Deliverable 7.3

Final version of Central database / harvesting

Authors and affiliation:

Blaž Bahar GeoZS,
Jernej Bavdek, GeoZS,
Maks Šinigoj GeoZS,
Andrej Vihtelič GeoZS,
Bjarni Pjetursson GEUS,
Frands Schjøth, GEUS,
Jonas Thyregod GEUS,
Martin Hansen GEUS,
Viktor Søgaard Rasmussen GEUS,
Feliachi Abdelfettah BRGM,
Roquencourt Jean-Baptiste BRGM,
Hegen, D. (Dries) TNO,
Huisman, E. (Erik) TNO,
Sohier, P. (Paul) TNO,
Dana Čápová ČGS,
Pavla Kramolišová CGS,
Ángel Prieto Martín IGME
Héctor Sánchez Molinero IGME.

[BENEFICIARY]

E-mail of lead author:
andrej.vihtelic@geo-zs.si

Version: 26-03-2021

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	D7.3	
Dissemination level	Public	
Deliverable name	Final version of Central database / harvesting	
Work package	WP7, Developments (central)	
Lead WP/Deliverable beneficiary	GeoZS	
Deliverable status		
Submitted (Author(s))	18/03/2021	Andrej Vihtelič
Verified (WP leader)	26/03/2021	Andrej Vihtelič
Approved (Coordinator)	09/04/2021	Jørgen Tulstrup

TABLE OF CONTENTS

1. DEFINITIONS 4
2. INTRODUCTION 5
3. CENTRAL DATABASES AND STORAGES 6
3.1 EGDI database 6
3.2 EGDI 3D Database 6
3.3 Mineral Inventory Database 6
3.4 M4EU System Info database 7
3.5 Metadata Catalogue Database 7
3.6 Solr Cores 7
3.7 European Geoscience Registry Jena Triplestore 7
3.8 Digital Archive (filesystem) 7
4. HARVESTING 8
4.1 Mineral Inventory Data Harvesting (M4EU harvesting) 8
4.2 Search System harvesting 8
4.3 Metadata Catalogue harvesting 8
5. OTHER COMPONENTS 9
5.1 EGDI Portal Main Entry Point 9
5.2 Web GIS 9
5.3 3D Viewer 9
5.4 Search System 9
5.5 Feature Distribution Viewer 9
5.6 Feature Distribution API for PostgreSQL 9
5.7 Feature Distribution API for Solr 9
5.8 Solr Search Engine 10
5.9 Repository Search 10
5.10 European Geoscience Registry 10
5.11 Metadata Catalogue 10
5.12 Harvesting Info 10
5.13 Monitoring System 11
5.14 Administration Module 11
6. RELATED DOCUMENTS 12

1. DEFINITIONS

Below there are several definitions, conventions and abbreviations that are used.

Application programming interface (API): a computing interface that defines interactions between multiple software intermediaries. It defines the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conventions to follow, etc.

Backend: Front end and back end describe the layers that make up a computer program or a website which are delineated based on how accessible they are to a user. The back end refers to parts of an application or a program's code that allow it to operate and that cannot be accessed by a user.

Codelist: Predefined list from which some coded attributes take their values. The purpose of a codelist is to ensure that data values comply with controlled terminology. The items in a codelist usually have an URI as an identifier.

Connector: Web API allowing the Search System to search inside a resource. All connectors have JSON files with a defined structure as input and output. Different repositories may require different connectors.

Database: an organized collection of data, generally stored and accessed electronically from a computer system.

Feature: an object that can have a geographic location and other properties. Examples: A record in a spatial database or a document in a document repository can be considered as features.

Feature distribution: a special type of distribution created for the Search System. It shows the selected features from a resource obtained as a result of the query made by the user. Feature selection would not be available for all resources, probably only for the most relevant databases and repositories.

Harvesting: a process to automatically extract large amounts of data from websites.

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

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

Stopword: a word that is automatically omitted from a computer-generated concordance or index (such as the, is, at, which, and on. In this case).

URI: A Uniform Resource Identifier (URI) is a unique sequence of characters that identifies a logical or physical resource used by web technologies.

