Deliverable 2.3.1

Mapping and describing the needed extensions to EGDI directly related to the task 2.2

Authors and affiliation:
Martin Hansen GEUS,
Andrej Vihtelič Geo-ZS,
Lucie Kondrová CGS,
Pavla Kramolišová CGS,
Ángel Prieto Martín IGME,
Bjarni Pjetursson GEUS,
Marianne B. Wiese GEUS,
Jonas Thyregod GEUS,
Rob van Ede TNO,
Viktor Rasmussen GEUS
Frands Schjøth GEUS
David García Moreno RBINS

[BENEFICIARY]

E-mail of lead author:
mh@geus.dk

Version: 23-04-2020

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	D2.3.2	
Dissemination level	Public	
Deliverable name	Filling the gap between the first extensions to EGDI and the actual geoinformation produce by the projects	
Work package	WP2, User Requirements	
Lead WP/Deliverable beneficiary	GEUS	
Deliverable status		
Submitted (Author(s))	23/04/2020	Martin Hansen
Verified (WP leader)	23/04/2020	David García Moreno
Approved (Coordinator)	23/04/2020	Jørgen Tulstrup

GENERAL INTRODUCTION

This report describes the extensions to the European Geological Data Infrastructure (EGDI) that the GeoERA Information Platform Project (GIP-P) is implementing in order to meet the user requirements of the different geoscientific projects of GeoERA. The requirements analysed in the present report come from descriptions provided by the projects and gathered in reports D2.1.2, D2.1.3 and D2.2.2.

TABLE OF CONTENTS

DEFINITIONS 6

1 INTRODUCTION 8

2 EXTENSIONS TO THE WEB GIS 9

2.1 Web portal tabular functionality 10

2.2 Web portal geospatial functionality 10

2.3 Handling of time component 11

3 DEVELOPMENT OF A GENERAL SEARCH SYSTEM 12

3.1 SOLR search engine 12

4 DEVELOPMENTS ON THE 3D DATABASE 14

5 DEVELOPMENT ON A 3D VIEWER 15

6 EXTENSIONS TO THE ADMINISTRATION MODULE 17

6.1 User management 17

6.2 Uploading spatial data 17

6.3 Document repository 18

6.4 Registering services 18

6.5 Upload of 3D models 18

6.5.1 Metadata on whole 3D Geological Models 19

6.5.2 Metadata on each Mapped Feature / Model Part 19

6.5.3 Metadata on Mapped Features belonging to each Model 20

7 DEVELOPMENT OF A DOCUMENT REPOSITORY 21

8 INTEGRATION OF VOCABULARIES AND KEYWORD THESAURI 22

8.1 Keyword Thesaurus 22

8.2 Project Vocabularies 22

8.3 Knowledge system 22

8.4 URI design 23

8.5 Publication of Project Vocabularies 23

9 EXTENSIONS TO THE CENTRAL DATABASE 24

9.1 User management 24

9.2 Project schemas 24

10 MAKING DATA AVAILABLE THROUGH SERVICES 25

10.1 Making data available as WMS and WFS services 25

10.2 REST services 25

10.2.1 Map and layer configuration 25

10.2.2 The document repository 25

10.2.3 The 3D viewer 25

10.2.4 Download services 25

11 CONNECTION TO THE EGDI METADATA CATALOGUE 26

12 EXTENSIONS TO THE HARVESTING SYSTEM 27

12.1 M4EU harvesting improvements and extensions 27

12.2 Mineral Yearbook (E-MYB) harvesting extensions..............................29
12.3 MIcKA harvesting extensions..30
13 MONITORING SYSTEM...31
REFERENCES...33

DEFINITIONS

Application Programming Interface (API): a computing interface to a software component or a system, that defines how other components or systems can use it.

camelCase: practice of writing phrases such that each word or abbreviation in the middle of the phrase begins with a capital letter, with no intervening spaces or punctuation. Common examples include "iPhone" and "eBay"

Functionality: the range of operations that can be run on a computer or other electronic system.

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

Metadata: data that provides information about spatial and non-spatial data (e.g., purpose of the data, time of creation, authors, etc.)

Non-spatial data: documents (PDFs, text files, etc.), photos/images (JPGs, PDFs, etc.), datasets (TXT, CVS, etc.), URL (DOI, etc.), etc. These data can or cannot be linked to spatial data.

Product: any deliverable generated by a GeoERA project that will be available via EGDI. Projects will deliver 4 types of products:

Project vocabulary: collections of terms with short descriptions, bibliographic citations and links to unstructured web contents used to define scientific parameters and concepts.

Representational state transfer (REST): software architectural style that defines a set of constraints to be used for creating Web services.

REST services: Web services that conform to the REST architectural style.

Spatial data: data concerning phenomena implicitly or explicitly associated with a location within Earth. These typically are:

- **2D, 2.5D and 3D GIS data:** shapefiles, GeoPackages, GeoTiffs, ASCII grids, etc.

- **Geographically localized 3D models**

- **Open Geospatial Consortium (OGC) Web services**: services defined by the OGC, allowing all kinds of geospatial functionality, e.g., WMS, WFS, ATOM. They include services for data access, data display and data processing.

Simple Knowledge Organization System (SKOS): a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary.

SPARQL: Resource Description Framework (RDF) query language. That is, a semantic query language for databases, which is able to retrieve and manipulate data stored in RDF format.

Triple store: purpose-built database for the storage and retrieval of triples through semantic queries. A triple is a data entity composed of subject-predicate-object, like "Bob is 35" or "Bob knows Fred".

