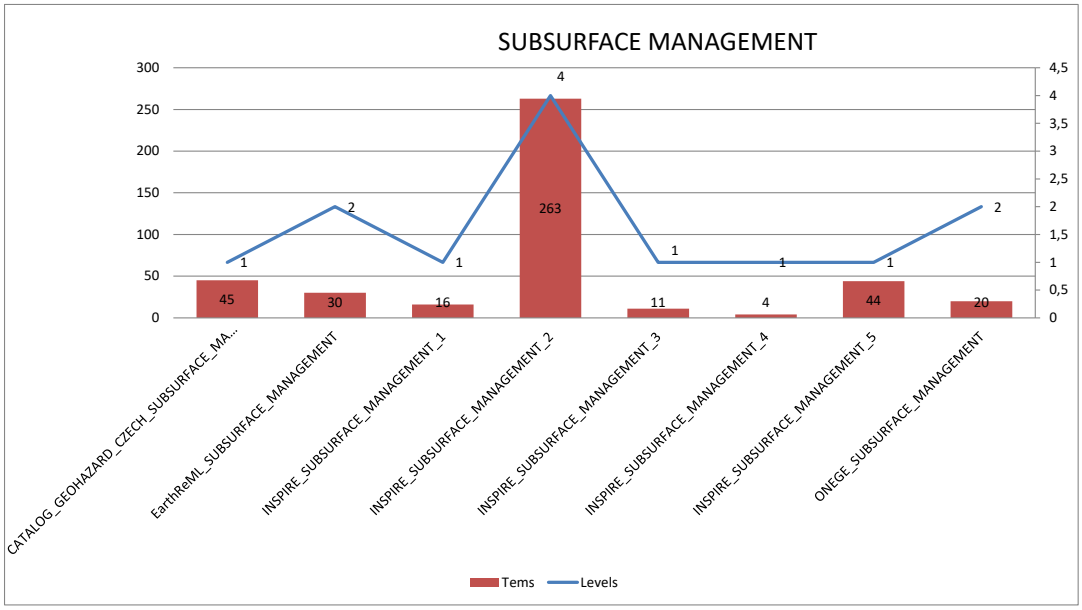

Figure 2.2.32: Number of levels and terms by search category-thesaurus for Subsurface Management

2.3 Conclusions

In the previous paragraph the selected codelists/thesauri were presented. In some of the search categories, a second evaluation, through test keywords extracted from Annex II, was necessary. In this section the results of this second evaluation are presented, as well as the final selection.

2.3.1 Test keywords

Search Category	Test Keywords	GeoEra Project
FOSSIL RESOURCES	surface deformation	HIKE
	uplift, land slide	
	land slide	
	induced seismicity	
	seismic event	
	earthquake	
	fault movement,	
	leakage	
	contamination	
	ground water pollution	
	surface water polution	
	Gas Hydrate	GARAH
	HC resources	
	reservoir	
	Unconventionals	
	Conventionals	
	environmental issues	EuroLithos
	massive sulphides	MINDeSEA
	phosphorites	
	marine placers	
	polymetallic nodules	
	Commodities	Mintel4EU
LITHOLOGY	facies distribution	HOTLIME
	carbonates	
	conventional and unconventional resources	GARAH
	reservoir rock parameters	
	ornamental lithotypes	EuroLithos
	petrographic information	
	predictive and mineral exploration potential map	MINDeSEA
	Phosphorites	
	Marine Placer Deposits	
	Polymetallic Nodules	
	Metallogenetic map	
	phosphate deposits	FRAME
	Mining regions	
	graphite, lithium, cobalt	
	mineral belts	

Search Category	Test Keywords	GeoEra Project
SUBSURFACE MANAGEMENT	resource management	HOTLIME
	UNFC	
	hydraulic properties, groundwater potential, groundwater chemistry, groundwater temperature, corrosion, scaling	
	Groundwater management	MUSE
	Shallow Geothermal Energy	
	Groundwater Temperature	
	Legal Framework	
	Management Strategies	
	Groundwater Quality	
	Urban Areas	
	Conflict of use	
	Resource Assessment	
	Monitoring	
	Land Use Planning	
	Sustainable Energy Use	
	Spatial Planning	
	Best Practice	
	Strategy	
	Policy	
	climate change assessment	TACTIC
	Water management recommendations	RESOURCE
	geo-environmental	VOGERA
	monitoring contaminant	HOVER
	thermal and mineral water	
	vulnerability of the upper aquifer to pollution	

2.3.2 Second evaluation

2.3.2.1 Fossil Resources

The four selected groups have been evaluated. As none of the test words were present in the INSPIRE_FOSSIL_RESSOURCE_3 (Figure 2.3.1) group it was excluded from the selection.

Group	Codelist/thesaurus	Urls	Test Keywords existing	number of test word existing
EarthReML_FOSSIL_RESOSOURCE	EarthReML_CommodityCode	https://vocabs.ands.org.au/viewById/55	gas hydrate	3
			hydrocarbon	
			reservoir gas	
INSPIRE_FOSSIL_RESSOURCE_1	INSPIRE Codelist BoreholePurpose	http://inspire.ec.europa.eu/codelist/BoreholePurposeValue	hydrocarbon	2
			pollution	
INSPIRE_FOSSIL_RESSOURCE_2	INSPIRE Codelist ClassificationAndQuantificationFrameworkValue	http://inspire.ec.europa.eu/codelist/ClassificationAndQuantificationFrameworkValue	none	2
	INSPIRE Codelist FossilFuelClassValue	http://inspire.ec.europa.eu/codelist/FossilFuelClassValue	resources	
	INSPIRE Codelist FossilFuelValue	http://inspire.ec.europa.eu/codelist/FossilFuelValue	Gas	
INSPIRE_FOSSIL_RESSOURCE_3	INSPIRE Codelist OtherContaminatingActivityValue	http://inspire.ec.europa.eu/codelist/OtherContaminatingActivityValue	none	0
	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	http://inspire.ec.europa.eu/codelist/SoilContaminationSpecialisedZoneTypeCode	none	

Figure 2.3.1: Results through test keywords evaluation for fossils resources search category

2.3.2.2 Lithology

To decide if GeoSciML_LITHOLOGY_2 and INSPIRE_LITHOLOGY_2 groups should be included in the Lithology category the test word for lithology where looked inside every codelist.
Only in the GeoSciML_LITHOLOGY_2 three keywords where found, therefore it was included in the final selection.

Group	Codelist/thesaurus	Urls	Test Keywords existing	number of test word existing
GeoSciML_LITHOLOGY_2	GeoSciML_CompositionCategory	https://vocabs.ands.org.au/viewById/54	carbonate	3
			phosphate	
	GeoSciML_GeologicUnitType	https://vocabs.ands.org.au/viewById/50	none	
	GeoSciML_MetamorphicFacies	https://vocabs.ands.org.au/viewById/90	facies	
	GeoSciML_ParticleType	https://vocabs.ands.org.au/viewById/88	none	
INSPIRE_LITHOLOGY_2	INSPIRE Codelist EventEnvironmentValue	http://inspire.ec.europa.eu/codelist/EventEnvironmentValue	none	0
	INSPIRE Codelist EventProcessValue	http://inspire.ec.europa.eu/codelist/EventProcessValue	none	
	INSPIRE Codelist HydroGeochemicalRockTypeValue	http://inspire.ec.europa.eu/codelist/HydroGeochemicalRockTypeValue	none	

Figure 2.3.2: Results through test keywords evaluation for GeoSciML_LITHOLOGY_2 and INSPIRE_LITHOLOGY_2 groups

2.3.2.3 Subsurface Management

To discriminate whether the three groups listed in paragraph 2.2.12 should be included in the final selection, it was searched whether the test words existed among their terms.