Web application (or web app): application software that runs on a web server. Web applications are accessed by the user through a web browser with an active network connection.

2. INTRODUCTION

Members of GeoERA Information Platform Project (GIP-P) team are developing and establishing a common platform for organising, disseminating and sustaining digital harmonised data from the GeoERA geoscientific projects on subsurface energy, water and raw material resources from all over Europe. The aim is that the data to a high degree will be Findable, Accessible, Interoperable and Reusable (FAIR) at one place and thereby be as valuable as possible for the stakeholders.

The Platform consists of web applications, databases, a digital repository, opensource tools and services that connect all together.

The GeoERA Information Platform is built as an extension to the EuroGeoSurveys (http://www.eurogeosurveys.org/) European Geological Data Infrastructure (EGDI, http://www.europe-geology.eu) which will care for its future sustainability.

This document focuses on the backend part, especially on the central databases and harvesting system of the EGDI that 'WP7 development-central' team of the GIP-P has been able to achieve and give a basic overview of the whole system. It is intended to give an overview of the backend part of the system and describes the system as it was at the date of deliverable.

The full version with in-depth information, which also includes description of services, is prepared as an internal document, available only for GIP-P development team members and its intention is to support EGDI IT teams to maintain the backend part of the system.

3.CENTRAL DATABASES AND STORAGES

In the following paragraphs the basic information for components with data (databases, solr cores, triplestores and storages) used by new EGDI system is described.

3.1 EGDI database

The EGDI database is used by EGDI Web GIS, EGDI Admin, Search System and Solr. The data in the database are divided into schemas which contain:

- Information needed to define the different maps and layers in the system. This includes configuration of the map, which layers to include on the map and how to arrange them.
- Location for the vector data uploaded by the projects through Shapefiles or GeoPackages (GeoTiffs are saved on the EGDI file system). Each project has their own schema (e.g. "project_hike") and each uploaded data set is represented by a table in the schema.
- Metadata for unstructured documents (pdf, jpg, csv etc.) and metadata for doi-links. The main source for this data is saved in Solr cores. The data in the database is stored for restoring Solr cores.
- Data which are exposed by PostgREST tool, so that data is available as REST web services.
- Search system data (resources metadata, configuration of catalogues, feature distributions, data for selecting areas in spatial search, replication of GeoERA thesaurus for optimizing text string processing with suggester and autocomplete functionality, auxiliary data).

3.2 EGDI 3D Database

The EGDI 3D Database is used by the EGDI 3D Viewer and EGDI Admin.

The data in database is divided with schemas which contain:

- 3D datasets tables, views and functions for PostGIS and pgPointcloud extensions.
- Metadata, geometry and geological description for the 3D models.
- Log data for all performed inserts, updates and deletes.
- Routines / functions for building point clouds from raster- tin- and voxel data, and temporary data calculated during import of geometry file sources.

3.3 Mineral Inventory Database

The mineral inventory database contains:

- Mineral inventory data which is collected from all data providers through harvesting process.
- Mineral yearbook statistical data relating to the mineral resources and reserves for country by years.

3.4 Mineral InventorySystem Info database

The mineral inventory system info database is used by harvesting info web application. It contains additional data concerning harvesting as information about available data providers and their installed versions of mineral inventory databases, harvested databases, alerts that occurred during harvesting and its status, records count history.

3.5 Metadata Catalogue Database

The Metadata Catalogue database store metadata which is managed by the EGDI Metadata Catalogue web application.

3.6 Solr Cores

The Solr cores contain indexed data i.e., searchable metadata for documents, images and data (csv files) and also searchable content in case of pdf files. 'Repository Search thematic application' and 'Search System' displays this data of performed search by Solr.

3.7 European Geoscience Registry Jena Triplestore

The European Geoscience Registry (EGR) triplestore is used by Metadata Catalogue, Search System, Repository Search application and EGDI Admin application. It contains the following registers with triples for:

- Keyword thesaurus: Used for tagging GeoERA project datasets,
- Project vocabularies: Collections of (linguistically labelled) scientific concepts. They can also be understood as an initial part of a future EGDI knowledge graph.
- Project specific codelists: Their goal is to provide a reference for the project to control the values used in the data.
- Projects codelist: It is a codelist gathering some details of the projects linked to GeoERA and other geoscience scopes.
- Organizations codelist: It is a codelist gathering some details about the organizations involved in the different projects of GeoERA.