1 INTRODUCTION

Even though the existing European Geological Data Infrastructure (EGDI) is fully functional, it does not have many of the functionalities required by the different geoscientific projects (GSPs) of GeoERA. In the present deliverable, we describe the extension of EGDI that the GIP-P has to undertake to meet the requirements necessary to archive, plot and share the data that the various GSPs will create in the framework of GeoERA. The different requirements that the geoscientific projects requested are described in D2.3.1 and D2.2.2. The information provided in the present report is based on those reports.

Deliverable D2.3.2 is divided in 13 chapters, each describing the different extensions of EGDI according to their functionality. Those are:

- Extensions to the web GIS
- Development of a search system
- Development of a 3D database
- Development of a 3D viewer
- Extensions to the administration module
- Development of a document repository
- Integration of vocabularies and thesauruses
- Extensions to the central database
- Making data available through services and for download
- Connections to the EGDI metadata catalogue
- Extensions to the harvesting system
- Setup of a monitoring system

Note that in this document, we do not discuss in detail the latest extension of the EGDI metadata catalogue performed by the GIP-P, as that extension is described in deliverable 7.1. We do however provide a brief description of it.

2 EXTENSIONS TO THE WEB GIS

The functionalities described in this chapter covers the user-oriented enhancements of the system that will be available through EGDI.

Many of the requirements identified by the GSPs refer to the web portal's user interface. Some of them are covered by existing functionalities of EGDI. However, several requirements needed the addition of new functionalities to the user interface; the new functionalities are:

- Go to location by typing city or coordinates (lat. / long. in decimal degrees) in the search field
- Small overview map
- Legend with hierarchy / tree view
- Switching layers on / off
- Multiscaling; showing more and more details when users zoom in to an area
- Transparency of layers
- Create simple queries and filters from the web GIS interface

Other functionalities must be hand-tailored to specific projects/datasets, such as:

- Display graphs with time series data (piezo levels, rain….)
- Creation of statistical diagram, rose diagrams, histograms, time series…

The maps will, by default, be shown in EPSG:3034 in EGDI webGIS portals. We thus recommend that the GSPs send their data in that projection or in a projection that can be transformed to that one by using standard tools (e.g., Proj4). It will however be possible to set up specific maps in other projections. However, maps with other projections might not be able to plot data from other sources if these does not support those projection.

Other functions require a background functionality, for example, to export a map view as a high-resolution image for publication. For this functionality to work properly, external WMS services (those set up by the thematic projects) must be able to deliver high resolution images.

Also, the "getFeatureInfo" must work in different ways depending on how many objects the user selects. If the user selects multiple features the result will be a list, whereas, if the user only selects a single feature, the result will be shown in a popup. For example, if the user selects an array of points from the map representing boreholes, the user interface will return a list of boreholes from where the user can get their descriptions by clicking on each of them. If the user only selects one point the user interface will return a popup window with the description of that specific borehole.

2.1 Web portal tabular functionality

This is a functionality conceived to display concepts and related concepts from EGDI's triple store (geoscience.earth/ncl/geoera) in a way that makes hierarchical sense if needed. That is, all the information available about a concept is displayed in a tabular structure, including links to related concepts, like broader and narrower concepts and bibliographic citations. The information available for each concept can be multilingual if the concepts are available in multiple languages.

An example of this functionality can be found at: https://thesaurus.geolba.ac.at/index.html?uri=http://resource.geolba.ac.at/structure/187&lang=en

2.2 Web portal geospatial functionality

This functionality will provide the user with all the information available on a dataset exposed in EGDI from different data sources, like vocabularies, Keyword Thesauri and the central database. Examples of this functionality are provided below.

1. From the map viewer:

   Dynamically displaying information on geometries from the spatial data stored in EGDI based on SPARQL query results. The process of displaying information about the geometries shown in the map is performed by a data-driven webpage that is built around the contents of the central database, which is enriched with additional contents from relevant information stored in project vocabularies and the keyword thesaurus. The additional contents are retrieved from the SPARQL endpoint by using SPARQL queries. Hence, by clicking on a feature shown in one of EGDI map viewers, the users cannot only see the attributes comprised in that feature, but also have access to their definitions and other information stored in the Project Vocabularies, Keyword Thesaurus and/or EGDI database.

2. From attributes shown in a table view:

   Relevant (tabular) data is shown after selecting one or more features from the map. If the dataset contains a link to a Project Vocabulary, a SPARQL query is executed. This query parses attribute values for the feature based on a getFeatureInfo response. The results of the query are displayed in a tabular form in a popup window or another page. Structured information from the central geo-database, combined with linked data from the triple store is displayed. In some cases, a URI web link from the getFeatureInfo request could be used directly for displaying the relevant data.

3. PDF for download:

   If the data related to a map feature (an attribute or a vocabulary entry) contain a link to a document archived in the document repository, the user should be able to download that document through that link. This is just another link to the document repository (see Chapter 7 of this document), but the link is stored in the triple store and retrieved through a SPARQL query.

Example of this functionality:
https://thesaurus.geolba.ac.at/structureViewer.html?uri=http://resource.geolba.ac.at/structure/186&lang=en

A prototype of this functionality has already been built into the EGDI portal:
http://egditest01.geus.dk/egdi/?mapname=egdi_geoera_muse#baslay=baseMapGEUS&optlay=&extent=3659870,2027340,5025520,2702340&layers=muse_vienna_basin_fault_system

2.3 Handling of time component

If a dataset with time component are delivered as WCS, the user will be able to show it on the webGIS as a map and select which time step to show from the user interface.