Group	Codelist/thesaurus	Urls	Test Keywords existing	number of test word existing
EarthReML_SUBSURFACE_MANAGEMENT	EarthReML_ExplorationActivityType	https://vocabs.ands.org.au/viewById/79	resource assessment	2
	EarthReML_ReserveAssessmentCategory	https://vocabs.ands.org.au/viewById/72	none	
	EarthReML_UNFCCode	http://resource.geosciml.org/classifier/cgi/unfc	UNFC	
	EarthReML_WasteStorage	https://vocabs.ands.org.au/viewById/69	none	
INSPIRE_SUBSURFACE_MANAGEMENT_5	INSPIRE Codelist RiskAssessmentStageValue	http://inspire.ec.europa.eu/codelist/RiskAssessmentStageValue	pollution	4
			assesment	
	INSPIRE Codelist RiskReceptorValue	http://inspire.ec.europa.eu/codelist/RiskReceptorValue	none	
	INSPIRE Codelist RiskTypeValue	http://inspire.ec.europa.eu/codelist/RiskTypeValue	none	
	INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	http://inspire.ec.europa.eu/codelist/SoilContaminationSpecialisedZoneTypeCode	management	
			monitoring	
	INSPIRE Codelist LayerTypeValue	http://inspire.ec.europa.eu/codelist/LayerTypeValue	none	
	INSPIRE Codelist OtherContaminatingActivityValue	http://inspire.ec.europa.eu/codelist/OtherContaminatingActivityValue	none	
ONEGE_SUBSURFACE_MANAGEMENT	OneGeology-Europe keywords database		pollution	2
			climate change	

Figure 2.3.3: Results through test keywords evaluation for Subsurface Management groups

Given that the three groups present some occurrence, they are included in the final selection.

2.3.3 Final selection

After all the processes and tests passed to the vocabularies, the final selection is shown below. Experts in the different search categories should validate this selection.

Search Category	Group	CodeList	Web	Organization	Person Of contact
APPLIED GEOPHYSIC	INSPIRE_APPLIED_GEOPHYSIC	INSPIRE Codelist BoreholePurpose	http://inspire.ec.europa.eu/codelist/BoreholePurposeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist CurveModelTypeValue	http://inspire.ec.europa.eu/codelist/CurveModelTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist ProfileTypeValue	http://inspire.ec.europa.eu/codelist/ProfileTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist StationTypeValue	http://inspire.ec.europa.eu/codelist/StationTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist SurveyTypeValue	http://inspire.ec.europa.eu/codelist/SurveyTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist SwathTypeValue	http://inspire.ec.europa.eu/codelist/SwathTypeValue	JRC INSPIRE Registry Team	
	ONEGE_APPLIED_GEOPHYSIC	OneGeology-Europe keywords database	http://gemet.bnhelp.cz/thesaurus/getTopmostConcepts?thesaurus_uri=http://www.onegeology-europe.eu/concept/&language=en	Czech Geological Survey	lucie.kondrova@geology.cz
FOSSIL RESOURCES	EarthReML_FOSSIL_RESSOURCE	EarthReML_CommodityCode	http://resource.geosciml.org/classifierscheme/cgi/2016.01/compositioncategory or https://vocabs.ands.org.au/viewById/55	CGI Geoscience Terminology Working group	Tim-McCormick or Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_FOSSIL_RESSOURCE_1	INSPIRE Codelist BoreholePurpose	http://inspire.ec.europa.eu/codelist/BoreholePurposeValue	JRC INSPIRE Registry Team	
	INSPIRE_FOSSIL_RESSOURCE_2	INSPIRE Codelist ClassificationAndQuantificationFrameworkValue	http://inspire.ec.europa.eu/codelist/ClassificationAndQuantificationFrameworkValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist FossilFuelClassValue	http://inspire.ec.europa.eu/codelist/FossilFuelClassValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist FossilFuelValue	http://inspire.ec.europa.eu/codelist/FossilFuelValue	JRC INSPIRE Registry Team	
GEOCHEMISTRY	INSPIRE_GEOCHEMISTRY_1	INSPIRE Codelist EventProcessValue	http://inspire.ec.europa.eu/codelist/EventProcessValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist HydroGeochemicalRockTypeValue	http://inspire.ec.europa.eu/codelist/HydroGeochemicalRockTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist NaturalGeomorphologicFeatureTypeValue	http://inspire.ec.europa.eu/codelist/NaturalGeomorphologicFeatureTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist WaterSalinityValue	http://inspire.ec.europa.eu/codelist/WaterSalinityValue	JRC INSPIRE Registry Team	
	INSPIRE_GEOCHEMISTRY_2	INSPIRE Codelist ProcessingActivityTypeValue	http://inspire.ec.europa.eu/codelist/ProcessingActivityTypeValue	JRC INSPIRE Registry Team	
	INSPIRE_GEOCHEMISTRY_3	INSPIRE Codelist ProfileElementParameterNameValue	http://inspire.ec.europa.eu/codelist/ProfileElementParameterNameValue	JRC INSPIRE Registry Team	
GEOCHRONOLOGY/STRATIGRAPHY	GeoSciML_GEOCHRONO_STRATIGRAPHY_1	CGI - Geologic Time Vocabulary - International Chronostratigraphic Chart - 2017	http://vocabs.ands.org.au/repository/api/lda/csiro/international-chronostratigraphic-chart-2017/2017/collection	Research Vocabularies Australia - Linked Data API	

GEOLOGICAL PROCESSES	GeoSciML_GEOLOGICAL_PROCESSES	GeoSciML_DeformationStyle	https://vocabs.ands.org.au/viewById/46 or http://resource.geosciml.org/classifierscheme/cgi/2016.01/deformationstyle	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_EventEnviroment	https://vocabs.ands.org.au/viewById/59	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_EventProcess	http://resource.geosciml.org/classifierscheme/cgi/2016.01/eventprocess or https://vocabs.ands.org.au/viewById/58	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_GeneticCategory	http://resource.geosciml.org/classifierscheme/cgi/2016.01/geneticcategory	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_GeologicUnitType	https://vocabs.ands.org.au/viewById/50	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_MetamorphicGrade	https://vocabs.ands.org.au/viewById/91	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_GEOLOGICAL_PROCESSES	INSPIRE CodelistAnthropogenicGeomorphologicFeatureTypeValue	http://inspire.ec.europa.eu/codelist/AnthropogenicGeomorphologicFeatureTypeValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistEventEnvironmentValue	http://inspire.ec.europa.eu/codelist/EventEnvironmentValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistEventProcessValue	http://inspire.ec.europa.eu/codelist/EventProcessValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistGeologicUnitTypeValue	http://inspire.ec.europa.eu/codelist/GeologicUnitTypeValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistNaturalGeomorphologicFeatureTypeValue	http://inspire.ec.europa.eu/codelist/NaturalGeomorphologicFeatureTypeValue	JRC INSPIRE Registry Team	
HAZARD, RISK AND IMPACT	EarthReML_HAZARD_RI	EarthReML_EnvironmentalImpact	http://resource.geosciml.org/classifier/cgi/environmental-impact	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_HAZARD_RI_1	INSPIRE CodelistAnthropogenicGeomorphologicFeatureTypeValue	http://inspire.ec.europa.eu/codelist/AnthropogenicGeomorphologicFeatureTypeValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistNaturalGeomorphologicFeatureTypeValue	http://inspire.ec.europa.eu/codelist/NaturalGeomorphologicFeatureTypeValue	JRC INSPIRE Registry Team	