3.8 Digital Archive (filesystem)

The EGDI Digital archive is the Linux filesystem used by EGDI Admin, the EGDI Web GIS and the Solr System. The filesystem is used to store unstructured documents like documents (pdfs), images (jpeg files), data (csv files) and raster data (GeoTiff). The data is stored as files in the Linux file system.

4.HARVESTING

4.1 Mineral Inventory Data Harvesting

The Mineral inventory data harvesting system collects the minerals inventory data from data providers local databases into one common harvested database. The harvester program sends requests to data provider services and collects, validates and stores data from responses into harvested database. The harvested database is then transferred to the central mineral inventory database (Chapter 3.3) but only if it is without gaps.

Data gaps in harvested database can occur when the providers' servers are not online, when the server is too busy and not responding correctly, due to errors because of the incorrect data entry by the provider or because data are modifying data by provider at the harvesting time. But harvested database must have no data gaps to show raw mineral data on Web GIS (Chapter 5.2) site.

Additional functionalities provided by GIP-P project are: supporting multiple providers by country, improved logging, geometry checking, providing a quick check of harvesting results, tracking data amount progression for providers, removal of data gaps, providers service monitoring, providing remote check for providers database (provider service ability, last date of data modification, updating status, counting provider data table records, number of unsatisfied xor constrains).

4.2 Search System harvesting

To complete/update information in the database that supports the Search System (Chapter 5.3) the harvesting is used for occasionally getting the data from:

- Metadata Catalogue for searching and describing resources and
- European Geoscience Registry keyword thesaurus to enrich search.

4.3 Metadata Catalogue harvesting

In the EGDI Metadata Catalogue (Chapter 5.11) harvesting is used for occasionally getting the providers remote catalogues metadata to complete/update information in the Metadata Catalogue database (Chapter 3.5). Metadata from remote national, project or other metadata catalogues are harvested through CSW service and there are 11 active remote catalogue providers.

Harvesting from remote catalogues and other sources can only be set by the EGDI Metadata Catalogue administrator on request from the data provider. It is possible to harvest just once (and update metadata manually) or set a regular harvesting interval (preferred option). Each harvesting session is documented by a harvesting report with a validation status that is sent to relevant contact points.

Metadata contact person from an organization that wants to harvest their metadata must send request to the administrator (egdi.metadata@geology.cz).

Harvesting is regularly processed, mostly every night. Only changes after the date of the last harvest will be processed.

5.OTHER COMPONENTS

5.1 EGDI Portal

EGDI portal is the main entry point which connects all web applications (Chapter 5) developed by GIP project. It is developed with WordPress.

5.2 Web GIS

The Web GIS is the component showing all the datasets (uploaded by the projects) as maps. On the map you can click the objects / cells and get access to their attributes. The Web-GIS System disseminates data products and data sets as online interactive maps including various tools to further perform data analysis (see D7.2 for further details). It uses the MapServer installation, which is installed at the same server as WebGIS, on the "EGDI database" and the "EGDI filesystem".

5.3 3D Viewer

Shows 3D models represented by point clouds and its metadata from central 3D database. The data is provided by PostgREST services from egdi3d database (Chapter 3.2).

5.4 Search System

The Search System allows the user to discover and access available geoscientific information, see their metadata, select and display subsets of features from those resources. It offers a possibility to perform spatial search, provides suggestions while the user types into the search field (suggestions are from European Geoscience Registry multilingual Keyword Thesaurus). The search string typed by the user is processed to improve and enrich the search (tokenization, removal of punctuation tags and stopwords, searching also for related terms and translations). The information is fetched from PostgreSQL databases and Solr cores and is shown in the detailed ranked list of results.

5.5 Feature Distribution Viewer

Displays feature distribution results of a search performed by using the search system web application.

5.6 Feature Distribution API for PostgreSQL

The feature distribution API for PostgreSQL creates a valid connector for any information stored in EGDI database (Chapter 3.1). It is a middleware between Search System (Chapter 5.4) and the PostgreSQL logic.

