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

Deliverable 5.2
Preliminary version

GeoERA Central System specification

Authors and affiliation:
François Tertre, Abdelfettah Feliachi, Sylvain Grellet [BRGM]

Carlo Cipolloni [ISPRA]

Martin Schiegl [GBA]
Patrick Bell [BGS]

E-mail of lead author:
s.grellet@brgm.fr

Version: 10/10/2019

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	D5.2 – version 1.1	
Dissemination level	Public	
Deliverable name	GeoERA Central System specification	
Work package	WP5, Architecture	
Lead WP/Deliverable beneficiary	BRGM	
Deliverable status		
Submitted (Author(s))	08/10/2019	Sylvain Grellet
Reviewed	10/10/2019	Marc Urvois
Approved (Coordinator)	06/12/2019	Jørgen Tulstrup

GENERAL INTRODUCTION

The GeoERA Information Platform project will build a complex infrastructure as an extension of the European Geological Data Infrastructure (EGDI) gathering heterogeneous components together. All these components will have to form a homogeneous system consisting of a mix of distributed and centralised components. It shall lead to seamless user experiences for the GeoERA partners (from the GeoERA Scientific Projects - GSP) and also for all the potential users of the platform (researchers, European Commission, stakeholders, general public…).

In order to fulfil the objective to have the most integrated platform, all the components shall be well thought out. An aim is that their specificities, relations and interactions shall be written down to extract the best from former projects that have been made during the past years, but also to follow the recent development around the best practices of data on the web. But first of all, the components must be chosen so that it is ensured that the specific requirements of the GSPs are honored. These requirements are mapped by the GIP-P WP2.

Deliverable D 5.1 defined the blueprint of the GeoERA Information Platform Project (GIP-P).
This deliverable is a dedicated focus on the GIP-P central system. Most of the required functionalities have been introduced in deliverable D 5.1.
The purpose of the present document is to describe more precisely the required components and the reference implementations that will be deployed in the development focused WPs.
As such it is divided into 2 clear parts:

- the first one describes each component one by one and tries to identify reference implementations to start from,
- the second is an IT analysis of many EU and international projects in which some IT component have been tried/matured in order to identify which pre-existing brick/experience could be reused for the GeoERA Information Platform Project (GIP-P).

Remark: *This first version of the report was planned for delivery at M6, which turned out to be the same time for delivering the first WP2 report (D2.2.1) presenting broadly the user requirements collected for the fourteen GeoERA Scientific Projects (GSPs). The D2.2.1 was delayed, mainly because many of the GSPs were not ready in M6 to decide on what they will deliver. It was therefore decided to postpone this report (D5.1) until a clearer picture of the GSPs' requirements was available. As a result, the present version rather targets the description of an "ideal" spatial data infrastructure based on the state-of-the-art and best practices in the domain.*

In the context of EGDI extension, the architecture design will be refined during the upcoming prototyping stage (M6-M18) against the D2.2.1 contents as well as D2.1.1 (data sets produced by the GSPs) and D2.3.1 (expected new functionalities).

In the second version of this specification report planned before the end of 2019, elements will be kept and further specified. Others may be moved to one or more annexes. And some others will be added when new needs are identified in the D2.x.x user requirements reports. In addition, the next revision will take into account updates based on the testing of several innovative approaches as well as interactions between WP2, WP3, WP4, WP6 and WP7. They will also clarify how the EGDI can be scalable with automated processes.

Finally, which extensions to the current EGDI system will be implemented will be evaluated in January 2020 based on the knowledge at that time of the requirements from the GSPs (WP2) and the findings in WPs 3, 4, 5, 6 and 7.

TABLE OF CONTENTS

TABLE OF FIGURES

1 COMPONENTS AND ARCHITECTURE

Deliverable D 5.1 listed the following functionalities and defined their position in the architecture.

They are based on an OGC standard stack with

- a metadata catalogue supporting OGC CSW,
- features using WMS, and application schema compliant WFS 2,
- observations using SensorThings API part 1.

They are supported by

- a URI resolver,
- a codelist registry tool,
- a shared data specifications environment,
- a data publication alternative for the data providers that don't have the IT capacity/know-how,
- validation services,
- a harvesting system,
- data front-ends,
- a spatial database.

And they are complemented by the capacities to expose

- metadata content using JSON-LD and DCAT_AP,
- features using JSON-LD,
- features using WFS3 (in combination with JSON-LD if feasible),
- observations using WFS3 (in combination with JSON-LD if feasible),
- the Information Platform content in a SPARQL endpoint.

The following parts of this report go through all those elements one by one to identify suitable open sources implementations.
If the identified implementations meet their limit due to GeoERA Information Platform Project (GIP-P) specific requirements, GIP-P partners will try first to seek to provide enhancements to those implementations so that they can benefit to the broader community.

1.1 Metadata catalogue

A Metadata Catalogue is the central access point to metadata concerning data relevant to GeoERA. The metadata catalogue for EGDI (and therefore also for the GSPs of GeoERA) is MICKA.

Only digital and structured information (e.g. non-geographic and spatial datasets or dataset series and spatial data services as Web Map Services (WMS), Web Feature Services (WFS), multidimensional models or other digital products) are to be described by metadata in this catalogue. The catalogue enables discovery, view and use of geological data across Europe. It provides tools for compilation of those metadata in a standardized format. Metadata are freely

accessible to the public for viewing and searching, but inserting and editing is for authorized users only.

The main principles that should be followed in implementation are the following:

- The catalogue will support the distributed system of metadata administration and be fully compliant with the following international standards :
    - ISO 19115 Geographic Information: Metadata,
    - ISO 19119 Geographic Information: Services,
    - ISO/TS 19139:2007 Geographic MetaData XML (gmd) encoding, an XML Schema implementation derived from ISO 19115,
    - ISO 19110 Geographic Information: Methodology for feature cataloguing,
    - ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set,
    - Open Geospatial Consortium Catalogue Service for Web 2.0.2 ISO AP 1.0
- Metadata will be compliant with the latest INSPIRE Metadata Implementing rules: Technical Guidance for the implementation of INSPIRE dataset and service metadata based on ISO/TS 19139:2007, version 2.x (https://inspire.ec.europa.eu/id/document/tg/metadata-iso19139).
- Common vocabulary/registry/thesaurus of geological terms will be implemented and used as keywords source for metadata records, following the SKOS rules. The keywords are to be provided as URIs using gmx:Anchor element.
- There will be an option that the catalogue provides multilingual fulltext search. Bilingualism (English and national language) is an option widely used in national catalogues.
- Thesaurus-based search may be further developed to include the function of ordering by relevancy.
- The catalogue will improve the link between resources (data – web services based on them etc.).
- Semantic search approach should be implemented (eg. geographic regions set as uri/name/coordinates),
- The available output formats will be
    - ISO 19139, JSON, ATOM, KML, HTML, RDF, DCAT-AP/Geo-DCAT-AP
- Based on D5.1 analysis of W3C Data on the Web and Spatial Data on the Web working groups recommendations the catalogue shall also expose its own description and its content using JSON_LD to enable metadata indexing by search engines (at the time of writing DCAT-AP is not considered, at least by Google dataset search when indexing)
  Several exercises are currently being conducted on that aspect (ex : Canada, France, Belgium/The Netherlands, …) and the geospatial metadata community (through INSPIRE and also GeoNetwork) is identifying common patterns to be applied so that they can be coded within Geonetwork

Metadata catalogue should be more effectively integrated into the GIP-P and the main principles for adding data and metadata to the portals should be defined to follow these rules:

- All data available within GIP-P shall be supplemented by metadata – "Top level" metadata.
- It is preferred that metadata are harvested directly from permanently maintained national or ongoing project catalogues.
- Each finished project shall have active GSO´s metadata contact for maintenance.

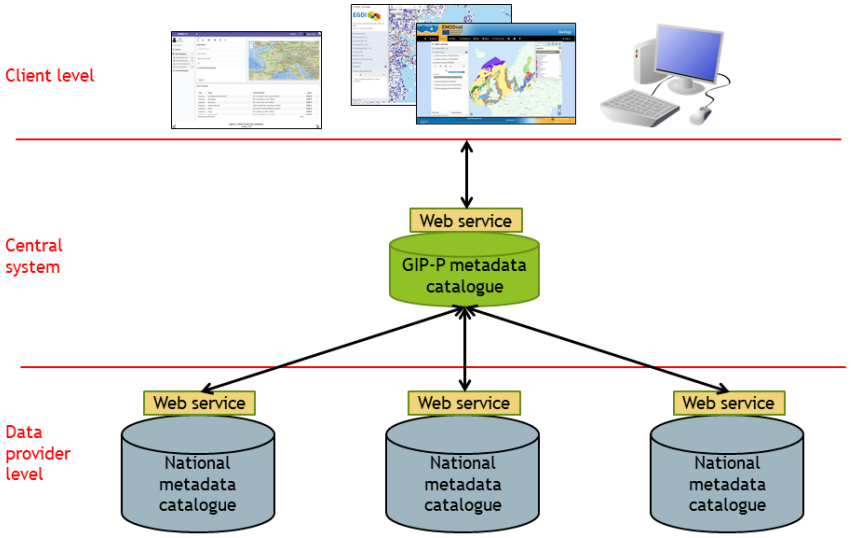

Figure 1 - Connection between Metadata Catalogue and Clients (e.g.: desktop and web map viewer)

The Metadata profile elements for which rules will be further elaborated:

- **Language** – proposed bilingual (national language + English). At least title and abstract should be filled in both languages.
- **Keywords –** specific GIP-P keywords should be used to distinguish data/services created for a specific project (project acronyms as Minerals4EU etc.).
- **Unified name of the organization** (Point of Contact) – within current EGDI project the solution is (*"English name of the organization (abbreviation)" example: Czech Geological Survey (CGS))***.** It is proposed to build on this. The use of URIs of organizations should be considered.
- **Unified name of the country** (Point of Contact) – currently not unified**.** The URI form should be used.

Country codes should be taken from the official European Union codelists as *http://publications.europa.eu/mdr/authority/country/*
(http://publications.europa.eu/mdr/resource/authority/country/skos/countries-skos.rdf)
Example:
http://publications.europa.eu/resource/authority/country/CZE

<gmx:Anchor
xlink:href="http://publications.europa.eu/resource/authority/country/ALB">Albania</gmx:Anchor>

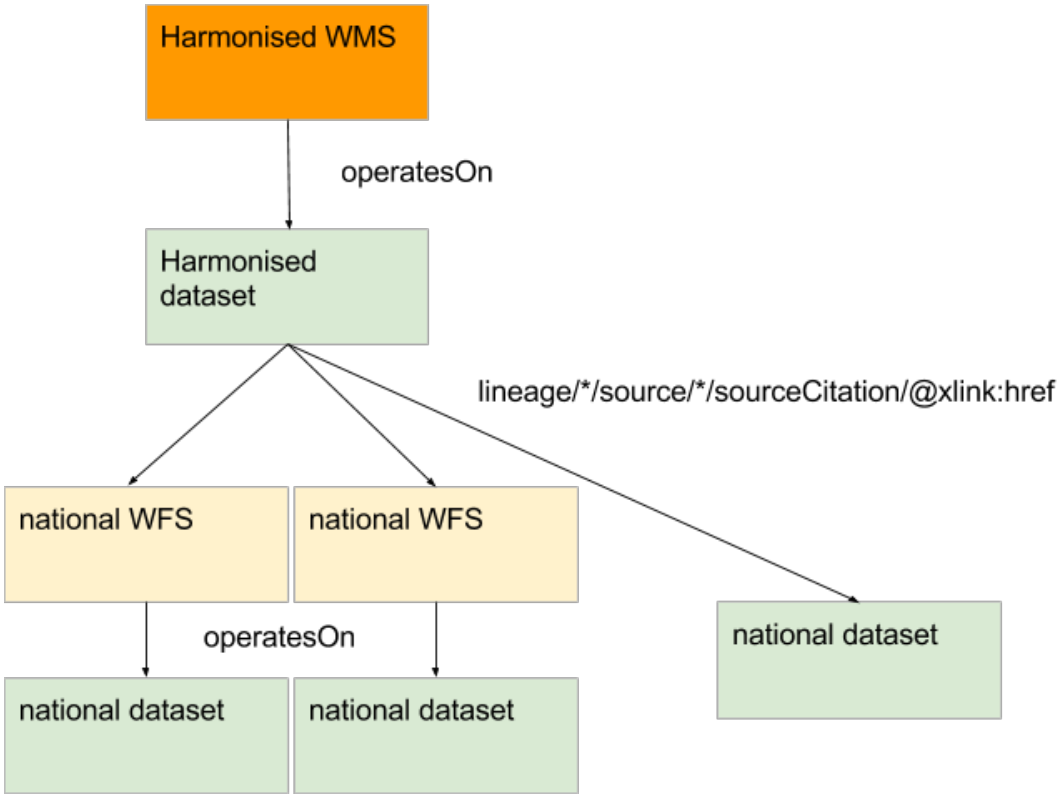

Figure 2 - Relations between metadata records of harmonized and national data sources

Description of mapping links between sources in metadata:

- Each national web service operates on one or more datasets. One-way relation from service to dataset is mapped with operatesOn metadata element, providing link to dataset metadata ISO 191139 record according to INSPIRE rules.
  If present in the metadata catalogue, the backward mapping from dataset to corresponding services shall also be available. Mapping from harmonized dataset to national resources (services as well as datasets are accounted for) is provided by dataQualityInfo/*/lineage/*/source/*/citation/*/xlink:href element, which contains URL to corresponding ISO 19139 metadata.
- The relations between harmonized services and harmonized datasets are mapped in the same way as relations between national services and datasets.