	INSPIRE_HAZARD_RI_2	INSPIRE Codelist ExposedElementCategoryValue	http://inspire.ec.europa.eu/codelist/ExposedElementCategoryValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist NaturalHazardCategoryValue	http://inspire.ec.europa.eu/codelist/NaturalHazardCategoryValue	JRC INSPIRE Registry Team	
GEOTHERMAL ENERGY	INSPIRE_GEOTHERMAL_ENERGY	INSPIRE Codelist ActiveWellTypeValue	http://inspire.ec.europa.eu/codelist/ActiveWellTypeValue	JRC INSPIRE Registry Team	
HYDROGEOLOGY	INSPIRE_HYDROGEOLOGY_1	INSPIRE Codelist ActiveWellTypeValue	http://inspire.ec.europa.eu/codelist/ActiveWellTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist AquiferMediaTypeValue	http://inspire.ec.europa.eu/codelist/AquiferMediaTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist AquiferTypeValue	http://inspire.ec.europa.eu/codelist/AquiferTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist ConditionOfGroundwaterValue	http://inspire.ec.europa.eu/codelist/ConditionOfGroundwaterValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist HydroGeochemicalRockTypeValue	http://inspire.ec.europa.eu/codelist/HydroGeochemicalRockTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist NaturalObjectTypeValue	http://inspire.ec.europa.eu/codelist/NaturalObjectTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist StatusCodeTypeValue	http://inspire.ec.europa.eu/codelist/StatusCodeTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist WaterPersistenceValue	http://inspire.ec.europa.eu/codelist/WaterPersistenceValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist WaterSalinityValue	http://inspire.ec.europa.eu/codelist/WaterSalinityValue	JRC INSPIRE Registry Team	
	INSPIRE_HYDROGEOLOGY_2	INSPIRE Codelist BoreholePurpose	http://inspire.ec.europa.eu/codelist/BoreholePurposeValue	JRC INSPIRE Registry Team	
	INSPIRE_HYDROGEOLOGY_3	INSPIRE Codelist OtherContaminatingActivityValue	http://inspire.ec.europa.eu/codelist/OtherContaminatingActivityValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	http://inspire.ec.europa.eu/codelist/SoilContaminationSpecialisedZoneTypeCode	JRC INSPIRE Registry Team	
LITHOLOGY	EARTH_LITHOLOGY	EARTh.	http://linkeddata.ge.imati.cnr.it/resource/EARTh/		
	GBA_LITHOLOGY	GBA Thesaurus Lithology	http://resource.geolba.ac.at/lithology	Geological Survey of Austria	
	GeoSciML_LITHOLOGY_1	GeoSciML_SimpleLithology	https://vocabs.ands.org.au/viewById/56	CGI Geoscience Terminology Working group	
	GeoSciML_LITHOLOGY_2	GeoSciML_CompositionCategory	http://resource.geosciml.org/classifierscheme/cgi/2016.01/compositioncategory	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_GeologicUnitType	https://vocabs.ands.org.au/viewById/50	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_MetamorphicFacies	https://vocabs.ands.org.au/viewById/90	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_LITHOLOGY_1	INSPIRE Codelist LithologyValue	http://inspire.ec.europa.eu/codelist/LithologyValue	JRC INSPIRE Registry Team	

MINERAL RESOURCES	EarthReML_MINERAL_RESOURCE	EarthReML_CommodityCode	http://resource.geosciml.org/classifierscheme/cgi/2016.01/compositioncategory or https://vocabs.ands.org.au/viewById/55	CGI Geoscience Terminology Working group	Tim-McCormick or Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_EarthResourceMaterialRole	https://vocabs.ands.org.au/viewById/78	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_EndUsePotential	http://resource.geosciml.org/classifier/cgi/end-use-potential	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_ExplorationResult	https://vocabs.ands.org.au/viewById/77	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_MineralOccurrenceType	https://vocabs.ands.org.au/viewById/76	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_MineStatus	https://vocabs.ands.org.au/viewById/126	CGI Geoscience Terminology Working group opertaing status	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_MiningActivity	http://resource.geosciml.org/classifier/cgi/mining-activity	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_ProcessingActivity	https://vocabs.ands.org.au/viewById/74	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_RawMaterialRole	https://vocabs.ands.org.au/viewById/73	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_ReportingClassificationMethod	https://vocabs.ands.org.au/viewById/125	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_ReserveAssessmentCategory	https://vocabs.ands.org.au/viewById/72	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	GBA_MINERAL_RESOURCE	GBA Thesaurus Rohstoffgeologie (Raw Material)	http://resource.geolba.ac.at/minres.html	Geological Survey of Austria	thesaurus@geologie.ac.at
	INSPIRE_MINERAL_RESOURCE_1	INSPIRE Codelist EndusePotentialValue	http://inspire.ec.europa.eu/codelist/EndusePotentialValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist ExplorationActivityTypeValue	http://inspire.ec.europa.eu/codelist/ExplorationActivityTypeValue	JRC INSPIRE Registry Team	

		INSPIRE Codelist ExplorationResultValue	http://inspire.ec.europa.eu/codelist/ExplorationResultValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist MineralDepositGroupValue	http://inspire.ec.europa.eu/codelist/MineralDepositGroupValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist MineralDepositTypeValue	http://inspire.ec.europa.eu/codelist/MineralDepositTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist MineralOccurrenceTypeValue	http://inspire.ec.europa.eu/codelist/MineralOccurrenceTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist MineStatusValue	http://inspire.ec.europa.eu/codelist/MineStatusValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist MiningActivityTypeValue	http://inspire.ec.europa.eu/codelist/MiningActivityTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist ProcessingActivityTypeValue	http://inspire.ec.europa.eu/codelist/ProcessingActivityTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist ReserveCategoryValue	http://inspire.ec.europa.eu/codelist/ReserveCategoryValue	JRC INSPIRE Registry Team	
		INSPIRE CodelistCommodityValue	http://inspire.ec.europa.eu/codelist/CommodityCodeValue	JRC INSPIRE Registry Team	
	Minerals4EU_MINERAL_RESOURCE	Minerals4EU Metadata Keywords	http://m4eu.geology.cz/codelist	Czech Geological Survey	egdi.metadata@geology.cz
STRUCTURAL GEOLOGY	EARTH_STRUCTURAL	EARTh.	http://linkeddata.ge.imati.cnr.it/resource/EARTh/		
	GeoSciML_STRUCTURAL	GeoSciML_ContactType	https://vocabs.ands.org.au/viewById/52 or http://resource.geosciml.org/classifierscheme/cgi/2016.01/contacttype	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_DeformationStyle	https://vocabs.ands.org.au/viewById/46 or http://resource.geosciml.org/classifierscheme/cgi/2016.01/deformationstyle	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_FaultMovementSense	https://vocabs.ands.org.au/viewById/63	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_FaultType	https://vocabs.ands.org.au/viewById/68 or http://resource.geosciml.org/classifier/cgi/faulttype	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		GeoSciML_GeologicUnitType	https://vocabs.ands.org.au/viewById/50	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_STRUCTURAL	INSPIRE Codelist EventEnvironmentValue	http://inspire.ec.europa.eu/codelist/EventEnvironmentValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist EventProcessValue	http://inspire.ec.europa.eu/codelist/EventProcessValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist FaultTypeValue	http://inspire.ec.europa.eu/codelist/FaultTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist FoldProfileTypeValue	http://inspire.ec.europa.eu/codelist/FoldProfileTypeValue	JRC INSPIRE Registry Team	