3 DEVELOPMENT OF A GENERAL SEARCH SYSTEM

The GIP-P is developing a search system capable of searching through the entire EGDI. The aim of the searching system is to provide multilingual searches through the different information available at EGDI and present the results sorted by relevance. This relevance will be calculated based on where the search word is found. When searching for datasets through their metadata, those that include the search word in their title will have a higher relevance than those that include it in their description. When someone searches for subsets of records within a dataset, the results will also be sorted by relevance. This relevance will be conditioned by the data model of each dataset and the importance that each attribute has. For example, when searching for documents within the document repository, if the word is found in a report title relevance will be high and, if it is found in the text body of the document, relevance will be low.

For the search system, the different components of EGDI to search through (datasets, project vocabularies, document repository, etc.) are considered as resources and to select subsets an API will be used. In this way it will be possible to add new resources in a later stage.

The search system will use the metadata from the EGDI metadata catalogue (run on MIcKA application) to discover the datasets. Through the APIs, the system will be able to select and display:

- Records within datasets stored at the EGDI database.
- Documents from the document repository: unstructured data (e.g. pdf files) and their metadata.
- Concepts from project vocabularies, being able to navigate through the concept hierarchy (broader/narrower concepts).
- The EGDI Keyword Thesaurus.

The system is supported by the EGDI multilingual thesaurus, which allows to enrich the search and find the datasets that are of interest to the user.

3.1 SOLR search engine

The Apache SOLR search engine will be used as a background tool for searching through the content of EGDI's document repository and their metadata.

The SOLR search system will be accessible as an API service, which will require a string as query parameter composed with special rules. The semantic search and spatial search with bounding box will be possible and ranking of search results based on relevance will be provided. Regarding querying parameters in string, the system will be able to search through several or all metadata elements and/or content. For the cases where the document is not stored in the EGDI document repository but available through a DOI link, the SOLR engine will be only able to search through the metadata elements provided for that entry in EGDI document repository.

The SOLR search engine will also support highlighting, Boolean operators (AND, NOT, OR), required operators, prohibited operators and grouping terms. The matched data will be returned in a JSON string with elements for all matching documents. Return elements will include the link to the original document.

4 DEVELOPMENTS ON THE 3D DATABASE

The EGDI3D database is installed in EGDI in a separate PostgreSQL database with PostGIS and PointCloud extensions. This database has the core functions from a 3D database (GEUS3D) developed by GEUS during the latest 5 years.

EGDI3D can store any geological model that a GSP would like to provide. The geological models will be available for the general public to discover, view and eventually for downloading. The EGDI3D database stores:

- Information on each geological 3D model, such as the location and purpose of the model, when it was made and by whom.

- Geological description of each geological feature of the model in the form of keywords, describing the geologic event that shaped the feature i.e., the event environment, the event process and the geological era in which the event took place. Additional information can also be stored on the type of geological feature, such as lithology of a geologic unit.

- Geometry of each geological feature. Geometries are produced by modelling tools in many proprietary formats. If a modelling tool can export modelled feature geometries in ASCII formats, EGDI3D database is able to import this and store any geometry from it in a common storage format.

GEUS has developed (or will develop) the following software for EGDI3D:

- A database model suitable for storing 3D models

- Services to read from and write to the database

- A program to read various ASCII geometry files for import

- A prototype of a web-based viewer, to be extended by the GIP project. This viewer should be able to show a legend with links to the project vocabulary if relevant.

- An administration module (work in progress) that will make it possible for the GSP to upload their own 3D geological models and add metadata about the layers of the models. Note that in order to upload a 3D model, the user must first create a metadata entry in EGDI metadata catalogue, describing the model in other to link it to that metadata entry during the upload process.

5 DEVELOPMENT ON A 3D VIEWER

There has been contacts with both the Polish and Austrian geological surveys (PGI and GBA) about using their 3D viewers. GEUS has also developed a simple viewer as a proof of concept. However, it is not yet decided what will be the final 3D viewer in EGDI. Currently, it is possible to make 3D models available through the EGDI3D database by a set of REST services, so external viewers can access the models directly from the database. This interface will be publicly available.

Ideally, both the Polish and Austrian surveys will make it possible to show models from EGDI in their viewers. However, this is still uncertain. Discussions are still ongoing.

The list of requirements for the 3D viewer are rather long (see D2.3.1 and D2.2.2) and it might not be possible to fulfil all of them. The requirements from the GSPs on this regard are listed in the table below.

Requirements from GSPs	**GIP-P comments**
Handling and displaying 3D models.	The EGDI3D database in EGDI handles import and export of 3D models from different text formats. The models are exposed through a basic REST interface.
Displaying virtual logs through models.	Hopefully, the ability to display virtual boreholes will be made by an external viewer from either the Austrian or Polish surveys.
Virtual cross section	Hopefully, the ability to display virtual cross sections will be made by an external viewer from either Austria or Poland
Virtual (horizontal) slice.	Hopefully, the ability to slice horizontally will be made by an external viewer from either Austria or Poland
Handling uncertainty.	This can be handled by colouring objects based on calculated uncertainties
Compass.	Both GBA and PGI have compass-like features in their 3D viewers. So, it should be possible, if they choose to contribute
Show camera direction.	In the viewer there must be axis / a compass showing the view direction