2 main open source solution exist and shall be compared against the functionalities described above before deployment takes place

- Geonetwork now under version 3.6.0 and supported by an important and ever growing international community

- Open Micka now under version 6 and implemented in EGDI (see EGDI metadata catalogue description in annex).

1.2 Standard OGC service stack for features

Geographical features will be exposed using a combination of WMS and application schema compliant WFS 2.

OGC WMS 1.3 specification will be implemented along with SLD/SE and tiling.
The recent MapServer branch 7.2.x will be the basis for this work.

OGC WFS 2.0.2 will be implemented along with Filter Encoding.
The recent GeoServer branch 2.16 will be the basis for this work along with app-schema and possibly SolR extensions. Indeed recent development GeoServer funded by the Geoscience community enabled to benefit both from the semantics conformity provided by app-schema extension and SolR indexes performances in combination with more classical Postgre/GIS database.

Both those services will be have their GetCapabilities responses conformant to INSPIRE requirements on services metadata.

1.3 Observation features OGC services strategy

In some simple situations, exposition observation (OM_Observation or specialization thereof) can simply be achieved deploying an application schema compliant WFS 2.

In other situations (e.g.: WaterML2-Part 1: time series…) it is wiser to deploy OGC Observation dedicated standards.

OGC Sensor Observation Service (SOS) 2.0 has been in the air for a more than a decade a performant implementation exists especially when it comes down to exchanging time series observation. However, the deployment of those implementation can be sometimes time consuming and is not the fastest to learn when compared to the recent OGC SensorThings API Part 1 : Sensing (which could be, wrongly, reduced to the REST binding of SOS 2.0).

As a result observation features will be exposed considering

- first OGC SensorThings API Part 1 implementation. The recommended implementation is the one from Fraunhofer IOSB,
- then OGC SOS 2.0 using 52° North implementation.

1.4 URI resolver

As proposed in D 5.1, EPOS TCS GIM and Minerals4EU for its codelist have already moved forward on a geoscience, yet project neutral, domain (https://data.geoscience.earth/) with corresponding URI patterns

- base/def : for specification (xsd, ontologies),
- base/id : for features ,
- base/ncl : for nomenclatures and codelists,
- etc.

There are numerous approaches to provide the URI resolving functionalities.
But a production resolver needs to be able to handle

- content negotiation using HTTP headers but also explicit media types requirement in the URL
- rewriting in proxy mode and permanent redirect
- the ability to cache replies to ensure performances
- and also to load balancing for heavily requested resources

As a result, the resolver implemented for https://data.geoscience.earth/ is already deployed and tested against those needs and proposed a combination of Apache with mod_rewrite, NGINX and Docker. It is proposed to build on this.

1.5 Codelist registry tool

There are so far only 2 open sources solutions identified to fulfill such need

- JRC Re3gistry
- and Epimorphics UKGovLD/LdRegistry

In a recent workshop of the INSPIRE register federation group (September 2018), JRC Re3gistry evolution wish list identified the need for the tool to behave according to Linked Data and Semantic Web as the first enhancement to be realized.
The tool is not considered to be mature enough with regards to Linked Data and Semantic Web principles to be the corner stone of the codeList registry tool functionalities.

It is thus suggested to build on Epimorphics UKGovLD. The tool was developed with the UKGovLD group (comprised of public sector and private sector membership. Input was provided from UK, EU (JRC and other groups), WMO and CSIRO.
There is now a progressive uptake in the Geoscience community (WMO, CSIRO, UK Met Office, BRGM, …) and also in public sector (UK DEFRA, UK Food Standards Agency, ...).
The code is maintained under Apache License 2.0 and the triple store used in the backend is Apache Jena.

It is already deployed in its V 2.0 as a response to https://data.geoscience.earth/ncl/ to expose EPOS GIM and Minerals4EU codelists. It is proposed to build on this

1.6 Shared data specifications environment

Data specification agreed on or patterns of data specifications to be used shall be shared in a single access point.

When data specifications are extended and/or defined the same practices shall be used.

It is proposed to build on the experience from many European and international geoscience data exchange initiatives and, when the data modelling is done in UML, implement ISO 191xx series of standards:

- do the data modelling in Enterprise Architect in combination with SVN to expose the models in XMI and version the edits
- expose the specifications in XMI but also HTML and generate the corresponding XSDs (and schematrons if needed)
- there is not, so far, a specific constraint for doing ontologies apart from providing both the vocabulary representation in machine language (.ttl, .owl, …) and in HTML (for example using Live OWL Documentation Environment - LODE)

All the specifications shall be exposed under https://data.geoscience.earth/def/ as already done for EPOS GIM for its complex borehole model (https://data.geoscience.earth/def/eposb).

1.7 Data publication alternative

This section corresponds to the need arising from D 5.1 section "6.2 Situation 2 - data provider does not have the IT capacity".

There is a specific effort to be carried on that aspect.
Some open source solutions already exist nationally like CARMEN (http://carmen.naturefrance.fr/) in France which is hosted by BRGM for the French Ministry of Environment but would require internationalization and updates to move to a Linked Data approach.

Some are proposed by companies and can be hosted in GIP-P like the web based solution from We Transform that build on HALE (https://www.wetransform.to/)

Some other may already exist but not be identified at the time of writing.

There will be a need for testing and evaluation before an actual IT decision/recommendation can be made.

1.8 Validation services

1.8.1 Metadata validation

- Built-in metadata validator is part of the metadata catalogue for metadata records to ensure compatibility with INSPIRE. It may be modified to the project metadata profile if a specific need reveals (e.g. mandatory project thesaurus keywords).
- Metadata validator may be used as standalone tool for on-line validation user metadata before posting to the catalogue.

An official INSPIRE metadata validator may be used.

1.8.2 Data validation

Validation services mainly targeting data exposed by data providers, those services will target first the XML/XSD/Schematron pattern.

Two solutions will have to be explored

- enriching the INSPIRE validator
- developing specific GIP-P validators

In case Semantic Web exchange validation are implemented, other approaches such as SHACL (Shapes Constraint Language) will be foreseen.

1.9 Harvesting system

Deploying a harvesting system is described in 2 of the 3 foreseen architectures.

Experience from pre-existing projects (e.g.: EPOS GIM, Minerals4EU, EGDI Surface Geology) show that it is often down to an XML parser development exercise.

However, it has to be noted that

- harvesting goes along with validation of the flows (thus a validation service) according to the data structure but also the proper use of the codelist registry tool entries for attributes that are based on codelists,
- harvesting is way simpler on SimpleFeature than on ComplexFeature based flows. This is both valid for the data diffusion endpoint (the more the feature structure is complex, the higher the stress is on the IT infrastructure) and for the harvesting system,
- harvesting strategies are also highly important for both the data diffusion endpoint and for the data consumer. In EPOS GIM a Pub/Sub approach based on Apache Kafka is implemented between the central node and some data providers. This to only harvest what changed and not and entire database through its WFS endpoint each time.

1.10 Central spatial database

Deploying a central spatial database is described in 2 of the 3 foreseen architectures.

Data will be one of the central points of the system. Although the new paradigm approach represented by the use of a Triple Store as described in the next chapters, a more classic spatial database is a need that shall fulfill the Information Platform.

From the previous project experience, PostgreSQL database with its spatial extension PostGIS is the best choice that come into being.

The latest version of the PostgreSQL database (v11.1 when these line are written) are already enough tested to be a good choice. Its accompanying PostGIS extension in version 2.4 might

present some limitations for very specific functionalities that will probably not be used in the Information Platform.

1.11 Data fronts ends / GUI

Writing recommendations on the GUI library to implement is always a moving target.

However, one shall remember that

- the GUI is not the Information System itself but only a visible part of the iceberg,
- the GUI can be web but also desktop based,
- the GUI can be a cartographic interface but also of completely different nature.

With the above in mind it is proposed to

- build on the experience of the previous (and running) projects that are described in Section 2 for the cartographic and CMS parts. Most importantly the EGDI web-GIS viewer is the starting point of the GUI work to display GSP data,
- consider new types of web-based application that enable to consume, process and use data in order to expose then in various representations like Jupyter Notebook (https://jupyter.org/) currently being implemented in EPOS,
- take into account development made in the open source GIS community to enable QGIS to consume OGC/INSPIRE based data flows and implemented a primer of Linked Data consumption (ex: https://plugins.qgis.org/plugins/gml_application_schema_toolbox/),
- take into account recent developments from the geological surveys when it comes down to 3D : the use of 3D web client like Cesium or WebGL to bring the 3D information to the browser of the user,
- consider recent developments from the geological surveys when it comes down linked data handling / visualization: as a support for ELFIE, BRGM prototype a tool called BLiV (for BRGM Linked Data Viewer). It's it demonstrated in 2 ELFIE demo videos
- (https://opengeospatial.github.io/ELFIE/demo/groundwater_monitoring.html and https://opengeospatial.github.io/ELFIE/demo/surface_groundwater_network_interaction). Its code is open on GitHub and it is planned to make it evolve over the course of the Second ELFIE (see the chapter below).
- consider the latest techniques to propose the user of the interface the better experience in term of web application (use of responsive technology like Angular or similar that lead to a fully integrated user experience).

1.12 Exposing features using JSON-LD

D 5.1 identified the requirement stemming from W3C Spatial Data on the Web working group to make data indexable by search engines as one of the most important to be achieved in order to enhance data discoverability and reuse

This chapter focuses on features as the exposition of metadata catalogue content has already been described in the chapter dedicated to metadata.

A recent OGC Interoperability Experiment called ELFIE (Environmental Linked Features Interoperability Experiment) tested the approach on top of the OGC philosophy and service stack. It's Engineering Report (OGC 18-097) is in the OGC portal pending documents at the time of writing that deliverable but it will be soon be made publicly available.

In the meantime ELFIE outcomes and demo are already accessible online: https://opengeospatial.github.io/ELFIE/

Given its success and importance with regards to the recent OGC/W3C collaboration, a Second ELFIE (SELFIE) project will be launched in 2019, thus in parallel to the GIP-P timeline

It is recommended to build GIP-P work on the findings from both ELFIE and SELFIE.

ELFIE already proposed JSON-LD contexts of interest for some features to be exposed by GIP-P. Various approaches have been discussed and tested in ELFIE to generate the corresponding JSON-LD payload: files (from DB or triple store content) or wrapper on top of application schema compliant WFS 2.
It is already planned in SELFIE roadmap to discuss how JSON-LD payload compliant to provided JSON-LD contexts can be natively generated from WFS 3 and SensorThings API implementations.

The next chapter dedicated to WFS 3 develops a planned strategy around WFS 3 and JSON-LD.

It also has to be taken into account that, parallel to SELFIE, tests between BRGM / Fraunhofer IOSB / DataCove are also approaching the same needs but between SensorThings API Part 1 and JSON-LD. Their conclusions will also feed in turn SELFIE.

1.13 Exposing features using WFS 3

As identified in D 5.1, WFS 3 is the upcoming OGC standard AIP to expose features to the web. At the time of writing this first version of D 5.2, the core part of the standard is in draft and opened to public comment with the goal to resolve all the pending comment beginning of 2019 and have a stable core specification available for mid-2019.

There are points heavily under discussion on WFS_FES GitHub (where WFS3 core specification is available) especially on the handling of identifiers and the compliance to data schemas (what was the normal expected behavior of an application schema compliant WF2).

Some implementations of the core draft exist: GeoRocket and GeoServer WFS 3 community module are amongst those.

Based on current test on going exposing in WFS 3 French water wells according to EPOS borehole specification it is advised to base GIP-P implementation on GeoServer WFS 3 community module. Indeed the development strategy around GeoServer and WFS 3 aims at allowing to use it in combination with the app-schema extension.