SUBSURFACE MANAGEMENT	EarthReML_SUBSURFACE_MANAGEMEN	EarthReML_ExplorationActivityType	https://vocabs.ands.org.au/viewById/79	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
		EarthReML_ReserveAssessmentCategory	https://vocabs.ands.org.au/viewById/72	CGI Geoscience Terminology Working group	Oliver Raymond: oliver.raymond@ga.gov.au
	INSPIRE_SUBSURFACE_MANAGEMENT_2	INSPIRE Codelist HILUCSValue	http://inspire.ec.europa.eu/codelist/HILUCSValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist LevelOfSpatialPlanValue	http://inspire.ec.europa.eu/codelist/LevelOfSpatialPlanValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist SupplementaryRegulationValue	http://inspire.ec.europa.eu/codelist/SupplementaryRegulationValue	JRC INSPIRE Registry Team	
	INSPIRE_SUBSURFACE_MANAGEMENT_5	INSPIRE Codelist LayerTypeValue	http://inspire.ec.europa.eu/codelist/LayerTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist OtherContaminatingActivityValue	http://inspire.ec.europa.eu/codelist/OtherContaminatingActivityValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist RiskAssessmentStageValue	http://inspire.ec.europa.eu/codelist/RiskAssessmentStageValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist RiskReceptorValue	http://inspire.ec.europa.eu/codelist/RiskReceptorValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist RiskTypeValue	http://inspire.ec.europa.eu/codelist/RiskTypeValue	JRC INSPIRE Registry Team	
		INSPIRE Codelist SoilContaminationSpecialisedZoneTypeCode	http://inspire.ec.europa.eu/codelist/SoilContaminationSpecialisedZoneTypeCode	JRC INSPIRE Registry Team	
	ONEGE_SUBSURFACE_MANAGEMENT	OneGeology-Europe keywords database	http://gemet.bnhelp.cz/thesaurus/getTopmostConcepts?thesaurus_uri=http://www.onegeology-europe.eu/concept/&language=en	Czech Geological Survey	lucie.kondrova@geology.cz

3 COMPILATION OF THE GEOERA KEYWORD THESAURUS

By Matevž Novak

WP4 subtask 4.1.2 “Compilation of a keyword thesaurus” is led by **GeoZS** and the following partners are involved: GBA, ISPRA, SGU, TNO, CGS, GIU, MBFSZ, LfU, BRGM, GTK, GEUS, BGR, HGI-CGS, and LNEG.

The WP4 subtask 4.1.2 on the compilation of a keyword thesaurus is building on the previous completed evaluation of existing vocabularies (see heading 2, subtask 4.1.1) applicable for subject heading systems. This subtask includes the modelling of the subject heading system, concerning the selected and tested vocabularies. The aim is the creation of a new keyword thesaurus and the completion and translation of missing keywords.

Semantic Text Search for data is the basic task for all data infrastructures. It needs to put all keywords used to tag datasets into a single hierarchy like a thesaurus. Data queries can use this kind of a word net also to get search results for similar keywords within a “semantic radius”.
For metadata descriptions, the clarification of the meaning of textual attributes applies mainly to keywords and the implementation of a semantic search within a metadata catalog. Here WP4 subtask 4.1.2 strives for a compilation (RDF/SKOS thesaurus) of keywords with URIs suitable for tagging metadata (-> use case: Multilingual Semantic Text Search).

A Linked Data approach 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.

3.1 Compilation process

3.1.1 Modelling and generating RDF file structure

The first GeoERA Keyword Thesaurus conceptual model (Figure 3.1.1) displays two constituent parts of the thesaurus. The uppermost hierarchy in the Keyword Thesaurus were Search Categories representing generic categories established through the thematic grouping of most of the GeoEra project deliverables included in the proposals. They represent the main headlines of the future GeoERA subject heading system (keyword thesaurus). The main objective was to have a common group of terms by topic for its application in the selective search of the GeoERA product catalogue. This GeoERA subject heading system will facilitate the labelling of products in the GeoERA metadata catalogue. The second constituent part represented thematic grouping to enable keyword tagging with widely used terms and concepts such as scientific topics.

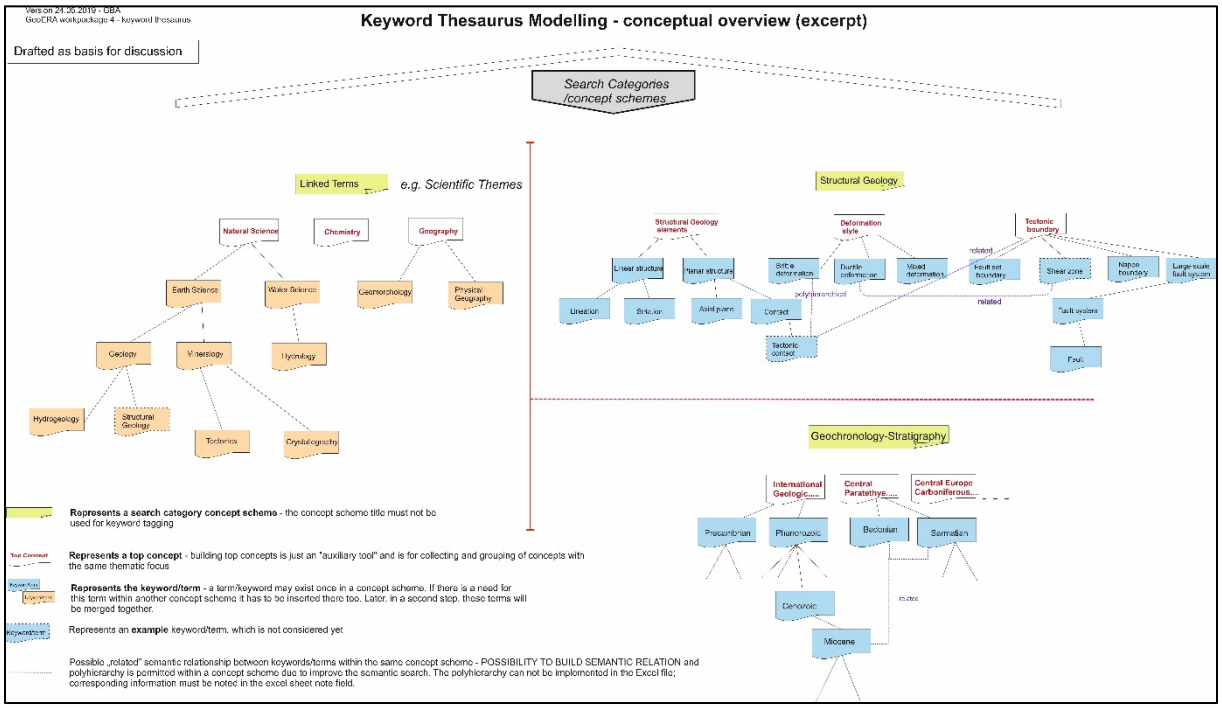

Figure. 3.1.1 Conceptual model of the keyword thesaurus

In the conceptual keyword thesaurus model each Search Category consists of a single hierarchically structured Concept Scheme that can be composed of one or several Top-concepts.