Colour / Alpha mapping functions to render attributes.	The 3D database will be able to store attributes for colour/alpha mapping
Glyphs for data representation.	It can be handled in some amount but not in huge amount (e.g., a BIM). That is, the system can handle the data, but displaying a whole BIM (Building information modelling) will not be possible.
Visualize different models at the same time.	Two models can be visualized simultaneous in two separate windows. This should be possible in EGDI, but we cannot promise to visualise them in the same viewer though. From the map view, it will be possible to create a virtual borehole describing the content of more than one model.
Possibility to display objects	The EGDI 3D model database can store points, surfaces and closed volumes.
Grid lines.	Depending on which viewer the EGDI3D viewer will be based, gridlines will be possible.
Exploded views of detailed part of 3D model (like the Polish viewer)	It provides details on parts of the model while the user is navigating across it. This visualization renders complicated models in a way that different parts do not occlude each other. This will be possible if the Polish viewer will be extended to display models from EGDI.
Change transparency for layers / surfaces in the 3D viewer	This functionality is available in both the Polish and Austrian viewers.

6 EXTENSIONS TO THE ADMINISTRATION MODULE

The administration module will be the future data entry point for all data coming into EGDI. It must be able to perform the following functions:

- To upload spatial data (GeoPackages and Shape Files).
- To register external services.
- To upload unstructured documents (reports, pictures and tabular data).
- To register documents stored elsewhere using DOI’s.
- To Upload 3D models.
- Entry of metadata for objects in 3D models.
- To handle users from different projects and make sure that users from one project cannot tamper with data belonging to other projects.
- To ensure that data uploaded to the platform have valid metadata in EGDI’s metadata catalogue.

6.1 User management

It has been decided to use the PostgreSQL databases user management system as the user management system for the administration module of EGDI. This means that, in order to be able to upload data to EGDI, a user must be created in the database. All users will belong to a user group, reflecting the different projects (e.g. a HIKE group). Users belonging to the same group will be able to edit data belonging to their project. This simple approach has been selected as only a few users for each project are expected.

Only users belonging to one or several GeoERA projects will be able to use this tool. All data uploaded by a user (spatial data, configuration of maps, layers and services and unstructured data) will belong to a specific project and can only be altered by users belonging to that project. However, all data uploaded to the system will be accessible to all projects. This means that a project can use data layers defined by another project in their map viewer.

6.2 Uploading spatial data

The spatial data uploaded to the system through the administration module will be saved in project specific schemas. Before uploading a dataset to EGDI a metadata entry must be created in EGDI metadata catalogue. Indeed, every time a spatial dataset is uploaded to EGDI database, the uploading system will require the establishment of a link to the pertinent metadata file of EGDI metadata catalogue (MIcKA) that describes that dataset.

It must be possible to upload spatial data as GeoPackages or Shapefiles. Upon upload of spatial datasets, the data will be stored in spatial tables in a database schema belonging to the project.

6.3 Document repository

The GIP-P is creating a document repository to where unstructured data can be uploaded (see chapter 7 of this report). The document repository will be able to archive, among others, documents, pictures and tabular data from every project. These files must, as the spatial data, belong to one or several GeoERA projects.

6.4 Registering services

If the data are delivered via web services, the user (i.e., the person who have logged into the administration module) can register the services and make them available as layers in the EGDI platform.

As with the spatial data, the services must also be linked to metadata entries previously created in the EGDI metadata catalogue (MIcKA).

6.5 Upload of 3D models

Most of the modelling tools currently in use are commercial products with their own proprietary formats and with no common open source export format. Also, for some of the tools, it is only possible to export the geometries and not the metadata describing them. Therefore, the import of 3D models into EGDI consists of importing geometries in ASCII formats.

According to D2.2.2, the GeoERA projects will create 3D models in:

- Seequent Leapfrog
- Schlumberger Petrel
- Paradigm GoCAD
- I-GIS GeoScene3D

Geometries from these modelling tools can be exported on the following ASCII formats that can be imported into EGDI's 3D model database

1. Raster file formats
   - *.asc
   - *.grd
2. Tin file formats
   - *.obj
   - *.ts
3. Voxel file formats
   - *.xyz

Besides the geometries, project must provide the metadata for each mapped feature via the administration module. However, the metadata for the model as a whole must be registered in MIcKA. The metadata entries for whole models and their parts that must fill out in MIcKA are listed hereunder (sections 6.5.1, 6.5.2 and 6.5.3).

6.5.1 Metadata on whole 3D Geological Models

- Model name
- Model type (geological, hydrogeological and so on)
- Model purpose
- Model access constraints
- Model location country
- Model description
- Model owner
- Model spatial reference system
- Model vertical reference system

6.5.2 Metadata on each Mapped Feature / Model Part

- Name of mapped feature
- Geological description of mapped feature
    - Feature type
        - Geologic contact
        - Geologic unit
        - Geologic fault
    - Geologic event that created this feature (INSPIRE code lists)
        - Geologic age
        - Geologic event environment
        - Geologic process
    - Specific characteristics, for example
        - Lithology for geologic unit
        - Displacement for geologic fault
- Geometry of mapped feature
    - Geometry
        - Geometry type (point, line, surface, volume)
        - Geometry as pgpointcloud
        - Geometry description
    - Geometry modelling process
        - Modeller
            - Email
            - Full name
            - Institution
        - Modelling tool
            - Software company
            - Software name
            - Software version
        - Modelling result / Model file

- File name
- File date
- File location / href to ASCII file imported to 3D database
- File spatial reference
- File format / File extension

6.5.3 Metadata on Mapped Features belonging to each Model

- Link between model and mapped Feature
- Legend text for mapped feature
- Legend color for mapped feature

7 DEVELOPMENT OF A DOCUMENT REPOSITORY

Several projects have requested the possibility to upload different types of unstructured data. The GIP-P has thus decided to develop a simple document repository capable of storing reports, pictures and tabular data. All documents will be available via links from the platform. It has also been decided to keep the metadata for these unstructured data stored at the EGDI platform and not integrate them into the EGDI metadata catalogue (MIcKA). This has been done to make data entry as easy as possible, as some of the data can contain their own metadata stored within the file, which can be reused during upload.