Which will allow to

- deploy an application schema compliant WFS2 (which is expected anyway from GIP-P)
- and when activating the WFS 3 module, de facto have an OGC WFS 3 compliant endpoint still respecting the semantics configured via the app-schema extension.

The current test will also evaluate to which extent GeoServer WFS 3 output could be compliant to predefined JSON-LD contexts. This activity will be conducted in parallel to the Second OGC ELFIE (SELFIE) and with shared people within both exercise to ensure consistency of the results. It will also provide feedback to the WFS_FES GitHub.

1.14 SPARQL endpoint and triple stores

Various open source triple stores and each year public benchmark results are made available to the public.

Exposing GIP-P content from triple stores will trigger several constraints amongst which:

- volumetry as European datasets will easily reach millions of arcs (ex : test made on EPOS EU BoreholeIndex),
- and involve big (and sometimes complex) geometries.

Building on the experience from previous projects (e.g.: H2020 MICA) it is advised to start the triple store implementation on one of the following implementations: Apache Jena-Fuseki / Virtuoso.

2 REACHING THE TARGET - PROJECTS THAT SERVE AS A BASIS FOR GEOERA

2.1 Former projects for Geo-energy

Remark: *for this preliminary version of the deliverable, no former projects for Geo-energy have been described.*

2.2 Former projects for Groundwater

Remark: *for this preliminary version of the deliverable, no former projects for Groundwater have been described.*

2.3 Former projects for Raw Materials

The last years have been punctuated with different projects in the domain of Raw Materials with always the same goal, centralize the information or at least the access to it, and homogenize how the data and the knowledge is structured (and to a lesser extent, harmonize this data at European scale).
Through European projects from the 6th Framework Program to the Horizon 2020 work program, projects have explored different facets of architecture to fulfil these goals.
With the first projects ProMine and EuroGeoSource (both in FP6), the basis for a central database with homogeneous data for the whole Europe have been settled up, and the mechanisms for an harvesting of European data provider (mostly Geological Surveys) to fill a database with a common model using the same toolstack.
The following projects (both FP7 and H2020) have been the opportunities to pair up these approach to have an integrated way of proceeding. The Minerals4EU (with its declination in EURare for REE, SCRREEN for CRM and partially ProSUM for SRM) has been one of the first knowledge data platform for Raw Materials that have coupled the principle of a central and homogeneous database for Europe and the principle of harvesting national data providers to fill up this database with always up to date data. The architecture that has been created from these principles can be seen in Figure 6. Apart from this central database and its harvesting system, the Minerals4EU platform proposes also an access to statistics for each European countries on Production, Trades, Reserves, and Resources: the e-Minerals YearBook. This tool, like the Urban Mine Platform (UMP) of ProSUM portal, doesn't rely on the central database nor harvesting system but is dependent of another database (or schema of the database to be precise) which is fed with a different mechanism (for the e-Minerals YearBook, some Excel files are created and manually integrated to a dedicated schema of the diffusion database; for the Urban Mine Platform, some Excel files – based on templates – are created by the different data provider – per waste group – and are manually added to an integration database, then this database is copied in the diffusion database and specific views – materialized views – are used to feed the services requested by the web platform).

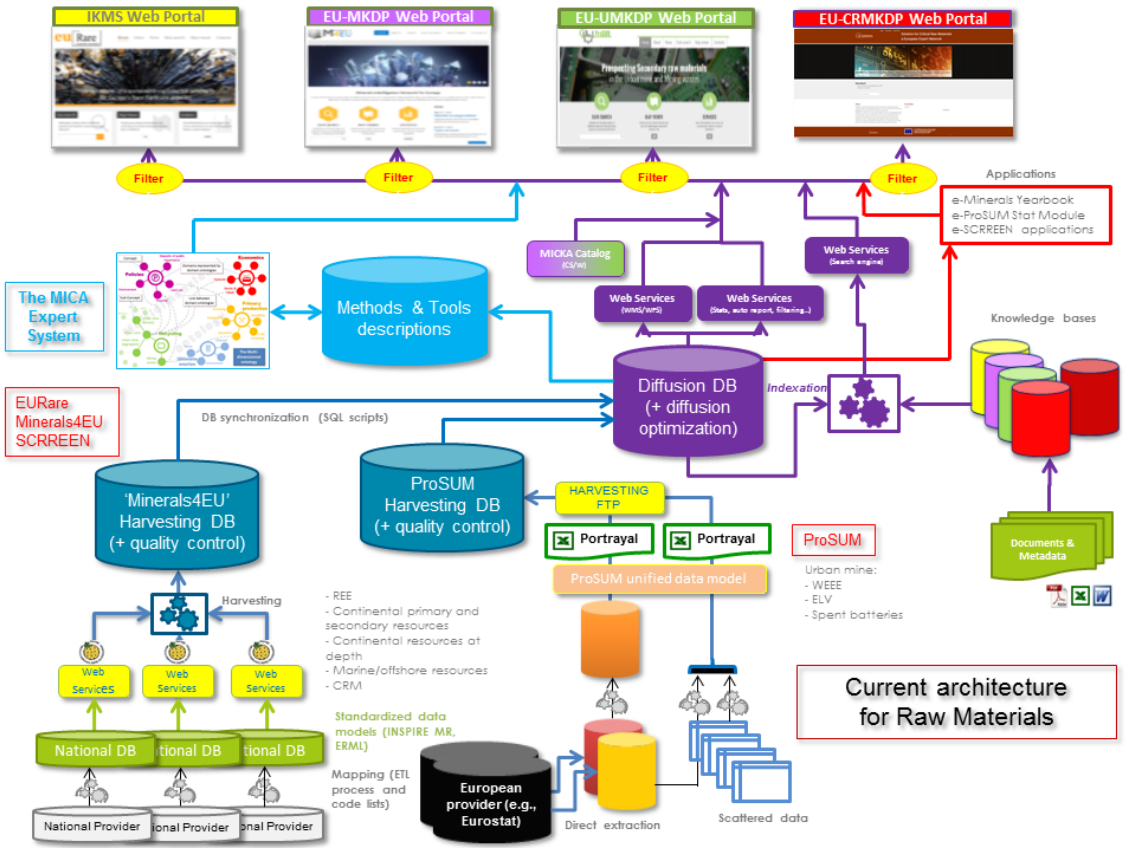

Figure 3 - Current architecture of raw materials projects

These knowledge base (EU-MKDP – European Minerals Knowledge Data Platform – for Minerals4EU, IKMS – Integrated Knowledge Management System – for EURare, EU-CRMKDP – European Critical Raw Materials Knowledge Data Platform – for SCRREEN and EU-UMKDP/UMP – European Urban Mine Knowledge Data Platform/Urban Mine Platform – for ProSUM) integrate also a component for the management of non-structured knowledge (documents, reports, books, videos, graphs, etc…). In the knowledge bases, the metadata for these documents are stored in a repository (related documents can be also stored inside the repository but can also – and preferably – be outside the repository in a place accessible through an URL. The metadata and the related document are indexed by a search engine that can be requested through a simplified interface in the different portals.

2.3.1 FP6 ProMine

One of the main objectives of the ProMine project was to develop the first pan-European GIS-based database containing the known and predicted metalliferous and non-metalliferous resources, which together define the strategic reserves (including secondary resources) of the EU.

2.3.2 FP6 EuroGeoSource

The aim of the EuroGeoSource project was to provide information on oil and gas fields, including prospects and mineral deposits, in order to stimulate investment in new prospects for geo-energy resources, as well as in renewing production at mines undergoing economic decline or closure, contributing this way to the independence of the EU having to import valuable minerals from outside resources.

2.3.3 FP7 Minerals4EU

The Minerals4EU project was designed to meet the recommendations of the Raw Materials Initiative and to develop an EU Mineral intelligence network structure delivering a web portal, a European Minerals Yearbook and foresight studies.

The Minerals4EU project has built around an INSPIRE compatible infrastructure that enables EU geological surveys and other partners to share mineral information and knowledge, and stakeholders to find, view and acquire standardized and harmonized georesource and related data.

For that purpose, the Minerals4EU project has developed a web portal (http://minerals4eu.brgm-rec.fr) with some components to give access to the users to the data and knowledge created during the project.

2.3.3.1 Content Management System

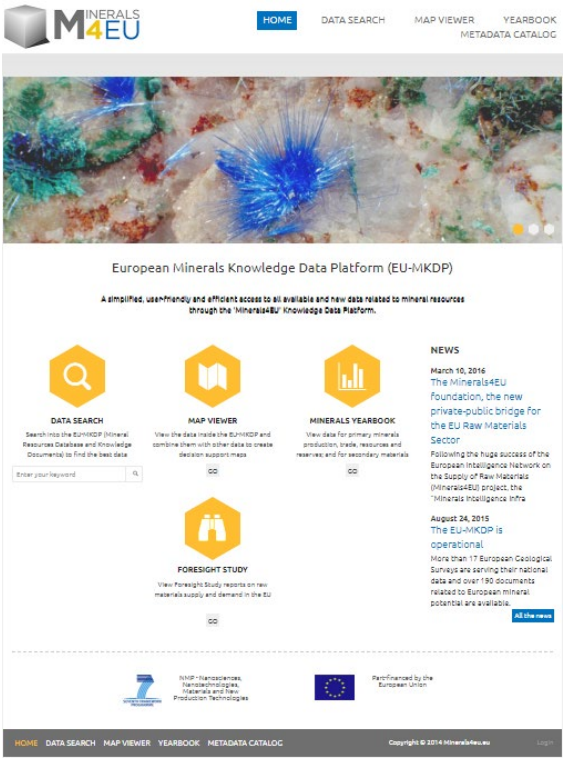

Figure 4 - Home page of Minerals4EU Portal - the CMS

The base component for the Minerals4EU portal is the CMS (Content Management System) that displays the latest news from the project and the portal itself. This home page also leads to the different tools/parts of the system. This component is based on a Drupal 7 with few enhancements especially for the look-and-feel.

2.3.3.2 Map Viewer

The Map Viewer is the way to view the data in a geographical way. This component proposes a map with different base layers where the user can add the Minerals4EU layers for the data (Mines and Mineral Occurrences, available with different representations). Some other layers can be added, from previous or ongoing project.

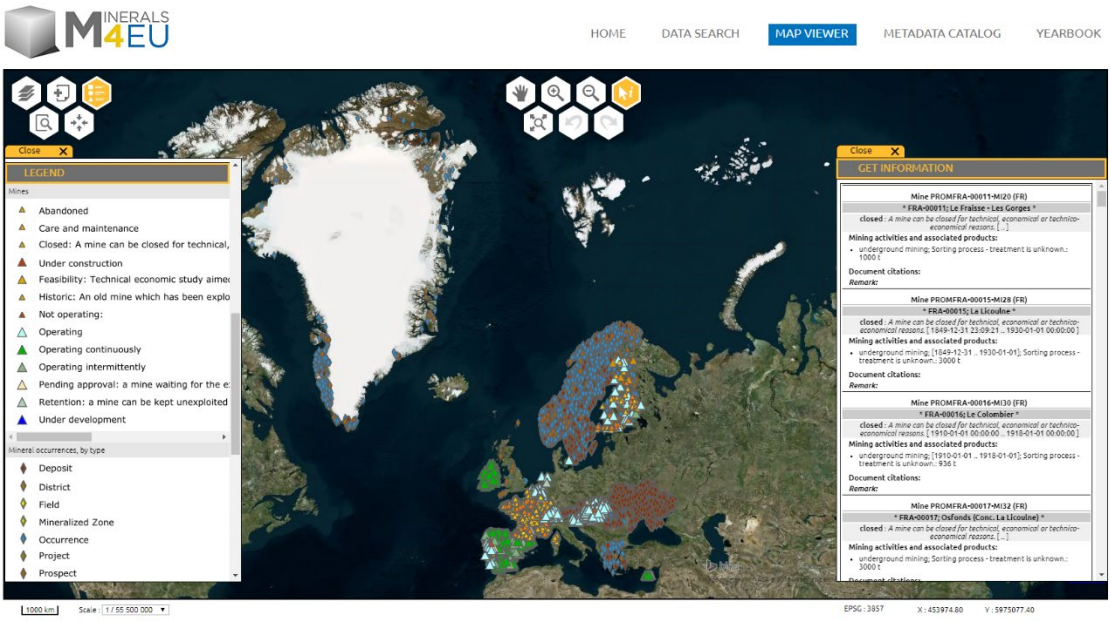

Figure 5 - Minerals4EU's Map Viewer

The Map Viewer is based on the OpenLayers 2 library, and rely on OGC services for the displayable maps. To avoid problem of cross domain, the viewer embed a java proxy to retrieve external content in a seamless way.