Most concept schemes are based on the existing hierarchically ordered codelists and vocabularies (e.g., INSPIRE, GeoSciML vocabulars, GBA thesaurus, etc.), selected within the WP4 subtask 4.1.1. Each term/keyword is linked with as many as possible codelist URIs forming the Linked data structure (Figure 3.1.2).

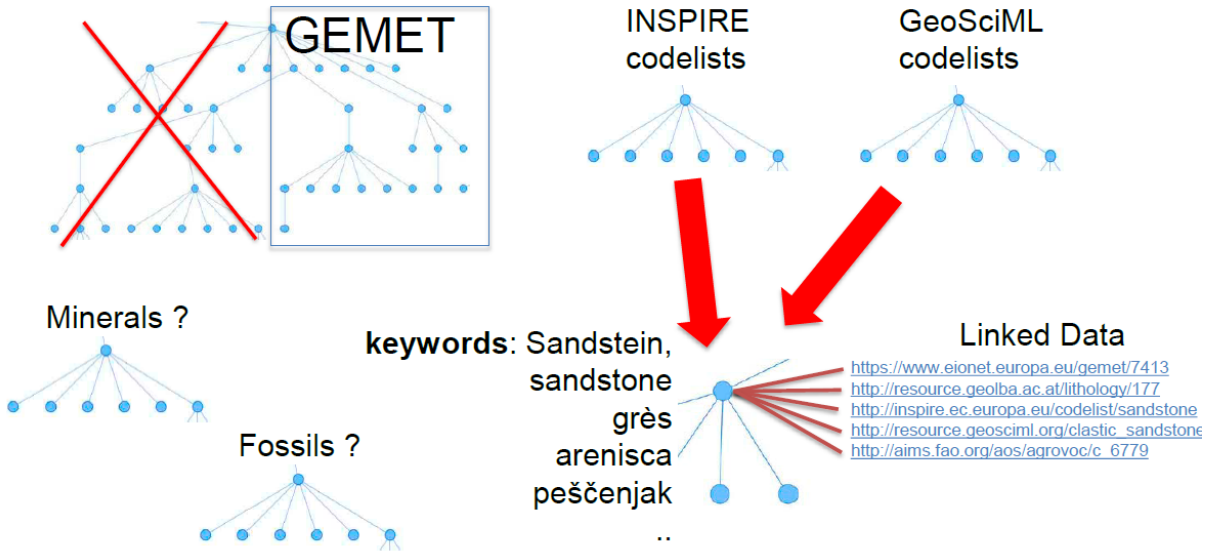

Figure 3.1.2 The conceptual model of Linked Data structure

The first draft template of the Keyword Thesaurus Excel file comprised individual Excel worksheets for each search category, 15 altogether. All worksheets were structured in the same way according to the model, of hierarchically structured keyword concepts with character encoding UTF8 to generate an RDF file.

The Concept Schemes were named according to the Search Categories and supplemented by a brief description of corresponding topics or disciplines. The columns of Top-concept and lower ranked concepts were followed by columns for mapping codelists/vocabularies URIs and columns for translations into several languages. The headers for the translations were called “prefLabel@” together with the language tags, e.g. prefLabel@de, prefLabel@si, ... Using “altLabel” refers to synonyms (Figure 3.1.3).

Scheme	dcTerms description	(Top)con-cept	concept	concept	GeoSciML_ MappingURI _exactMatch	INSPIRE_ Mapping URI _exactM atch	GBA_The saurus_ Mapping URI _exactM atch
Search Category name							

NOTES skos:rela ted info	prefLabel@de	prefLabel @es	prefLabel @fr	prefLabel @it	prefLabel @cz	prefLabel @si	etc…
	Geochronologie-Stratigraphie					geokrono logija-stratigrafi ja	

Figure 3.1.3 The structure of Keyword Thesaurus Excel worksheets.

For the test entry we chose to fill the Geochronology-Stratigraphy Search Category as the simplest scheme which is based on the International Chronostratigraphic Chart with well established hierarchy.

The draft template and the test entry results raised several questions, discussed by project partners. Problems and solutions are listed here:

3.1.1.1 Implementing related keyword/concept hierarchies from different sources into one scheme

Besides the subdivision of the International Chronostratigraphic Chart, several other regional subdivisions are widely used, such as the Carboniferous subdivision in Central Europe, the Neogene subdivision in the Central Paratethys or Alpine glacial cycles.

Ad 1.: To implement keywords from different sources on the same topic there are two possibilities. One is the incorporation of those keywords into the existing basic hierarchy (in this case into the Chronostratigraphic terms from the ICS). The second possibility is to formulate another Top-concept within the scheme (e.g., Central Paratethys Stages) and create the separate hierarchy for it. The semantic relationship is then specified by the labels to be linked in the column skos:related info, which has been added (Figure 3.1.3).

3.1.1.2 Keywords topic issues

Several keywords/concepts do not fit within the existing hierarchies, neither are they directly related to the topic of hierarchies, but they do relate to the same Search Category. E.g., names of the lithostratigraphic units, important orogenic events, or geochronological age dating methods, all relate best with the Geochronology-Stratigraphy Search Category.

Ad. 2.: In general, keywords in a Search Category should reflect the same topic. If the keywords in a Search Category are concerning a topic that does not fully fit into this search category it should be decided whether it is better to create a new Top-concept (and a related hierarchy) for these keywords or to create a new separate Excel sheet called "Loose concepts" ("Linked Terms" in the current version). Especially, for more general and widely used keywords, which are important for tagging, but do not fit into any Search Category. This Excel sheet has been renamed into "Linked terms".

3.1.1.3 The level of detail – how many hierarchical levels are needed to be modeled

Ad 3.: It is up to the compiler to consider, which terms may be required by the project data providers to tag the different dataset products and which terms might be required to obtain a satisfactory search result. There is no general rule on the number of levels.

3.1.1.4 Establishing keyword relationships

The basic problem is that none of the existing thesauri is consistent in parent/child relationships of the concepts – even within those thesauri. To be more specific; igneous rocks can be subdivided either based on the "environment" and the related structure (e.g. plutonic/volcanic or fine grained) or based on chemical composition (e.g. acidic/intermediate/ultrabasic or ultramafic). As consequence, one concept can have more parents. How/if to incorporate this?

Ad 4.: Keywords are semantically linked either by a broader/narrower relationship within the hierarchy, or by specifying and entering related keyword terms (one or more, separated by a comma) in the skos: related info column.

3.1.2 First phase of compilation

The first phase of the Keyword Thesaurus compilation was finished and, supplemented with an internal report, sent to project partners on 3rd April this year. This phase comprised more of a collection than the selection of terms. Only based on names and links of Code Lists selected for each Search Category within the WP4 subtask 4.1.1 it was difficult to see the overall picture of what is included in the existing vocabularies. The total number of entries after the first phase of compilation was 1,704 keywords (Figure 3.1.4).

Search Category	Number of entries
GEOCHRONOLOGY/STRATIGRAPHY	214
LITHOLOGY	209
GEOLOGICAL PROCESSES	56
STRUCTURAL GEOLOGY	94
APPLIED GEOPHYSIC	74
GEOCHEMISTRY	62
HYDROGEOLOGY	75

Search Category	Number of entries
GEOTHERMAL ENERGY	19
SUBSURFACE ENERGY STORAGE	24
MINERAL RESOURCES	498
FOSSIL RESOURCES	57
HAZARD, RISK AND IMPACT	74
SUBSURFACE MANAGEMENT	125
INFORMATION SYSTEM	66
MODELLING	57
Loose concepts	Not yet relevant
SUM	**1,704**

Figure 3.1.4: The number of entries per Search Category in the Keyword Thesaurus after the first phase of compilation.