In order to store these data in open formats, the document repository will support the following formats:

- Documents – pdf
- Documents by reference – DOI
- Pictures – jpeg, jpg, png and tiff
- Tabular data – CSV (comma separated files)

For documents for which the projects have no ownership, it must be possible to register them through a DOI, so that it can be accessed through the portal. Documents with DOI links will not be stored in the repository. Therefore, they will be searchable through their metadata, but not through their content.

For the different types of unstructured data, there will be different types of metadata. The user interface will reflect these different demands for metadata. The metadata will be stored in the repository schema of EGDI database and, when possible, it will also be written into the document (mainly in PDF). The metadata will be used by the search system to find the documents.

When a user from a project uploads a file into the EGDI document repository, the user will get a static URL to the document. This URL can be used to link to the document to specific spatial data shown in the EGDI portal, to the project vocabularies or elsewhere. The final URLs have not been created yet, but it will be similar to this one: http://repository.europe-geology.eu/projectName/fileUUID. In this way, the unstructured data will be available through semantic search from the search system. To be able to make readable URL's, special characters (characters not in a-z, A-Z, and 0-9) will be replaced with URL friendly characters.

The EGDI document repository will also allow users to retrieve a list of "documents" and filter this list by project and type of document. The list will be available both at a homepage and as a REST service.

8 INTEGRATION OF VOCABULARIES AND KEYWORD THESAURI

The EGDI platform has been extended to integrate Keyword Thesauri and project vocabularies in order to search and link data and scientific concepts across GeoERA.

8.1 Keyword Thesaurus

Search for data is the basic task for all data infrastructures. To enable a comprehensive and powerful search functionality, all keywords used to tag datasets must be collected into a single hierarchy like a thesaurus. Data queries can then use this 'word net' to get search results for similar keywords within a certain "semantic radius".

For metadata descriptions, the Keyword Thesaurus will help to clarify the meaning of textual attributes, as well as to enable semantic search functionality within the metadata catalogue.

For the GeoERA projects, the use of the GeoERA Keyword Thesaurus developed by GIP-P WP4 can be used to store keywords, which can be tagged to the data produced by GSPs (see D4.2). This will facilitate the work of the EGDI search system.

8.2 Project Vocabularies

Project Vocabularies are collections of controlled dictionaries containing essential information about scientific concepts relevant for a project. The primary goal is to support projects and datasets with linguistically labelled terms. Project Vocabularies provide stable and reusable links to concepts (units of thoughts) that can be referenced whenever unambiguity is important. Behind such links alternative names, translations, definitions synonyms and additional information about other related concepts are made available. In any situation when something must be unambiguously named, a concept from a Project Vocabulary can be used. A Project Vocabulary can facilitate search and information access in a linked data environment. The GIP-P thus encourages all projects to create projects vocabularies.

8.3 Knowledge system

The EGDI knowledge system consist of a website that sends SPARQL queries to the keyword thesaurus SPARQL endpoint in order to find related keywords that can be used to retrieve related documents. The response consists of a list of keywords that can be used to query the document repository and return the relevant documents.

A raw version of this functionality has been tested by WP7, who reported that a website must be built around this functionality to provide a user interface. This interface is currently being built and tested for 9 categories of interrelated subjects that are defined in the HIKE project, which will be hosted in the keyword thesaurus triple store. A document management system with a web-API is essential for this to work. This will be handled by the EGDI document repository.

8.4 URI design

All scientific concepts published in the "GeoERA project vocabularies" framework will be accessible and online available in the Semantic Web and in a sustainable and long-term storage. The GIP-P will create an ID (URI, a resolvable HTTP web address) for each single term to index project data. It will be possible to integrate all this information (e.g., multilingual translations) live in a web application, project portal or a simple webpage via a web service (SPARQL endpoint).

Once published together, the terms and their global identifiers (URIs) cannot longer be deleted, being permanently available and resolvable. The GIP-P project team agreed on a domain name – https://data.geosience.earth (/ncl/) – to be used for URIs. The structure concerning subdomains and paths will be further discussed and decided by GIP-P WP5, as a "URI naming policy" or recommendation, in order to unify the way vocabs and concepts are named (camelCase writing, number of levels in the URI path, etc.). This solution will be applied not only for the GeoERA project vocabularies, but also for the GeoERA Keyword Thesaurus editions.

There is a need for a responsible organization that ensures a long-term online availability of the project vocabulary concepts. This organization also ensures an access to a detail page (website) to get human readable information (browsing facilities for Linked Data included) and machine-readable information by creating an RDF suffix to the URL. It is the central contact point which has to agree when a project creates new concept URIs too.

8.5 Publication of Project Vocabularies

The GeoERA projects decide whether there is a need to publish scientific concepts (with obligatory bibliographic references and INSPIRE mappings if applicable) by using semantic web and Linked Data. The publication, so called project vocabulary, is put online at the European Geoscience Registry, which is a Linked Data Registry software installed on a Jena triple store plus Sparql endpoint operated by BRGM. This is done by using the base URI https://data.geoscience.earth/ncl/geoera and numbers for concepts.