2.3.3.3 Search engine and indexation of data and documents

The search engine in Minerals4EU can handle both data from the database (stored in ERML 2 with M4EU extension in a PostgreSQL database), and documents that are stored in the CMS of the portal. The CMS can handle metadata records (based on Dublin Core Metadata Element Set) that are implemented as a user defined type of content of the CMS. In the metadata record, the user shall define a link to the resource described, it can be a local link (in that case the document itself is stored in the portal) or it can be (and this solution is encouraged) on the original location of the document (website of the original producer, website of the publisher…). The search engine, based on SolR, is made to handle the documents in the CMS by reading and indexing the

metadata record, and if the document is accessible (locally or remotely) by indexing its full text. The data from the database is also indexed and the content of the index can be searched using a very simple search form or a specific form (with predefine search fields corresponding to the structure of the database or the available metadata records).

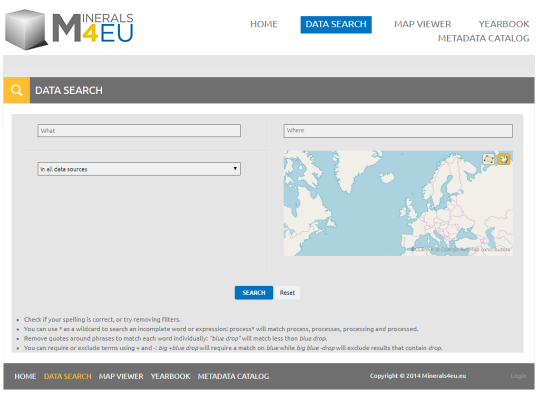

Figure 6 - Minerals4EU's search interface

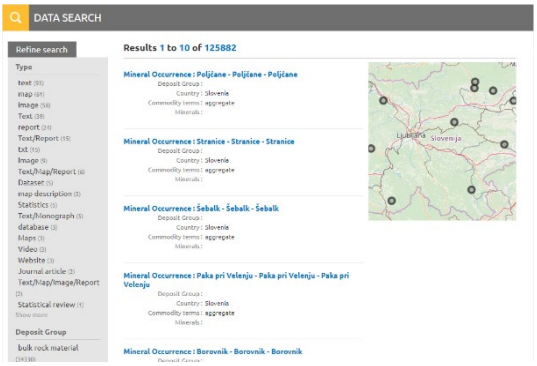

Figure 7 - Minerals4EU's search results

2.3.3.4 Metadata catalog

The Minerals4EU Metadata Catalogue is available on URL address http://m4eu.geology.cz/metadata/. The catalogue was based on Micka software (version 5) as the central access point to metadata concerning European mineral resources and related topics.

Only digital and structured information were described by metadata in this catalogue (data, services, applications). Minerals4EU metadata profile is compliant with the requirements of the INSPIRE Directive for metadata; the EN ISO 19115 terminology has been implemented.

Figure 8 - The interface of the Minerals4EU Metadata Catalogue

The specific **Minerals4EU Metadata Keywords Codelist** was compiled and integrated into the Minerals4EU Metadata Catalogue.

The codelist can be accessed on http://m4eu.geology.cz/codelist/# and used for filling in keywords.

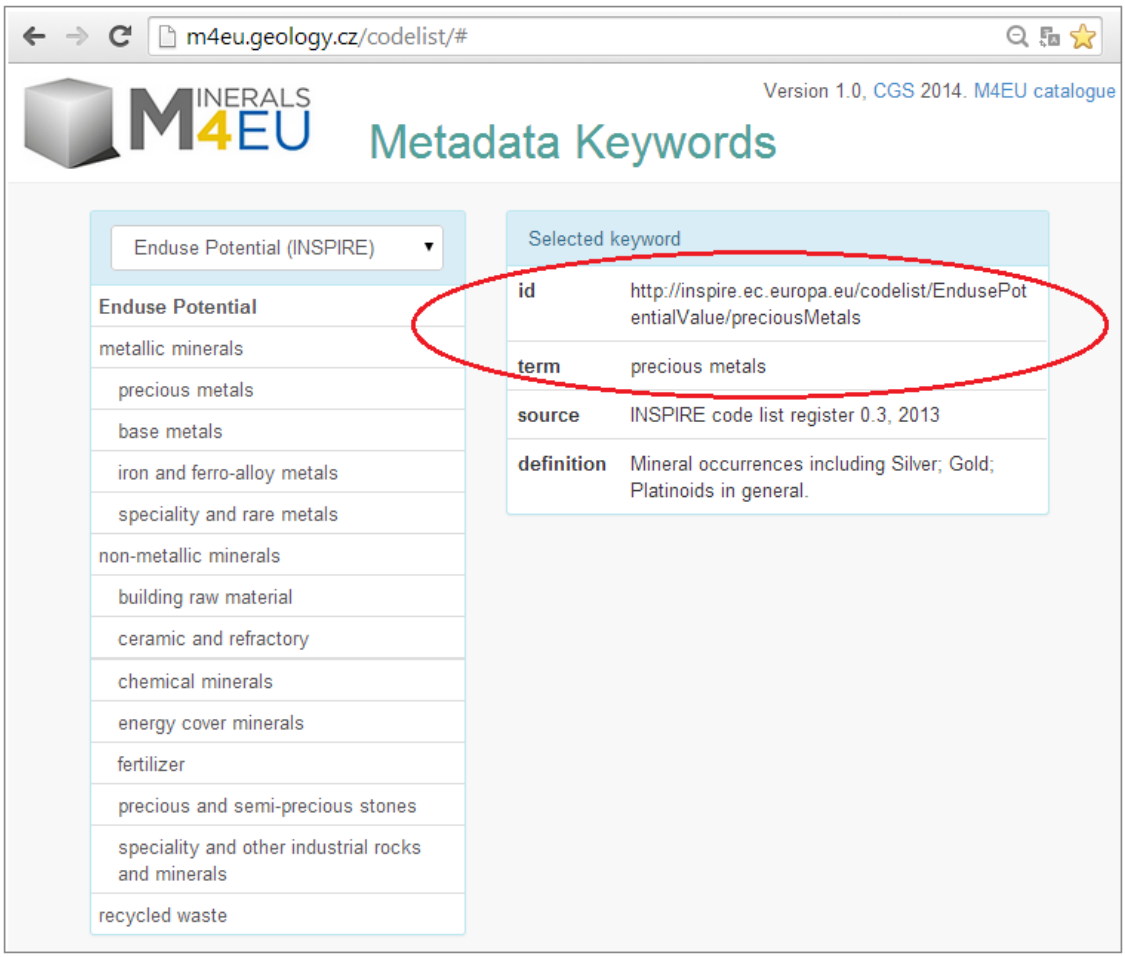

Figure 9 - Interface of the Metadata Keywords application created specifically for of the Minerals4EU project

2.3.3.5 Minerals Yearbook

The European Minerals Yearbook is a comprehensive source of minerals statistics for Europe. It presents official national level statistical data for 40 European countries, including all European Union (EU) Member States, countries in accession and potential candidates, as well as European Free Trade Association (EFTA) members. All the data is accessible digitally through the EU-MKDP.

It contains data for primary minerals production, trade, resources, reserves and exploration sourced from national authorities in each country. It includes also available data for mineral-based waste generation, treatment and trade. It also contains case studies relating to the resource potential of 8 commodities from secondary raw materials.

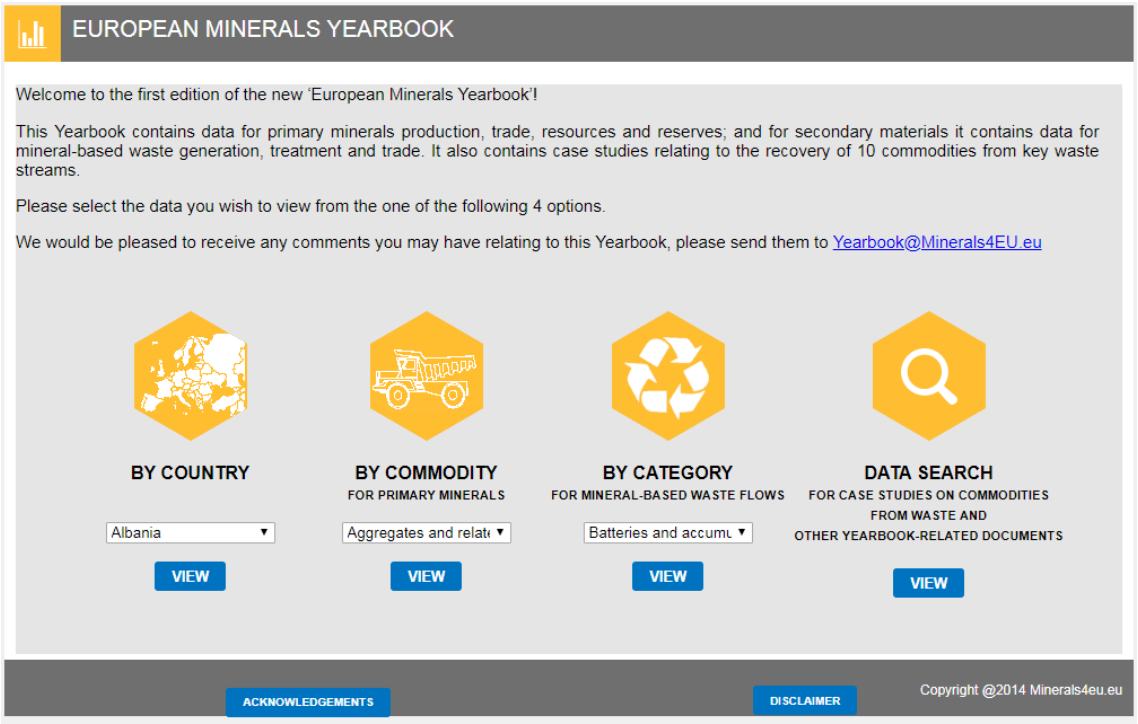

Figure 10 - European Minerals Yearbook - Home page

The European Minerals Yearbook includes data for 65 commodities for primary minerals and 13 mineral-based waste categories. The data are searchable by country or by commodity/category through a user-friendly interface. The retrieved statistics are presented in a range of tables, interactive maps and graphs enabling the user to compare certain data between countries. For production and trade, data are presented for the ten-year period 2004 to 2013.

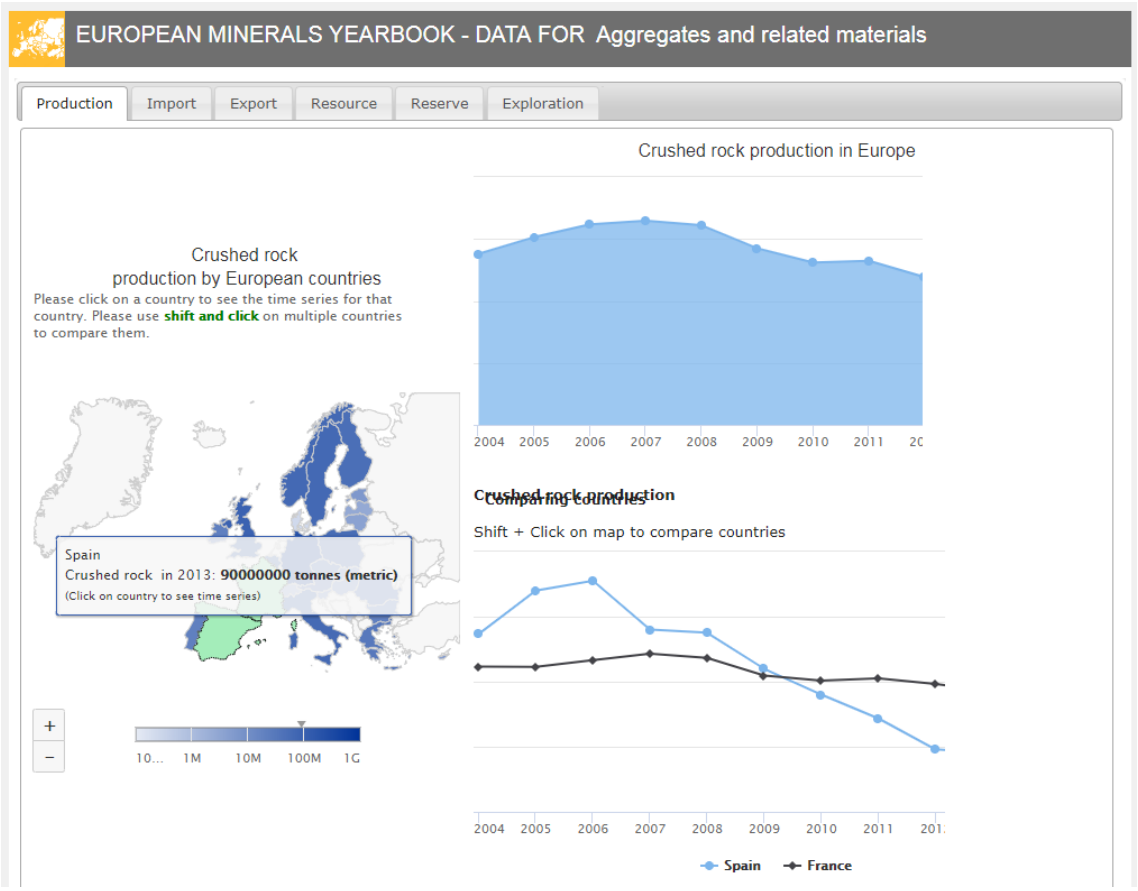

Figure 11 - European Minerals Yearbook - Viewing the data

The European Minerals Yearbook relies on a dedicated schema in the EU-MKDP diffusion database. This part of the database doesn't follow a standardized data model but is based on a specific data model created for the purpose of the Yearbook and close to the data provided as input. The input provided is not correlated with the data provided by each member state partner of Minerals4EU project, but created from external sources managed by BGS.