Because of the open questions that arose within this phase, the selected terms from the GEMET thesaurus, KINDRA vocabulary and VogERA terms have not been included yet. Neither were included the requirements of the project leaders collected in the Annex II of the Subtask 4.1.1 report.

The analysis of the compilation results raised new questions, discussed by project partners at the Teleconference on June 5. Problems and solutions are listed here:

3.1.2.1 Defining Top-concepts within Search Categories

It's inconsistent. E.g. in Search Category Lithology every major rock group is a separate Top-concept, while in Structural Geology all structural elements are under a single Top-concept. In many cases there are several possible ways to group concepts in higher ranks (Top-concepts). E.g., from GEMET: Soil and Soil process are two concepts within Top-concept Pedosphere. However, they can also be defined as two Top-concepts.

Ad 5.: It does not matter how many Top-concepts are within one Concept Scheme.

3.1.2.2 Overlapping Search Categories; several Top-concepts would fit into different Search Categories

E.g., overlapping between:

- Hazard, risk and impact and Subsurface management: e.g., Categories of Hierarchical Supplementary Regulation Code List (HSRCL) (http://inspire.ec.europa.eu/codelist/SupplementaryRegulationValue) and several concepts related to pollution which also overlap with Geochemistry: (http://inspire.ec.europa.eu/codelist/OtherContaminatingActivityValue, http://inspire.ec.europa.eu/codelist/RiskTypeValue, http://inspire.ec.europa.eu/codelist/RiskAssessmentStageValue, http://inspire.ec.europa.eu/codelist/RiskReceptorValue)
- Hydrogeology and Geochemistry: e.g., Hydrogeological rock/groundwater type (http://inspire.ec.europa.eu/codelist/HydroGeochemicalRockTypeValue)
- Subsurface management and Geochemistry: e.g., Soil contamination specialised zone type (http://inspire.ec.europa.eu/codelist/SoilContaminationSpecialisedZoneTypeCode)
- Geochemistry and Mineral Resources: e.g., Processing activity

- Geological process and Geochemistry: e.g., Chemical weathering

This was one of the main issues. Can the same Top-concept belong to two or more different Schemas/Search Categories? And, if so, should all “sub-concepts” be included in each category?
E.g., “borehole purpose” or “active well type” – selected were only those types that are relevant to the Search Category and not all. Another example: terms collected under “pollution” relate to many Search Categories, i.e., Hazard, risk, impact, Geochemistry, Hydrogeology, Subsurface management.

Ad 6. Each keyword should be assigned to one search category. Only if the meaning is fundamentally different, keywords should be assigned to two categories. E.g., keyword "sand", which can be regarded as lithological concept (grain size, grain rounding ...) as well as mineral resource (kind of commodity...). In the search itself, the term "sand" is then displayed only once. Keywords may occur only once within a Search Category.

The keyword “pollution” may fit into several Search Categories, but the fundamental meaning would not be different by assigning this term to some other scientific topic. Therefore, it has to be decided to which Search Category this term shall be allocated.
It should always be remembered that the focus is on search logic rather than scientific modelling of the keywords.

3.1.2.3 Should keywords for which links (URIs) do not exist in any of the existing codelists/vocabularies or links are not active as in the case of OneGeologyEU, be considered and included

The developer of OneGeologyEU vocabularies informed us that 1G-E keywords were made in an old SW for GEMET, where URIs are in this form: http://www.onegeology-europe.eu/concept/15 – the terms are substituted by numbers and they are not referenceable (the web to which they are pointing doesn't exist). In MICKA you can only see the text description.

Ad 8.: Yes, the relevant keywords without links have to be included. Links/URIs are not required for vocabularies which will be used in the Keyword Thesaurus for tagging. The OneGeologyEU-keywords should also be considered.

3.1.2.4 Integration of keywords from GEMET thesaurus, KINDRA vocabulary and VogERA terms

Introducing GEMET's terms in the existing structure is a difficult task since the principle of hierarchies in GEMET is in many cases entirely different to the ones of INSPIRE, GeoSciML, GBA Thesaurus. As opposed to the mother-child relationship in these codelists, GEMET's hierarchies are merely based on relations - more like library bookshelf -, which in many cases are very vague (e.g., continent as sub-category of continental shelf / potash as sub-category of carbonate and this as sub-category of carbon dioxide, etc…).

Ad 7.: It is important to decide whether to split a concept hierarchy tree from GEMET (or from other vocabularies) or to use the entire concept hierarchy tree (see example in Figure 3.1.5). If appointment of GEMET's keywords to a specific Search Category causes the former hierarchy to "dissolve" too much, they should be included as complete parts into the Linked terms Search Category.