In a later stage, the project vocabularies or parts of them (published in separate SKOS concept schemes = registry lists) could be used to officially extend INSPIRE codelists via the federation of registries. However, the workflow or governance to extend INSPIRE codelists directly or via preceding project vocabularies is out of the scope of the GIP-P.

9 EXTENSIONS TO THE CENTRAL DATABASE

The central PostgreSQL database will be used to archive spatial data, unstructured data, data for the search system and data used to set up maps and map layers. The data will be divided into schemas so that the data for the unstructured data will be stored in a schema called repository and all project data will be stored in a schema named after the project.

9.1 User management

The user management have been setup to use a PostgreSQL's user management. Project users will be allowed to store data in their own schema and add datasets and services to their own maps.

9.2 Project schemas

All project will have their own schema, where all their spatial data will be uploaded to. The structure of the data in these project schemas will match the structure of the uploaded GeoPackages and/or Shapefiles.

10 MAKING DATA AVAILABLE THROUGH SERVICES

The EGDI system can be used to store data from different projects. Once in the EGDI database, the data will be accessible at the webGIS interface and from WMS and WFS services. Users will also be able to download the data from the platform.

10.1 Making data available as WMS and WFS services

To make available both the datasets stored in EGDI and the datasets delivered to EGDI by services for use in the GIS portal, the dataset can be configured to make the layers available as a WMS and as a WFS (if applicable) from the web GIS user interface and from a list of services. The datasets will be made available as WMS, WFS and in Shapefile format by the systems MapServer component. The GIP-P is also exploring the possibility of making available the data in GeoPackage format, although it is not yet clear whether that will be possible or not.

10.2 REST services

For some of the systems, data will also be available through REST services. These REST services will be served by the PostgreSQL addon PostgREST.

10.2.1 Map and layer configuration

All layers, all maps and all couplings between layers and maps will be made available as REST services.

10.2.2 The document repository

For the document repository, it must be possible to get a list of all unstructured data. It must be possible to filter this list on project and on type of unstructured data. For example, it should be possible to get a list of all documents uploaded by a specific project. This list should include metadata about the document and a download link to the document.

10.2.3 The 3D viewer

The 3D viewer reads directly from the 3D database when a user wants to preview the model. The model will be retrieved through a REST service delivering the model setup and binary REST service delivering the geometries.

10.2.4 Download services

The spatial data delivered to the EGDI platform should be available for download. It has not yet been decided how to do this. There are several possible options:

- Download from WFS
- Download from ATOM feed
- Download in Shapefile / GeoPackages format
- 3D models will be made available through REST services

11 CONNECTION TO THE EGDI METADATA CATALOGUE

The EGDI metadata catalogue (https://egdi.geology.cz/) uses the MIcKA application system, which follows international standards (INSPIRE TG 2.1x, ISO etc.). Metadata uploaded to EGDI must follow the requirements defined in the EGDI metadata profile, which are described in GIP-P’s deliverable D7.1.

All datasets available through EGDI must have metadata stored in the EGDI metadata catalogue. The ways how these can be created are explained in a cookbook shared with all geoscientific projects in April 2020. That document provides specific information on how to create metadata entries in MIcKA, both by manually editing the entries in EGDI metadata catalogue, or by providing the metadata via services for harvesting.

Note that when services are set up by the EGDI platform to make data available as WMS and/or WFS services, the system must create metadata for these services in the EGDI metadata catalogue. These metadata must also be updated when the services are updated or changed. This should be done automatically.

12 EXTENSIONS TO THE HARVESTING SYSTEM

The EGDI harvesting system is currently provided for the following groups:

- Minerals data harvesting (M4EU harvesting) and
- Metadata harvesting (MIcKA harvesting).

With the extension to the harvesting system, we will develop, upgrade, improve and optimize the current EGDI harvesting systems and add the following module(s):

- Mineral yearbook data harvesting (MYB harvesting)

- Spatial data (and other data) harvesting. This extension will be implemented only if it is specifically required by one or several GeoERA projects.

12.1 M4EU harvesting improvements and extensions

Here below, we list the currently known requests from the GeoERA projects:

- **To implement a solution to removing a data gap in the harvesting process:** the harvested database must have no data gaps. Currently, when harvesting has started, it is not rare that some providers harvesting fails. These failures happen, for example, when the provider's servers are not online, when the server is too busy and not responding correctly, due to errors because uncorrected data entry by the provider or owing to errors because the provider modify its data by following a ETL process at the harvesting time. In the occurrence of any failures listed in this paragraph, the system will keep the last successful harvesting data for those providers.

- **To implement a report for a quick checking of the harvesting results:** after harvesting is finished, currently, only the database dump is provided. This is insufficient; it should provide a report with an overview of the records harvested by the providers. In that report, the dates for provider harvesting should be shown.

- **To implement a solution for geometry checking:** if the provider enters the geometry data incorrectly, the data cannot be shown on the map. We need a tool for geometry checking and reporting irregularities.