From a technical point of view, the data is provided by BGS as Excel file and then integrated in the database using ETL scripts.

The user interface of the European Minerals Yearbook is made using HTML/Javascript that rely on custom web services (REST API).

2.3.3.6 Harvesting system

The harvesting system consists of a harvesting application and one or several databases.

The Minerals4EU harvesting application retrieves WFS formatted data from each provider, harvests data from it, transforms it into proper form and then stores it in the Central Harvesting database. It was decided to develop the harvesting system using JAXB and with NetBeans.

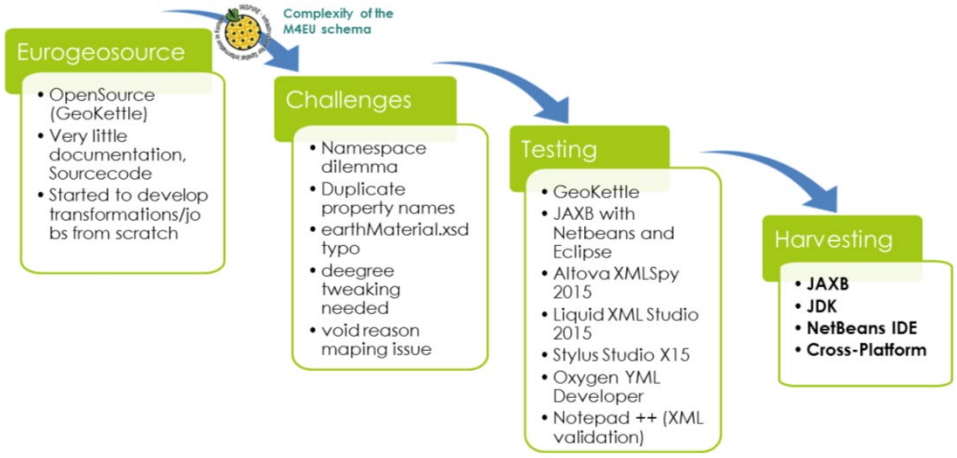

Figure 12 - The development of the Minerals4EU harvesting system and the challenges encountered

During the development of the Minerals4EU harvesting scripts it turned out that the GeoKettle ETL tool could not properly parse WFS 2.0 formatted XML files. The biggest obstacle was parsing XPath from XML files. For this reason it could not properly navigate through the XML elements and attributes of the XML document.

JAXB (Java Architecture for XML Binding API) was tested both with NetBeans and Eclipse SDK using their own JAXB components and several different JAXB libraries. The JAXB offers easy access to XML documents from within applications written with Java programming language. In order to access and harvest data from XML document with JAXB, the first task is to generate several Java classes that represent the project schema. Then it is necessary to unmarshal the XML document into Java content objects. The Java content objects represent the content and organization of the XML document. It is then possible to store this java objects in the database. One of the advantages is that the objects contain all the relations to other objects. Aggregation and other manipulation of data in the harvesting system, if needed, will be performed by procedures within the harvesting application and PostgreSQL/PostGIS database. The harvesting database will have the same structure as the database (or portrayal) that each provider will setup. The delivery of data from the harvesting system to the diffusion system will be carried out by sending SQL update scripts.

2.3.3.7 Database and spatial services

During Minerals4EU project (and EURare project), a ToolStack has been developed to fulfil the needs of the projects to help involved partners.

This ToolStack uses Enterprise Architect for the modelling of the creation/adaptation of the Data Model (UML) and for the modelling of the database, this is the only non-free software used in the ToolStack (nevertheless, a free viewer is available).

The other ToolStack elements are open source software but specific versions are used for compliance purposes:

- Apache Tomcat version 8.0.15

- Deegree version 3.3.13
- PostgreSQL version 9.2.9-1
- PostGIS 2.0.1-1
- GeoKettle version 2.6-r192

All open source software can be found in the Minerals4EU repository at:
http://data.geus.dk/svn/m4eu/

The modelling phases are the following:

1. Model /update the DB UML diagram;
2. Model/update the DB model diagrams;
3. Create the DDL data for the DB model;
4. Update the WFS schemas for the Deegree;
5. Create/update code list scripts;
6. Send out the update scripts to partners.

Once the modelling phases have been completed, the databases have been created at the different levels. Each partners of the project have their own database using the Minerals4EU model to be able to run the services. The database model is also used for the harvesting system and the diffusion system. The diffusion database also has some particularities for diffusion (for example: dedicated diffusion materialized views, used to speed up the display).

Each partners of the Minerals4EU project were able to implement the toolstack to have a functional feature services delivering their national data to the central system.

Their first step was to install the database using the cookbook provided by the project. Then, the partners have to fill the database (with the Minerals4EU schema) with the data coming from their own database, using ETL process (in the case of Minerals4EU, GeoKettle was proposed to partners with some examples of use). Finally, a version of the Deegree server with a preconfigured WFS compliant with INSPIRE RM/EarthResourceML was proposed to be installed in an Apache Tomcat to have the services available and working to be harvested.

2.3.4 FP7 EURARE

The main goal of the EURARE project was to set the basis for the development of a European Rare Earth Element (REE) industry. Establishment of an REE value chain in Europe would safeguard the uninterrupted supply of REE raw materials and products crucial for sectors of the EU economy (including automotive, electronics, machinery and chemicals) in a sustainable, economically viable and environmentally friendly way.

One of the goals of the EURARE project was the development of an Integrated Knowledge Management System (IKMS) for EU REE resources, which will provide information on REE and build up the knowledge to be developed within the frame of the project.

This project share the same basis as the Minerals4EU project (both projects ran during the same period of time and the development has been shared between them). The IKMS has the same

type of CMS, the same Map Viewer and the same search engine and share the same database than the EU-MKDP. The only difference is that the data displayed in EURARE portal has been filtered on REE, the Map Viewer has a filtering specificity and the corpus of documents is different than the one of Minerals4EU portal.

2.3.5 H2020 ProSUM

The goal of the ProSUM project was to deliver the First Urban Mine Knowledge Data Platform, a centralised database of all available data and information on arisings, stocks, flows and treatment of waste electrical and electronic equipment (WEEE), end-of-life vehicles (ELVs), batteries and mining wastes.

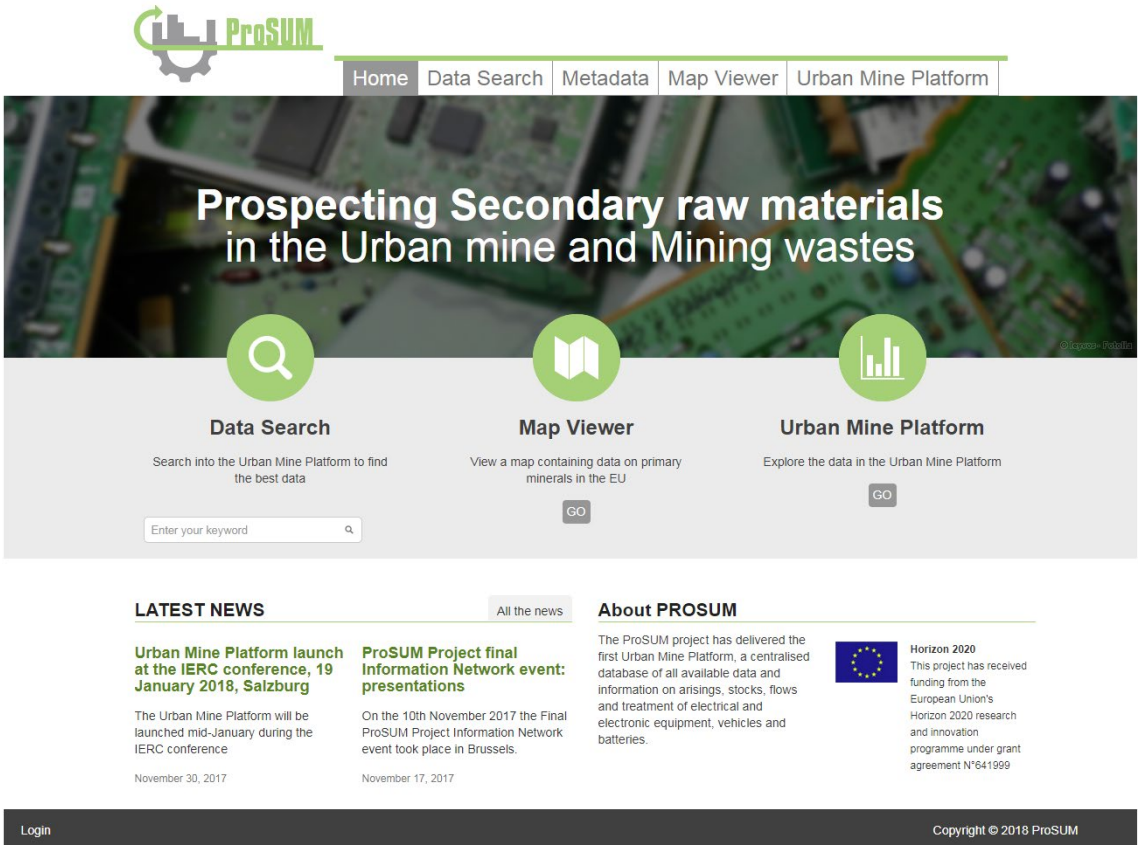

Figure 13 - The ProSUM portal homepage

The ProSUM portal shares the same base as Minerals4EU EU-MKDP. Only notable differences will be described here.

The system use a Content Management System to display the latest news, a Map Viewer to display the spatial data (datasets for primary and secondary raw materials) and a search capabilities to be able to retrieve document or data from the Knowledge base. These components are the same than in the EU-MKDP.

2.3.5.1 Metadata catalogue

The ProSUM Metadata Catalogue is based on Micka software (version 5). It is available on http://www.prosumplatform.eu (tab Metadata) or directly on http://prosum.geology.cz. It is the central access point to metadata concerning digital and structured data (non-geographic and spatial datasets or dataset series and spatial data services) on secondary sources of raw materials.

It is integrated to the Portal, and provides tools for compilation of those metadata in a standardized way. For each chart generated by the application, a button “more information” is available. This provides full and detailed explanation of the data including: the approach taken to produce the data, link to the source data, a description of the data quality and the author in ProSUM (Figure 17 and Figure 18). The catalogue is fully compliant with international standards.