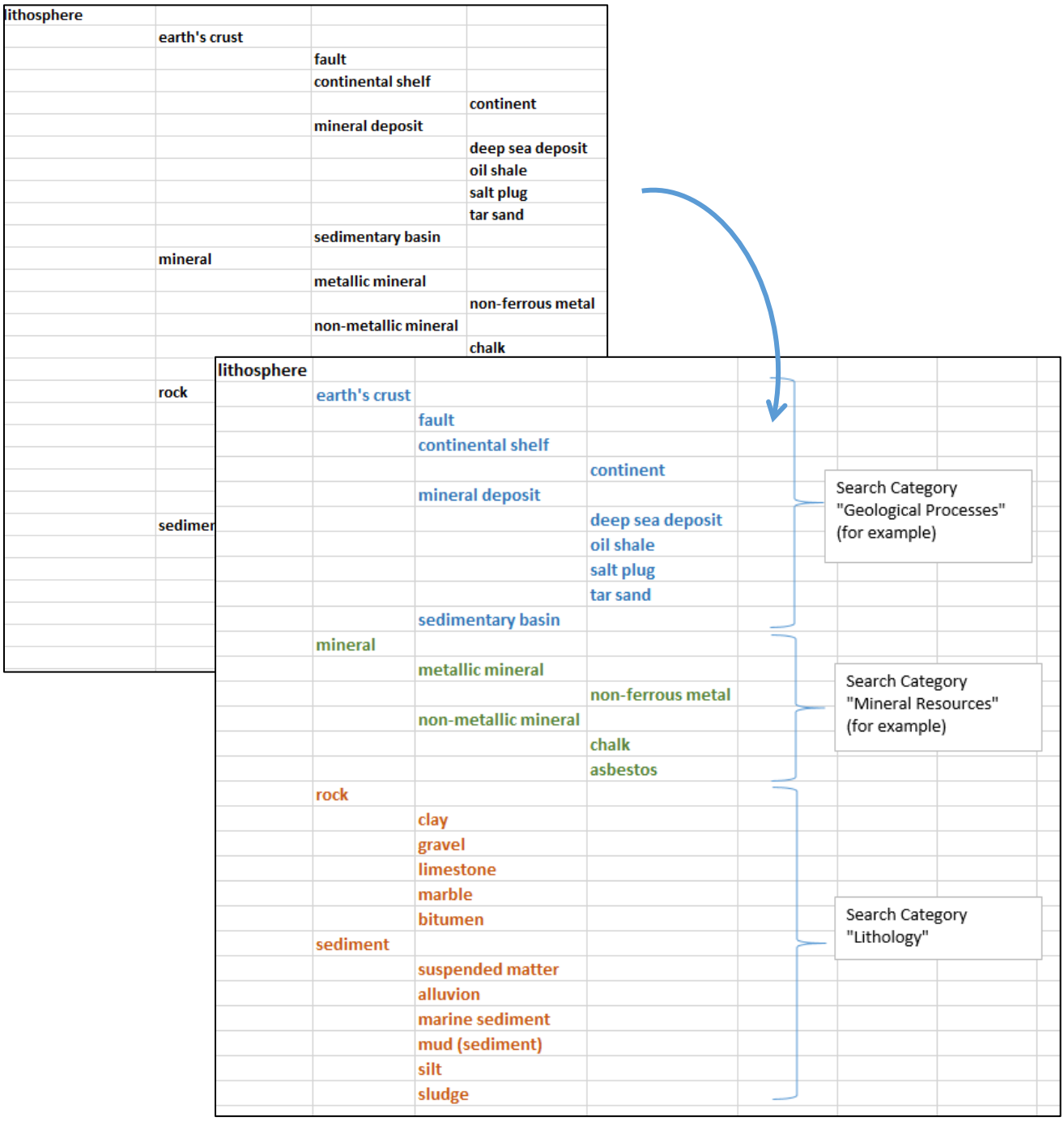

Figure 3.1.5: An example of how to divide GEMET's keyword/concept hierarchy tree "lithosphere" into different Search Categories.

3.1.3 Second phase of compilation

The second phase of the Keyword Thesaurus compilation was finished on July 1. It consisted of implementation of the selected keywords from the GEMET thesaurus, KINDRA vocabulary, VogERA terms and requirements of the project leaders collected in the Annex II of the Subtask 4.1.1 report.
The most important, however, was searching for the missing necessary terms and/or reference vocabularies and, most of all, the critical selection of the compiled concepts.
The result was the Excel document Keyword Thesaurus FINAL with the total number of 2,524 keywords (Figure 3.1.6).

Search Category	Number of entries
GEOCHRONOLOGY/STRATIGRAPHY	214
LITHOLOGY	209
GEOLOGICAL PROCESSES	57
STRUCTURAL GEOLOGY	94
APPLIED GEOPHYSIC	82
GEOCHEMISTRY	133
HYDROGEOLOGY	195
GEOTHERMAL ENERGY	20
SUBSURFACE ENERGY STORAGE	31
MINERAL RESOURCES	523
FOSSIL RESOURCES	66
HAZARD, RISK AND IMPACT	204
SUBSURFACE MANAGEMENT	191
INFORMATION SYSTEM	88
MODELLING	68
Linked terms	349
SUM	**2,524**

Figure 3.1.6: The number of entries per Search Category in the Keyword Thesaurus after the second phase of compilation.

3.1.4 Translation

It has been decided to move the keyword translation to the second part of the GeoERA Keyword Task (after submitting the RDF files at the end of August). We have filled in the missing translations (Google, GEMET) in a first step just to have a basic multilingual translation for the RDF and to test the whole structure. Due to the improvements still to be made, we will have to ask the project partners to revise the keywords a second time. Translation before that would not be very efficient.

3.2 Integration and validation

After the compilation process has been finalized, the integration and validation (SKOS) through the GBA semantic management tool (PoolParty - Semantic Web Company) started.
This implementation required several improvements of the compilation file:

- deleting multiplications, merging of concepts with the same meaning, changing synonyms, dealing with adjectives
- checking the keywords regarding their usage for tagging and search
- import the keywords in the Thesaurus management system (PoolParty) in order to validate (fix errors, links …) and to create an RDF file

The final product after this stage, the keyword thesaurus RDF file version 1.0 has been sent to the project leader for review on August 14.

- The Keyword Thesaurus now contains 2545 terms (concepts) in English, partly (+/- 750 terms) pre-translated into 22 other languages by GEMET or GeoSciML.
- Including links to sources like INSPIRE codelists, GeoSciML vocabs, GEMET thesaurus and GBA thesaurus
- Attributed with 16 “Search Categories” (dbpedia:category) available to filter search terms or results
- Finally, we also integrated terms from KINDRA and VogERA
- It is a Keyword Thesaurus RDF file **version 1.0** to be extended with future versions!
- It is now available via web API (Sparql endpoint) at: https://resource.geolba.ac.at/PoolParty/sparql/geoera_keyword (e.g. you can run the sparql.txt here, just change the language parameter “de” into selected languages like “en” or “si”)
- The domain name of URIs is still resource.geolba.ac.at, the path name is geoera_keyword – as long as the final domain name isn’t clear
- To play with the modeled logic you can use http://www.geolba.net/semantic-search/ in your set browser language
- Or view in GBA new thesaurus page with URI parameters like: https://resource.geolba.net/?uri=http://resource.geolba.ac.at/geoera_keyword/information
- We also tested by importing into Geonetworks 3.2.2 and tagging datasets (use English language to write English keywords!)
- Then it is clear that we need a modeling and test phase to connect similar keywords (e.g.: if somebody is interested in A – he/she may search for B too)
- As soon as we finalize merging disambiguate keywords (like borehole, borehole purpose or drilling) – GeoERA participants could finalize translations (supported by Google translate?)

4 GOVERNANCE PLAN AND WORKFLOWS AROUND THE KEYWORD THESAURUS

By Lucie Kondrová

WP4 subtask 4.1.3 “Compilation of a keyword thesaurus” is led by CGS and the following partners are involved: GBA, GeoZS, BGR, LfU.

For a multilingual semantic text search, this WP4 subtask 4.1.3 aims at the design of a governance plan for a keyword thesaurus including workflows for application, crosslinking to other Linked Data resources, and thesaurus maintenance - in order to establish a multilingual and semantic subject heading system for the GeoERA platform.

4.1 Governance Plan

According to ISO 19135-1, every register and subregister shall have a register owner, a register manager, and at least one submitting organization or community. Register owner is an organization that establishes a register and makes it available for the public. A register owner shall specify the criteria that determine which organizations may act as submitting organizations and can delegate the management of the register to another organization called register manager. The owner shall decide whether a control body is required for the register (or the owner himself can act as the control body). A register manager informs community bodies of subsequent additions, deletions or amendments to any included vocabulary. The submitting organization or community proposes changes and verifies the correct input of the terminology into the register and advises the register owner of any changes to the terminology. The roles and responsibilities are described in detail in Figure 4.1.1 from ISO 19135-1 (Organizational relationships).