- **To provide a support for more than one data provider per country:** because one country could have more than one provider, we need to modify m4eu database to add a country code and/or data provider acronym name to each table. In that way, we can assure the correctness of the data source

- **To provide a solution for minerals service monitoring:** currently, providers do not know if some of the required services are up or down. The harvesting process fails if the services are not up. By providing a service monitoring, the provider will immediately be aware of the problem and will be able to fix it. With the help from the monitoring services, the harvesting process will not start if services are not up and running

- **To provide a better logging**: currently, the harvesting process has one big log file, with a non-transparent content. With the current logging, it is impossible to see the history of logging for a provider. A better and more transparent logging is desirable for each data provider

- **The improvement of the current harvesting code:** the harvesting results are sometimes or somewhere strange or some data appear not to be harvested. In order to use the data, EGDI needs complete and correctly harvested datasets. Therefore, an extensive testing, discovering bugs, a code debugging and fixing is required for the harvesting process. Improvements in the harvesting code includes also all requested improvements in the database, code lists, views, services, etc.

- **To provide tests with providers:** providers must help to improve the harvesting process by comparing their source database with the harvesting results and report all identified differences to harvesting@geo-zs.si. They must also send information on data change in their databases. Any missing tables, differences in records count and different field values must be discovered and corrected.

- **To check for stable INSPIRE ID fields:** an identifier should correspond always to the same deposit and should not be changed by the provider. Providers should normally change just the version ID and the begin/end life spam. Providers usually change the data by ETL process, which first delete all their data in their database and create a new set. In that process, providers generally do not take into consideration stable INSPIRE id fields. We cannot just trust that stable INSPIRE id‘s were provided. We need an automatic mechanism to check and report deviations, so checking should be added to the harvesting process in some way.

- **To provide a history information for number of data records for provider:** To track a progression in amount of data provided by a provider, we need a solution for showing the historic for the number of harvested records by the m4eu database tables. This could be solved by creating a special table or database with the harvesting history data (history of count reports), which could provide us a provider progression (history of number of data entries in each table).

- **Option for recreating a provider local database:** that is, to provide a solution to data providers, so they can recreate their harvested database in their site. This solution will provide an additional option to providers to check if their database was harvested correctly.

- **Add m4eu version log table with service:** Information about the m4eu database version, code list version and views version. Maybe, we could also have a record where users can put their last database modification date. All that information must be provided by services. That information can be checked by the harvesting process.

- **To provide an information for the latest database modification by the provider:** it is not important only when the last harvesting was provided by the country, but also when a provider modifies the database. Providers will be required to manually change some fields in their database after each data update.

- **The integration of the next version of M4EU database:** The ORAMA project gave some recommendations to improve the M4EU database, the software versions of Java Developer Kit (open source), Apache Tomcat, Deegree3, and PostgreSQL and the mapping of the WFS-services. Many of the ORAMA recommendation has now been tested, but some still need to be tested before they will be implemented in the next version.

  Differences between current M4EU DB v1.1.2 and next version of M4EU DB:

  - 2 new DB tables 'commoditygrouptype' and 'totalproduction'
  - Updates of more codelists tables especially tables 'commoditytype', 'UNFCategoryType' and 'WasteTypeType'
  - Updates of INSPIRE schemas

12.2 Mineral Yearbook (E-MYB) harvesting extensions

The GIP WP7 is developing a program for mineral yearbook data harvesting from data providers. Their data will be provided with WFS 2.0 services. The whole E-MYB data and data transfer are part of the Mintell4Eu project and not of the GIP-P. At the end of the GIP-P and Mintell4EU projects, everything will be part of EGDI. For more information, see Mintell4EU *D5.3.1 Specification of steps needed for the integration of the E-MYB in the M4EU DB*.

The ORAMA project modelled a new E-MYB database using the same software versions as M4EU DB and developed the mapping for the WFS-services. Many of the ORAMA recommendations have now been tested. However, there are still some of them that need to be tested before they will be implemented in the next version.

M4EU YB mineral yearbook:

- Updates of more codelists tables especially tables 'commoditytype', 'UNFCategoryType'
- E-MYB DB including codelists

12.3 MIcKA harvesting extensions

The GeoERA projects requested the following harvesting functionalities from the EGDI/MIcKA metadata catalogue:

- harvest metadata from CSW

- making MIcKA compatible with other EU inventories.

The option of harvesting metadata from external CSW was already available in previous versions of MIcKA. Hence, this request is met by the extended EGDI metadatabase, which sends a query to the remote catalogue using GetRecords and GetRecordById via post request. However, in order for MIcKA to be able to harvest metadata via standard CSW, the export from the source catalogue must be built following an ISO 19139 structure. That also applies for other EU metadata inventories. MIcKA will be able to harvest metadata from any remote catalogue that have an ISO structure and provide standard CSWs that follow OGC Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile (1.0.1).

More information about this can be found in GIP-P's deliverable D7.1 and the GIP-P internal report "Cookbook for creating metadata records using the EGDI Metadata catalogue (MIcKA, version 6.0)".