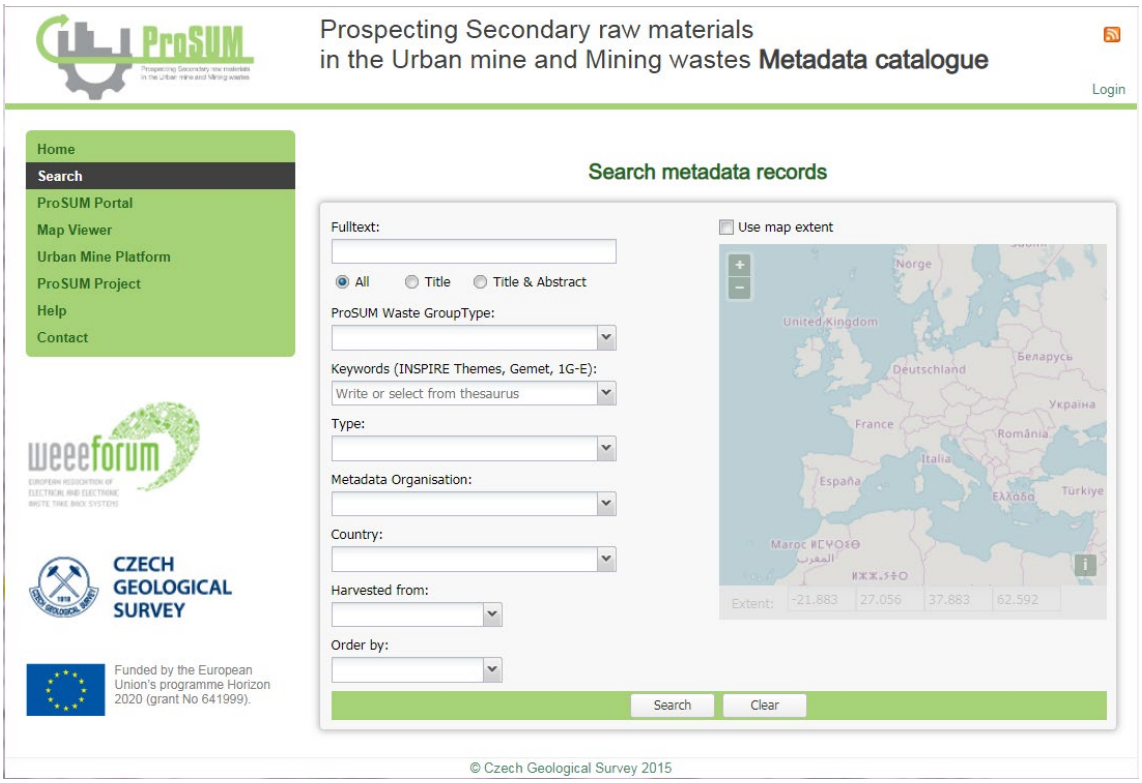

Figure 14 - The search page of the ProSUM Metadata Catalogue

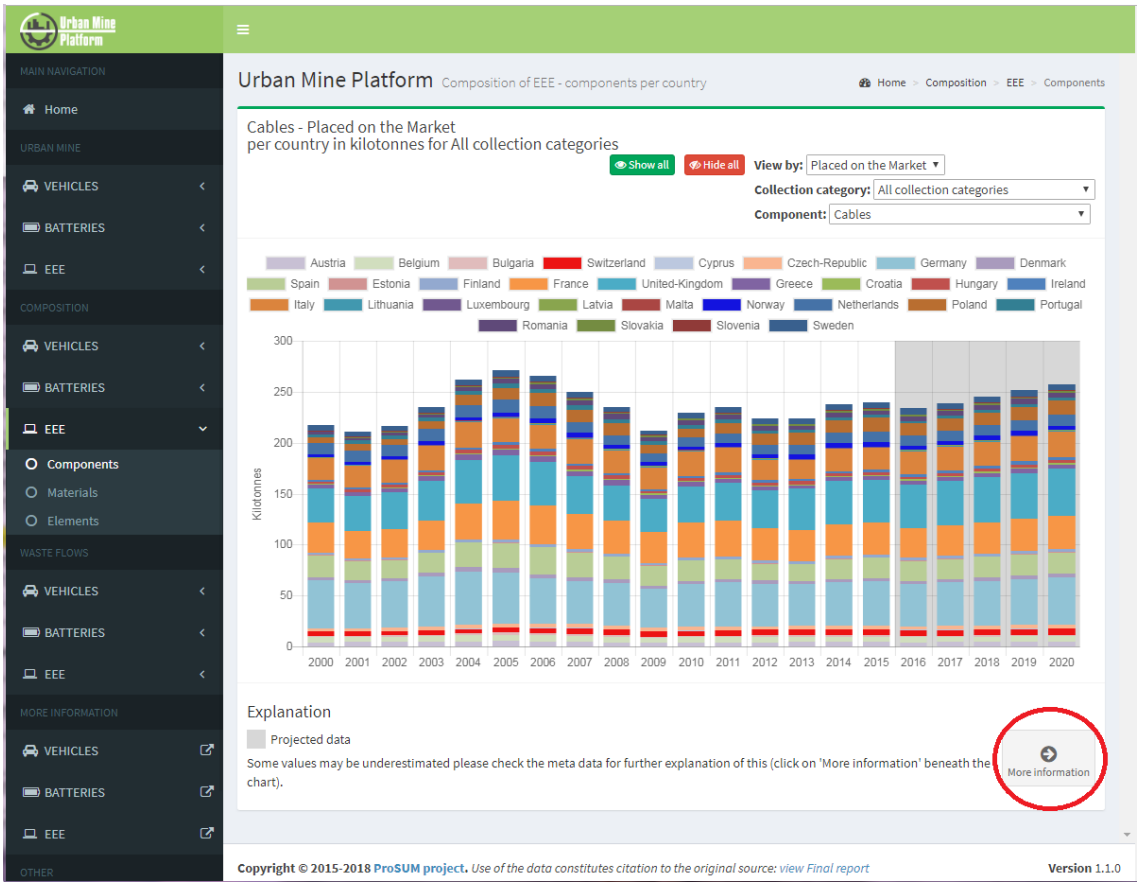

Figure 15 - Example of the link between the ProSUM Metadata Catalogue and Portal – link to the metadata record providing more detailed information on the displayed data

Figure 16 - Example of the link between the ProSUM Metadata Catalogue and Portal – link to the Portal

2.3.5.2 Urban Mine Platform

One of the major outcomes of the ProSUM project is the creation of the Urban Mine Platform (http://www.urbanmineplatform.eu) which displays all readily available data on products put on the market, stocks, composition and waste flows for electrical and electronic equipment (EEE), vehicles and batteries for all EU 28 Member States plus Switzerland and Norway (and Iceland for vehicles).

Through this platform, the user can select and produce charts. It is also possible to access the extensive library of more than 800 source documents and databases used to populate this platform (through the document repository already developed for Minerals4EU and adapted to ProSUM). The user can also see the metadata allowing him to review the key underlying information and to understand the data quality.

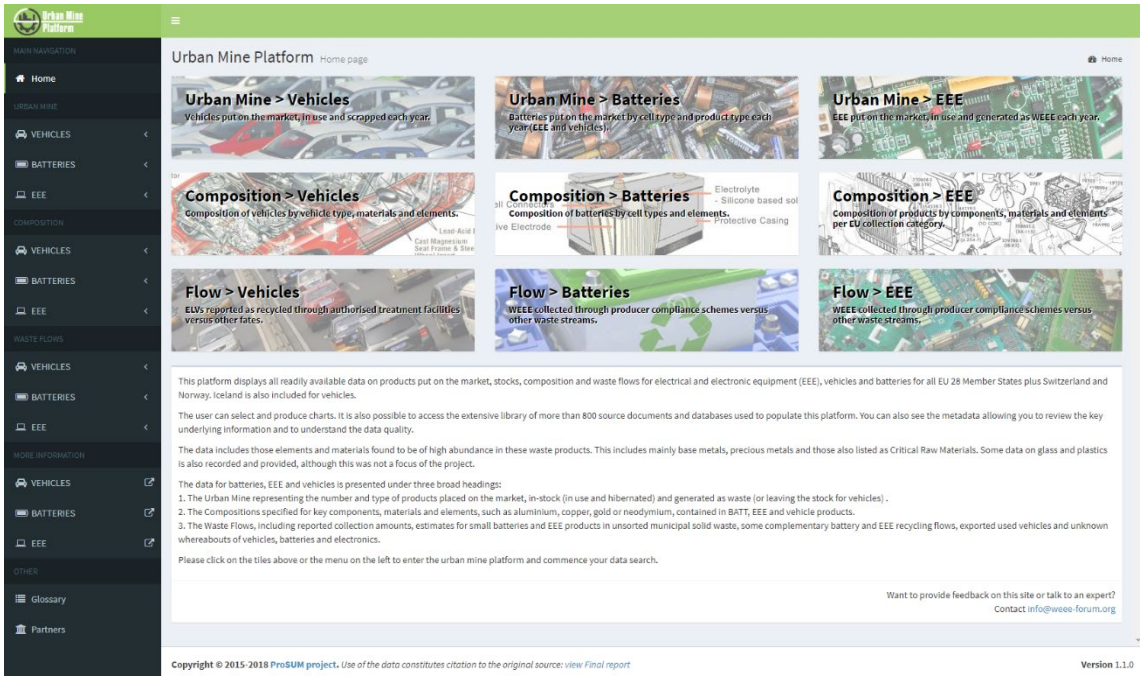

Figure 17 - Homepage of the Urban Mine Platform

This platform proposes some views composed of interactive graphs that relies on a dedicated API based on a custom database.

The database use the ProSUM data model that was developed during the course of the project. This data model allows to represent the different products with their composition (in components, materials or elements) and the flows or stocks of these products in and between the countries of Europe.

On this database is connected a REST API that give access to the aggregated data on the stock and flows and the composition of the products for the European countries.

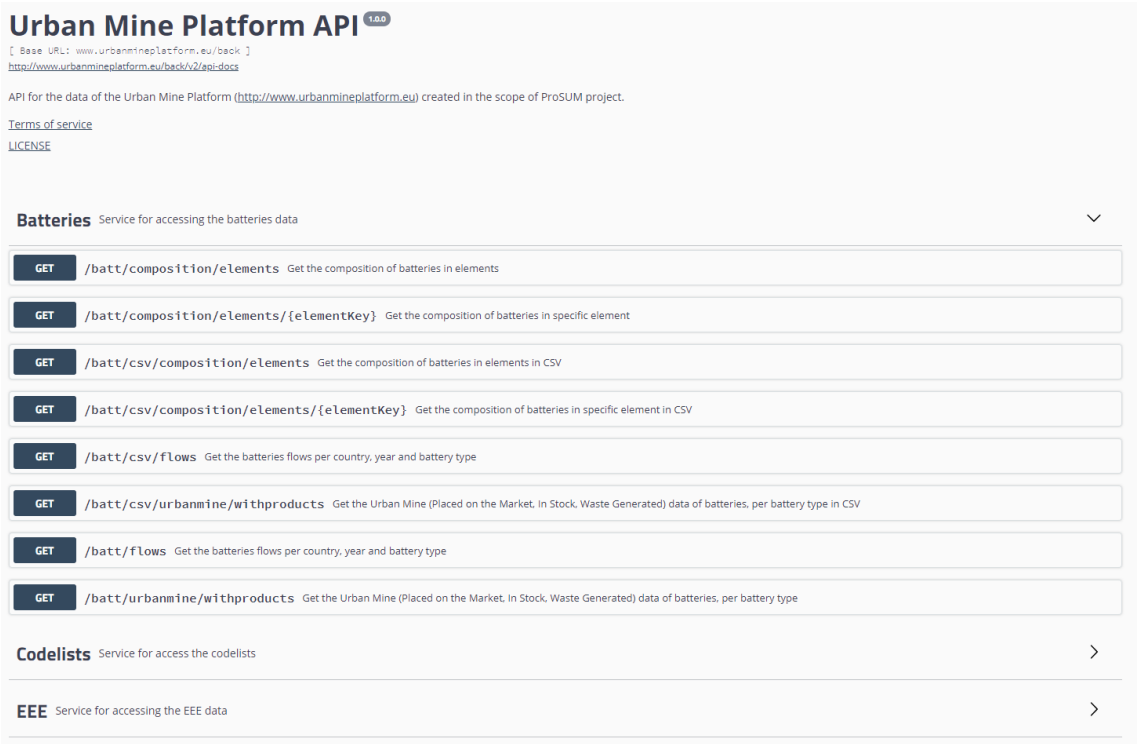

Figure 18 - Urban Mine Platform API

The platform itself uses this API service through an Angular application. This web application use the latest features of responsive technics to provide the user of the platform with the best available experience with an unique data loading for each data type (i.e. couple Waste group/Level of information). This solution allows the user to navigate into the data of a specific theme without any delay and other loading time, it allows also the user to filter directly the graphs in an interactive way.