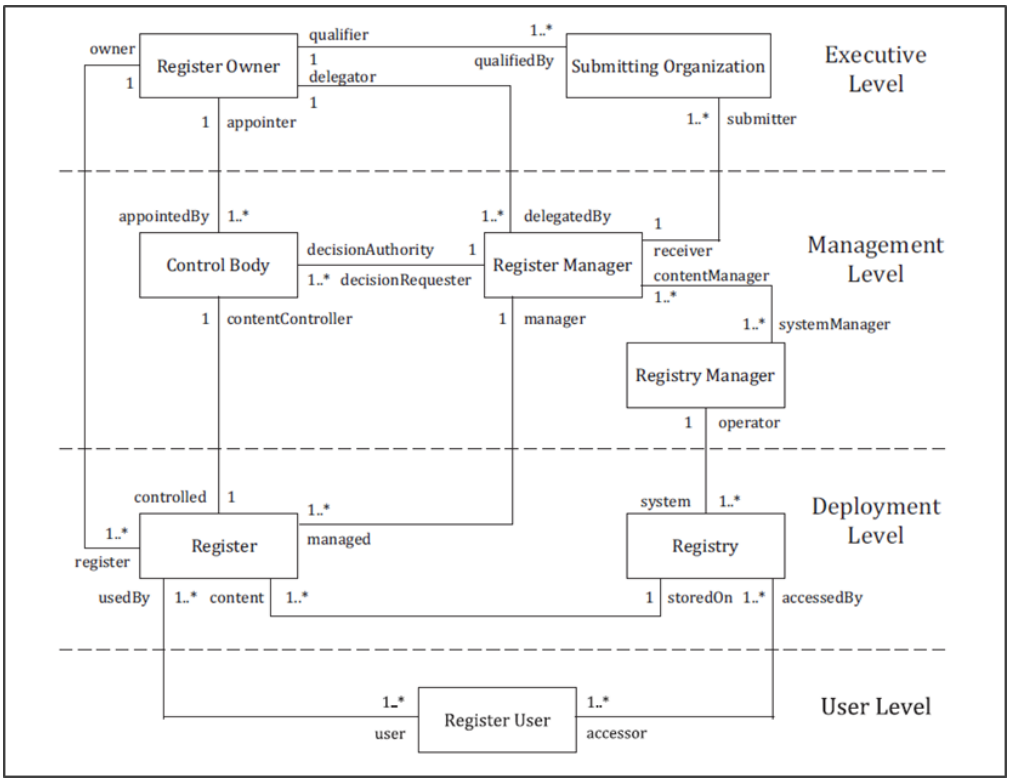

Figure 4.1.1: Organizational relationships for the management of registers (from ISO 19135-1)

4.2 Workflows

4.2.1 Management of changes, revision of keywords, translations, update and extension workflow during the project and after the project end

Running the thesaurus service includes three levels of maintenance: technical, content-related, and language-related. The proposed scheme of responsibilities and communication flows are described by Figure 4.2.1 below.

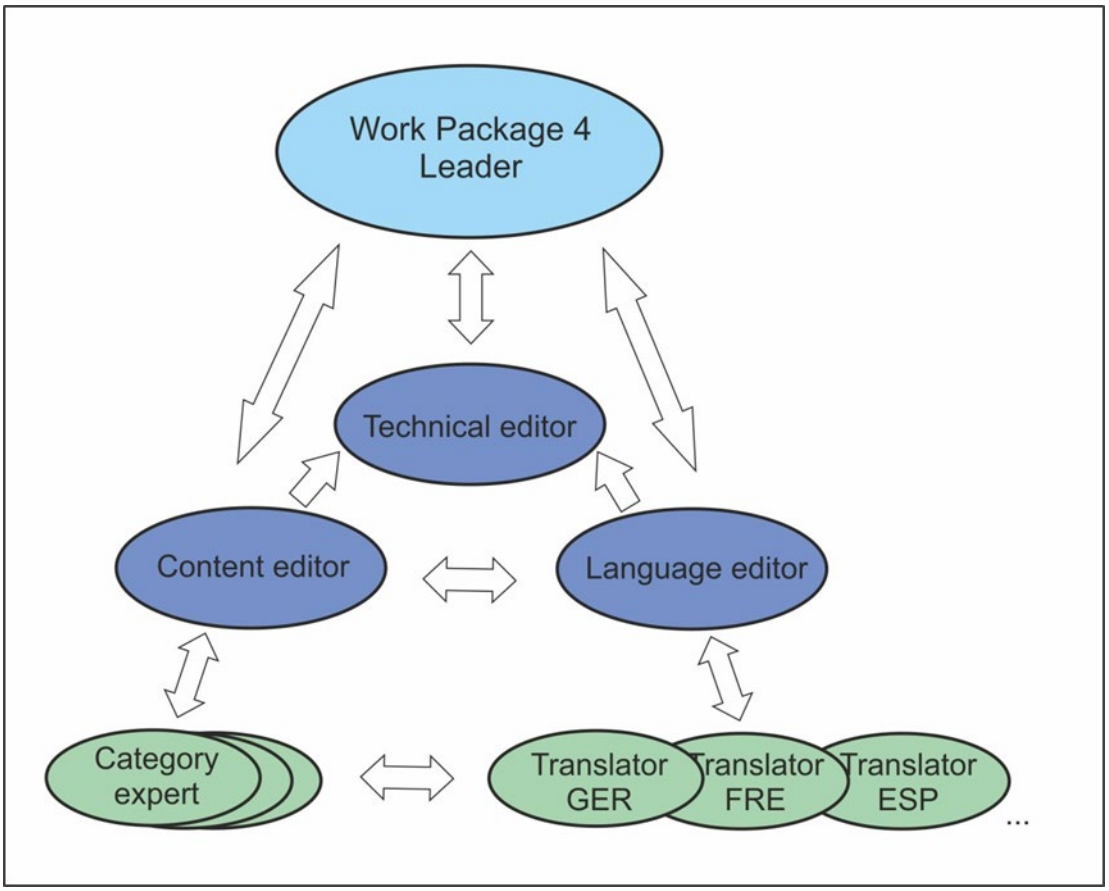

Figure 4.2.1: Responsibilities and communication flow in the process of creation and maintenance of the keyword thesaurus

The terms in the thesaurus should be revised regularly (to be discussed) – proposals for additions, amendments or deletion of terms can be results of the revision. According to ISO 19104, all candidate terms must be assessed and evaluated by the terminology maintenance group within two months of the received proposal. A similar concept is proposed for the duration of the GeoERA project – proposed changes will be evaluated by the WP4 group and then (if agreed) added to a new version of the thesaurus. No terms will be deleted from the thesaurus, they will only be marked as „deprecated“ in the following published version.

After the GeoERA project end, EGDI should be responsible for running the thesaurus – therefore, an expert group on the EGDI level should be established, that would be responsible for the technical, content-related and language-related aspects of running the thesaurus for future use in any geoscientific projects and research. This could either be a new group, or a group of experts selected from the existing EuroGeosurveys expert groups. If no such body is established, the thesaurus will gradually become obsolete.

4.2.2 Backup processes

All files that are necessary to run the GeoERA thesaurus should be stored in the EGDI central data store, which should be backed up regularly and managed sustainably also for the future use after the GeoERA project end.

4.2.3 Management of the domain of the terms

The domain defining the namespace for the terms from the thesaurus should be owned and managed by EGDI, which would guarantee its sustainable operation in the future, so that the domain won't expire and the URIs will remain unchanged.
This topic is discussed at the moment. For more information concerning URIs and URI design please have a look at 3.2 URI design chapter in the GeoERA WP4 deliverable D4.3 "GeoERA Project Vocabularies"

4.2.4 Maintenance of the service

We suggest the BRGM to be responsible for this task, but it has to be discussed.

4.2.5 Contact point/support/information

The CGS maintains the MICKA metadata catalogue for EGDI so they would be the favourable responsible party, but this has to be discussed.

4.2.6 Licensing

The thesaurus will be published as Linked Open Data under the free license Creative Commons Attribution (CC-BY 4.0) for free reuse.

4.3 Use of the keyword thesaurus in the metadata catalogue

EGDI metadata catalogue is able to consume either text strings or URIs as keywords, in order to describe spatial datasets or services. Therefore, there shouldn't be any problem using the terms from the GeoERA thesaurus, as long as it is available either as a web service, or as a stored RDF file. The optimal option will be selected after a testing phase at the end of 2019.

4.4 ISO References

ISO 19104:2010 Geographic Information – Terminology
ISO 19126:2009 Geographic Information – Feature Concept dictionaries and registers
ISO 19135-1:2015 Geographic Information – Procedures for item registration – Part 1: Fundamentals
ISO 19146:2018 Geographic Information – Cross-domain vocabularies