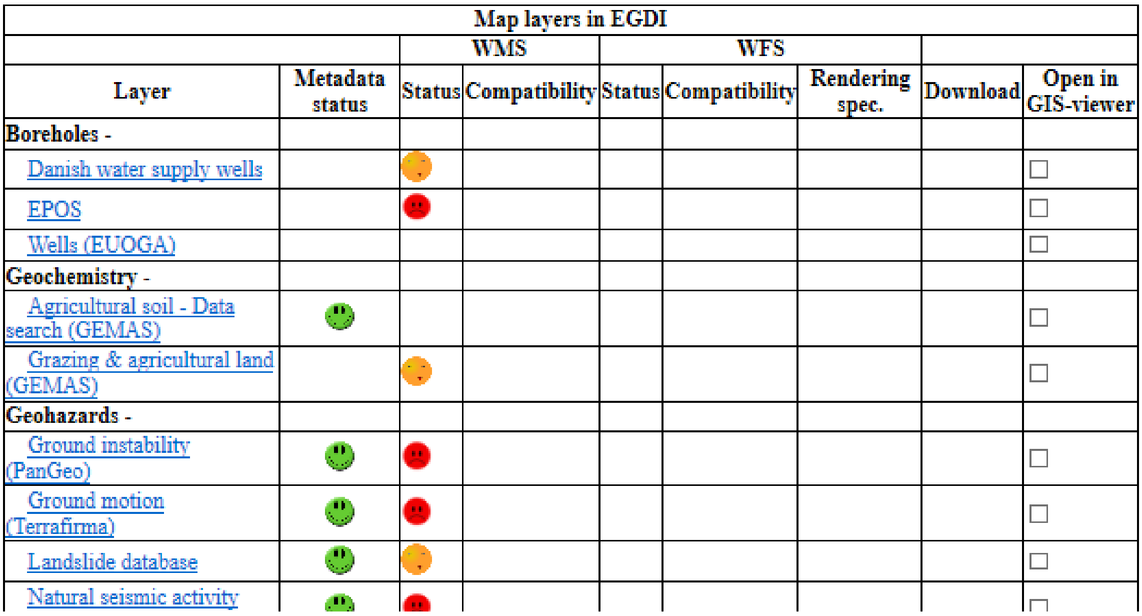
13 MONITORING SYSTEM

In order to monitor the different parts of the EGDI system and the web services delivering data to EGDI, a monitoring system must be developed.

The GIP-P has decided to use the Zabbix system for the monitoring and to develop a monitoring system where the current status of the system can be seen. The status of the system will be shown as a set of smileys going from green (all OK) to red (not working) for different parts of the system and services. This system must show information about metadata, WMS and their capabilities, WFS and their capabilities, and the link to download.

All layers (web services) that are registered at the EGDI portal will be added to the Zabbix monitoring system managed by GEUS. The Zabbix system calls each service with a given time interval and logs the result of the call. The result will show whether or not the service replied and how fast the answer was.

The user interface of the monitoring system will consist of two web pages coded in JavaScript. An overview page will show a list of all layers registered at the EGDI portal. This list can either show all layers grouped by the project that made them or grouped in the same way as the layers are grouped in the "All content" map at EGDI. For each layer, there will be fields showing whether the service is currently available (this information is retrieved from Zabbix). A column showing whether the layer has metadata registered at the EGDI metadata catalogue, and some columns showing different specifications of the layer. The suggested columns can be seen in the screenshot below.



Map layers in EGDI								
		WMS		WFS				
Layer	Metadata status	Status	Compatibility	Status	Compatibility	Rendering spec.	Download	Open in GIS-viewer
Boreholes -								
Danish water supply wells		🟠						□
EPOS		🔴						□
Wells (EUOGA)								□
Geochemistry -								
Agricultural soil - Data search (GEMAS)	🟢							□
Grazing & agricultural land (GEMAS)		🟠						□
Geohazards -								
Ground instability (PanGeo)	🟢	🔴						□
Ground motion (Terrafirma)	🟢	🔴						□
Landslide database	🟢	🟠						□
Natural seismic activity	🟢	🔴						□

It will be possible to select a number of layers on the overview page and have them shown in the EGDI map viewer.

The other part of the user interface will be a web page showing detailed information about a given layer. This will include all the registered information about that layer in the EGDI portal. It will include links to the full metadata record, archived at the EGDI metadata catalogue, and to the getCapabilities page of the service. The detail page will be accessed by clicking on a layer in the overview page.

REFERENCES

D 2.2.1: First report describing the requirements to the Information Platform by the Geo-energy, Groundwater and Raw Materials themes. January 2019. https://geoera.eu/wp-content/uploads/2019/01/D2.2.1-Requirements-to-the-Information-Platform.pdf.

D2.2.2: A second report refining the requirements after feedback exchanges related to the prototypes of the EGDI database and the display interface. January 2020. https://geoera.eu/wp-content/uploads/2020/01/D2.2.2-Refinements-of-requirements.pdf; https://geoera.eu/wp-content/uploads/2020/01/D2.2.2-Appendix-A.pdf.

D 2.3.1: First report mapping and describing the needed extensions to EGDI directly related to the task 2.2. March 2019.

D 2.1.1: First report highlighting the potential synergies and overlaps between the projects in terms of geoinformation. June 2019. https://geoera.eu/wp-content/uploads/2019/07/D2.1.1-Potential-synergies-and-overlaps.pdf.

D4.2: Keyword Thesaurus. October 2019. https://geoera.eu/wp-content/uploads/2019/11/D4.2-GeoERA-Keyword-Thesaurus.pdf.

D4.3: GeoERA project vocabulary. October 2019. https://geoera.eu/wp-content/uploads/2019/11/D4.3-GeoERA-Project-Vocabularies.pdf.

D7.1: Working version Metadatabase. December 2019. https://geoera.eu/wp-content/uploads/2019/12/D7.1-Working-version-Metadatabase.pdf

Kramolišová, P., Kondrová, L., Moravcová, O., and Kafka, Š.: Cookbook for creating metadata records using the EGDI Metadata catalogue (MIcKA, version 6.0). April 2020.

Mintell4EU D5.3.1 Specification of steps needed for the integration of the E-MYB in the M4EU DB.