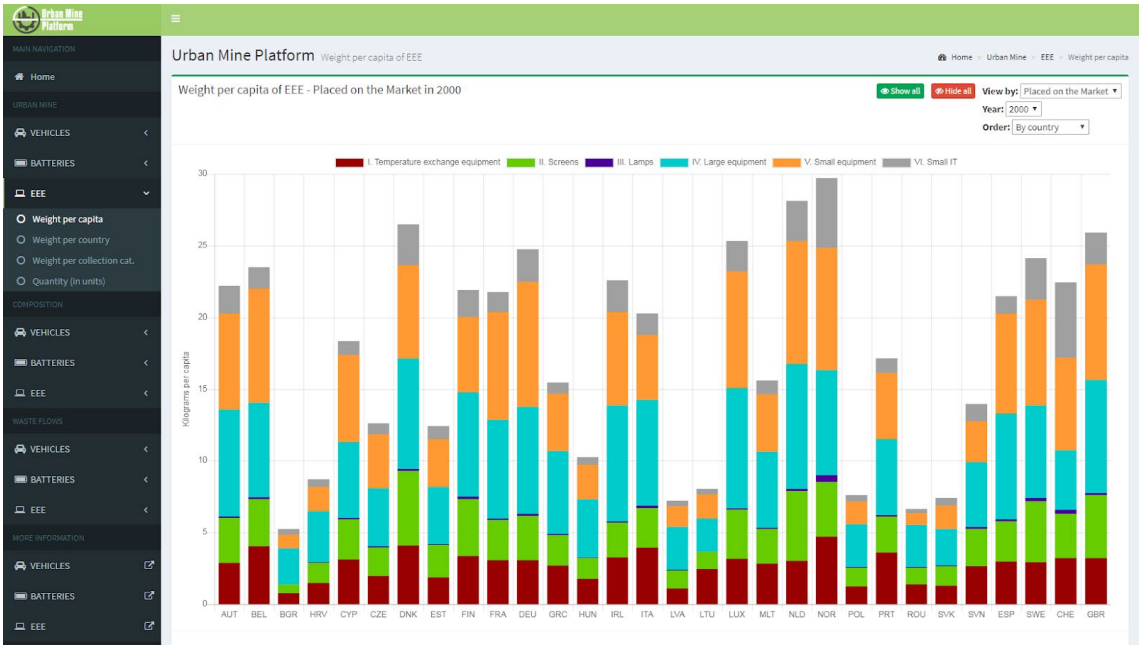

Figure 19 - Urban Mine Platform - Data visualization

2.3.6 H2020 SCRREEN

The goal of the SCRREEN project is to establish a EU Expert Network that covers the whole value chain for present and future critical raw materials; to analyze pathways and barriers for innovation, and identify the solutions for overcoming these barriers; to study the regulatory, policy and economic framework for the development of these technologies and to identify the knowledge gained over the last years and ease the access to these data widely and efficiently, beyond the project.

For these purpose, SCRREEN will collect and organize all of the data generated in other projects, associations, initiatives, and develop a knowledge data portal.

The EU-CRMKDP (European Critical Raw Materials Knowledge Data Platform) shares for its main functionalities the same architecture than Minerals4EU. Like the IKMS for EURARE, the portal has a CMS, a Map Viewer and a search engine base on the same technologies than for Minerals4EU. The corpus of documents is specific to the SCRREEN project and the data are filtered on the CRM. On top of that, the portal will have some applications that exploit the content of the database (the deployment of these applications is plan in late December 2018/beginning of 2019).

2.4 Other relevant projects

2.4.1 OneGeology Europe / OneGeology Global

The aim of the OneGeology projects is to bring together a web-accessible, interoperable geological spatial dataset for the whole Europe (OneGeology Europe) and World (OneGeology Global) at 1:1 million scale.
Unlike Minerals4EU, in OneGeology projects, the approach is not centralized with a harvesting system that retrieves the data from all the participating countries to feed a central database. OneGeology projects take the opposite approach, instead of harvesting everything to a single central database, all the data are let to their national providers, and the providers shall furnish harmonized services which are called by the portal, using WMS for the visualization on the map, or using WFS for the thematic queries.

This approach has pros and cons. As the data stays only at the provider level, updates of the data are directly visible by portal users, there is no delay lead by the harvesting procedure, but in the meantime, when a data provider has a failure, the data is no more accessible. Also, all the processing times stay at the provider level, but as the data are not centralized, some treatment cannot be doable (or easily doable).

Figure 20 - Former One Geology Europe portal

To fulfil the objective to have the data providers bringing interoperable and harmonized WMS and WFS to the OneGeology Europe portal was a challenge. When the project has started, the software were not able to handle functionalities like custom application schemas, cardinality of some attributes, multilingualism… A tool, eXows (for eXtended OGC Web Services), has been developed to bypass these limitations. This software was acting like a proxy in front of the standard OGC Web Services of the national provider (most likely a MapServer, GeoServer or Deegree server) to handle the multilingual parameters and then translate the requested layer name to another layer/server, and to handle a simple feature schema returned by the WFS to transform it to a complex schema according to the OneGeology XML Schema. This software was abandoned when the standard WMS/WFS software were able to handle the complex features and language parameters.

OneGeology Europe was dismantled to the benefit of EGDI and the data where transferred to EGDI database (abandoning by the same the approach of letting the data provider furnishing directly the data to the portal).

2.4.2 EGDI

EGDI stands for EuroGeoSurveys' European Geological Data Infrastructure. This infrastructure provides access to Pan-European and national geological datasets and services from the Geological Survey Organizations of Europe.

At its core, EGDI consists of the following components:

- A portal for dissemination of pan-European geological projects and data products.
- A metadata database for searching through records in INSPIRE compliant format.
- A product database for hosted products – most prominently: OneGeology Europe.

2.4.2.1 The EGDI portal

The portal serves as the main entry for all available data products. Mostly accomplished by serving web maps like this:

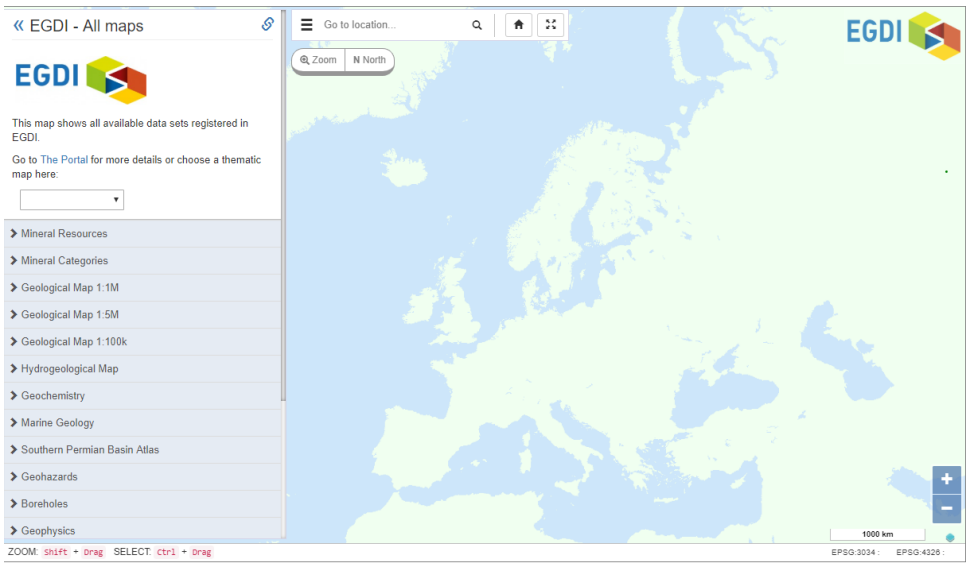

Figure 21 - View of the EGDI portal

EGDI has a dual approach, data products can either be hosted by EGDI through harvesting techniques or served from distributed sources (WMS and WFS).

Here is the harvesting approach illustrated:

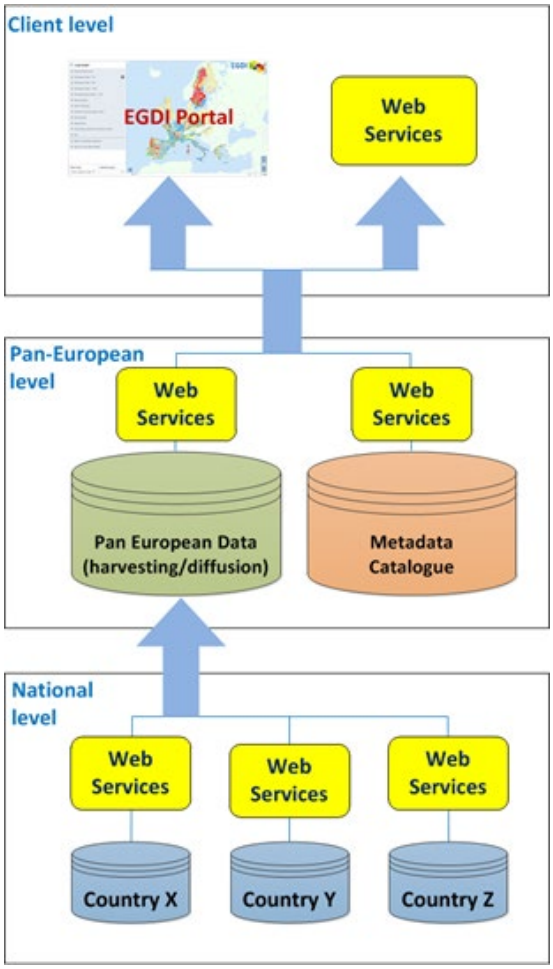

Figure 22 - Harvesting approach in EGDI

The portal and web maps use a three-tier architecture:

- A presentation layer using Wordpress and server-side application logic
- A business layer handling translations from database registration to browser output
- A persistence layer (database) using PostgreSQL database

The above can be illustrated as follows:

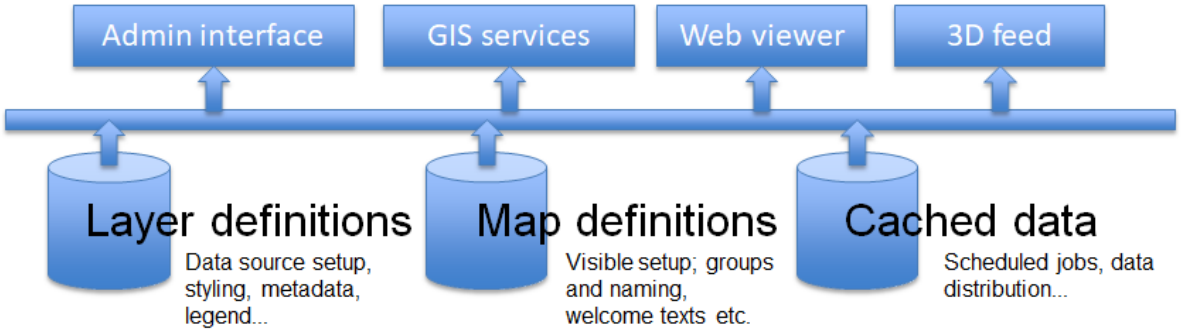

Figure 23 - Interactions between components and database in EGDI

Note: Currently, only the web viewer interface for presentation is supported. The rest is under development.

When a data provider wants to register their data products in EGDI, the following steps shall be taken:

1. Create a data export, shapefile or georeferences raster product
2. Define data layout, styling, legends, titles, attribute list
3. Define a use case for the web map including web map menu, groups, titles, layer filters
4. Register above in EGDI (data and metadata)

For the time being, step 4 shall be done manually by EGDI developers.

2.4.2.2 The EGDI Metadata Catalogue

The EGDI Metadata Catalogue is the central access point to metadata concerning structured geological data. It provides tools for compilation of those metadata in a standardized format that will allow users to effectively search through the metadata catalogue. In order to make the data discoverable in the most efficient way, the catalogue is fully compliant with international standards and supports the distributed system of metadata administration. Only digital and structured information (spatial datasets or dataset series, spatial data services - WMS, WFS and web applications) are described by metadata in this catalogue. In order to display a metadata record for which an on-line map service is available, the Metadata Catalogue is integrated into the EGDI Portal. The catalogue enables systematic discovery, view and use of geological data across Europe. Metadata from previous EU projects such as OneGeology-Europe and Minerals4EU are included.

The catalogue is based on Micka software (version 5). In addition to basic CSW functionality, the GeoDCAT-AP, KML, Atom, OAI-PMH and other outputs are available. The EGDI metadata profile is compliant with the requirements of the INSPIRE Directive for the metadata and the EN ISO 19115/19119 standards. Metadata are freely accessible to the public for viewing and searching, but inserting and editing is for authorized users only. Functions of transactions and harvesting are also supported.

The EGDI Metadata Catalogue is available from the URL addresses: http://egdi.geology.cz/ or http://www.europe-geology.eu/metadata/.