5.7 Feature Distribution API for Solr

The feature distribution API for Solr provides the connection between the Search System and Solr Search. It gets request from Search System (Chapter 5.4), converts it to Solr request, gets response from Solr Search Engine (Chapter 5.8) and converts it to a response for Search System.

5.8 Solr Search Engine

Solr Search is used by Repository Search application (Chapter 5.9) and through Feature Distribution API with System Search to perform search through document content and metadata for documents, images and data stored in Digital archive (Chapter 3.8). The Administration module calls Solr when the document is uploaded into Digital archive to perform indexing of the document's metadata (and content in case of pdf documents).

5.9 Repository Search

The Document Repository Search Thematic web application is used for searching through the items (pdf documents, doi documents, images, csv data and their metadata) uploaded through Administration module (Chapter 5.13) web application into Digital archive (Chapter 3.8). The searching is done by using Solr Search (Chapter 5.8). The application offers five different types of search: basic (through entered search terms), semantic (through semantically related words in European Geoscience Registry to the entered search terms), advanced (grouping terms, logical operators in user search, required and prohibited operators, search by certain field), UUID (fetch results of a stored search query condition by using PostgREST services) and spatial search (search based on a spatial component). Results are ranked to determine the most relevant results, search terms (and also semantically related terms) are highlighted.

5.10 European Geoscience Registry

European Geoscience Registry stores and publishes controlled vocabularies for the whole GeoERA program which supports multilingual semantic text search and project specific knowledge concepts. It is a Linked Data Platform based on Jena RDF triple store (Chapter 3.7).

5.11 Metadata Catalogue

The Metadata Catalogue is s a web application for management and cataloguing of structured, mostly spatial metadata. It uses the MIcKA system for management and publication of metadata on structured data and services. MIcKA technology enables entry, editing, harvesting, discovery, and view of metadata on geological data across Europe. It provides tools for compilation and export of the metadata in standardized formats. Metadata from remote national, project or other metadata catalogues are harvested from metadata catalogue through CSW service into metadata catalogue database (Chapter 3.5).

5.12 Harvesting Info

Harvesting info is a Django web application which supports data providers with an overview of harvesting results from mineral inventory system info database (Chapter 3.4). It also provides status checks of their mineral inventory services and with collecting data on resources, reserves and exploration as part of electronic yearbook (Online Minerals survey).

5.13 Monitoring System

The monitoring system lists and monitors if services that are listed on the EGDI portal with additional providers services for harvesting are up and running. For each service the status is shown with coloured smileys and additional detailed information is available. The monitored services are registered in the Zabbix system.

5.14 Administration Module

Registers and maintains the data sets, enables the upload of unstructured data, spatial data (GeoPackage, Shape files, GeoTIFF), NetCDF, 3D models, geological models (from LeapFrog, GoCad, Petrel, GeoScene3D). The metadata must be entered before upload. It is also possible to set up new maps and edit existing maps, define or add layers and groups of layers to maps.

6.RELATED DOCUMENTS

Apache Jena: https://jena.apache.org/index.html

Apache Solr: https://solr.apache.org/

D2.1.1: 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.

D2.2.1: 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.

D2.3.1: First report mapping and describing the needed extensions to EGDI directly related to the task 2.2. March 2019. http://geoera.eu/wp-content/uploads/2019/04/D2.3.1-Extensions-to-EGDI.pdf

D2.3.2 Mapping and describing the needed extensions to EGDI directly related to the task 2.2. April 2020. https://geoera.eu/wp-content/uploads/2020/04/D.2.3.2-Mapping-and-describing-the-needed-extensions-to-EGDI.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.

D6.4: Portal Version2. June 2020. https://geoera.eu/wp-content/uploads/2020/06/D.6.4-Portal-Version-2.pdf

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

D7.2: Finished testing the system and identifying problems. July 2020. https://geoera.eu/wp-content/uploads/2020/07/D.7.2-Finished-testing-the-system-and-identifying-problems.pdf

Deegree: https://www.deegree.org/

Django: https://www.djangoproject.com/

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.

MapServer: https://mapserver.org/

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

Zabbix: https://www.zabbix.com/