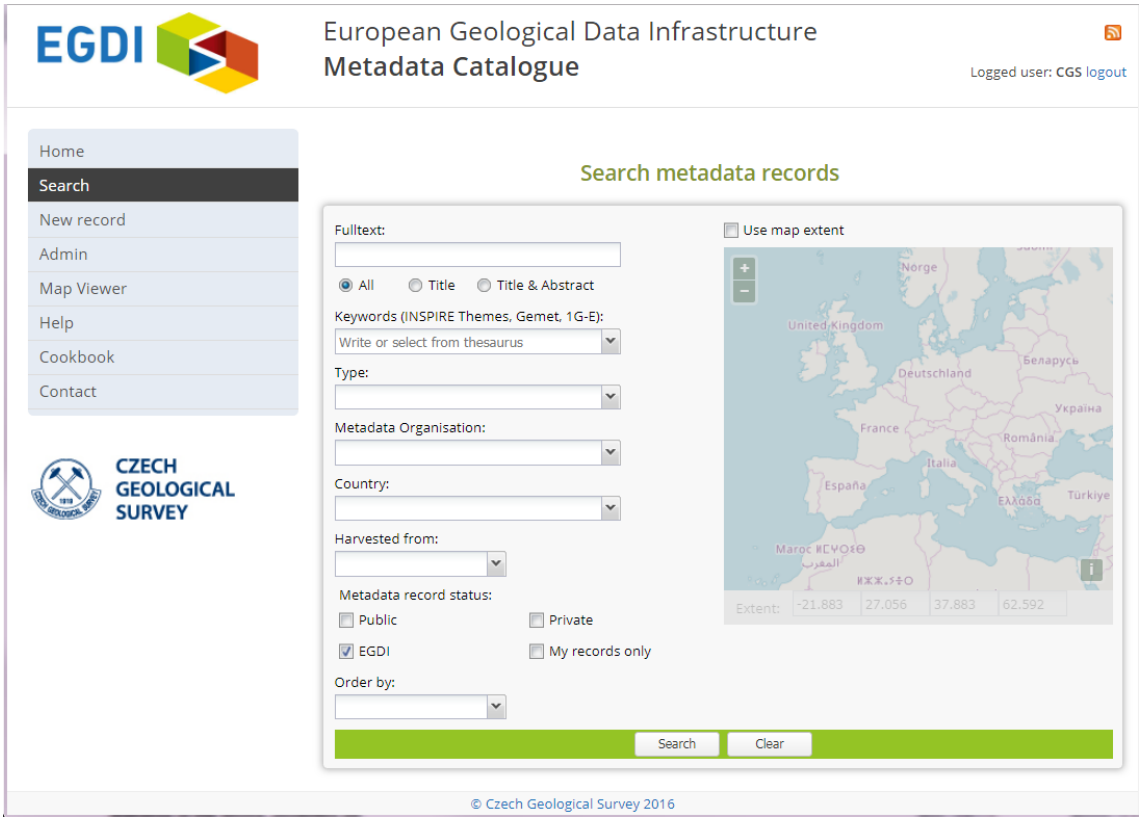

Figure 24 - The Search page of the EGDI Metadata Catalogue for editors

The EGDI metadata content:

- **Services from past projects** presented on the EGDI Portal (OneGeology-Europe, Minerals4EU, ProMine, EMODNET, PanGEO, Terrafirma, GEMAS, IGME 5000, IHME 1500, Geomind, EuroGeoSource eENVPlus etc.) – "Top level" metadata
- **Datasets** related to abovementioned services - "Top level" metadata
- Important national **datasets** and **services** according the geological survey´s priorities
- **Applications** to access to some important national data (boreholes etc. – according the geological survey´s priorities)

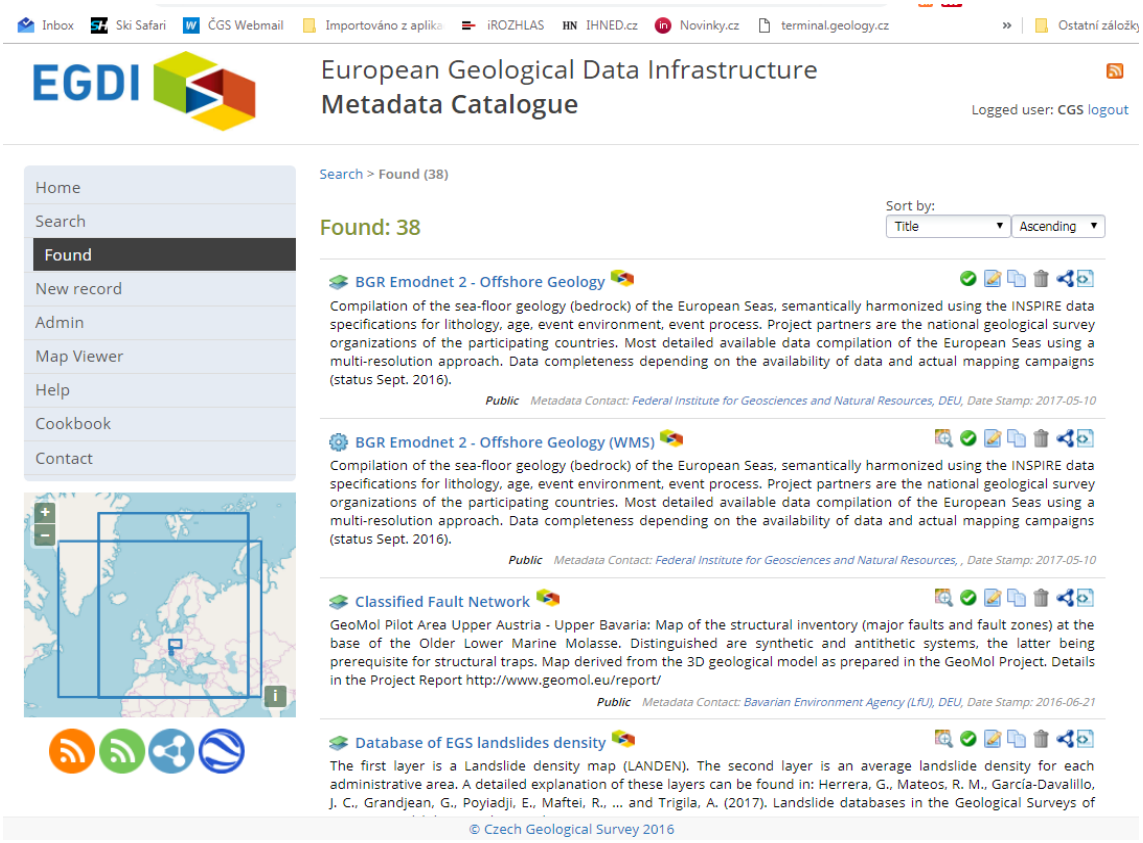

Figure 25 - EGDI catalogue: “Top-level” metadata

Life cycle of the metadata

- Scenario 1 (preferred): The participating organization (GSO) has its own national metadata catalogue. Selected metadata within their catalogue which are denominated by the keyword **EGDI** are harvested to the central EGDI metadata catalogue (more keywords are applicable). Harvesting rules may differ between organizations or catalogues.
- Scenario 2: The participating organization (GSO) does not have its own catalogue. Metadata can be inserted and edited directly on the EGDI catalogue.

The tool used for EGDI (Micka software) is now available as Open Source (Open Micka is a web application for management and discovery of geospatial metadata). Its version 6 is available, the figure below provides a screenshot of the new interface.

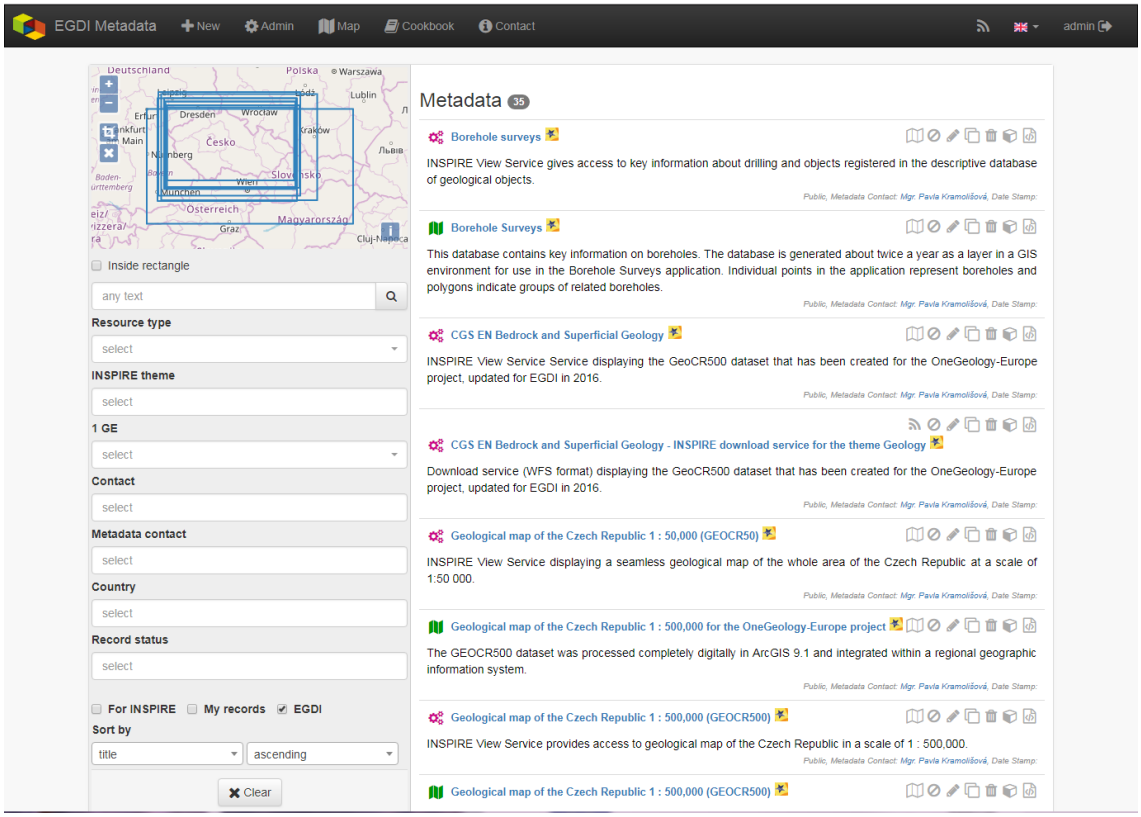

Figure 26 - Test EGDI Metadata interface based on new version 6 of Micka Catalogue

2.4.3 EPOS Thematic Core Service Geological Information and Modelling (TCS GIM)

EPOS TCS GIM architecture is based on a pure OGC service stack and implements, extends and define domain standards applying ISO 191xx series of standards

Its architecture aims to expose data index only to EPOS central part (named EPOS ICS-C). This choice, based on a decade of EU and international interoperability projects aims at lowering the entry barrier for data providers, simplify data exchange within the infrastructure and implement as fully as possible linked data principle on top of a classical OGC service stack (CSW, WMS, WFS).

The images below are extracted from an EPOS consortium meeting in March 2018 and provide a good overview of EPOS TCS GIM architecture.

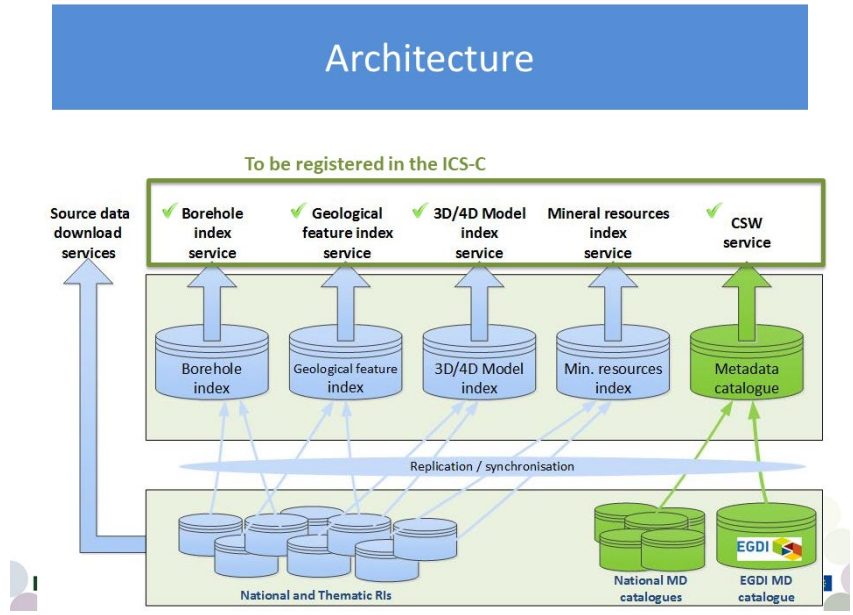

Figure 27 - EPOS Architecture

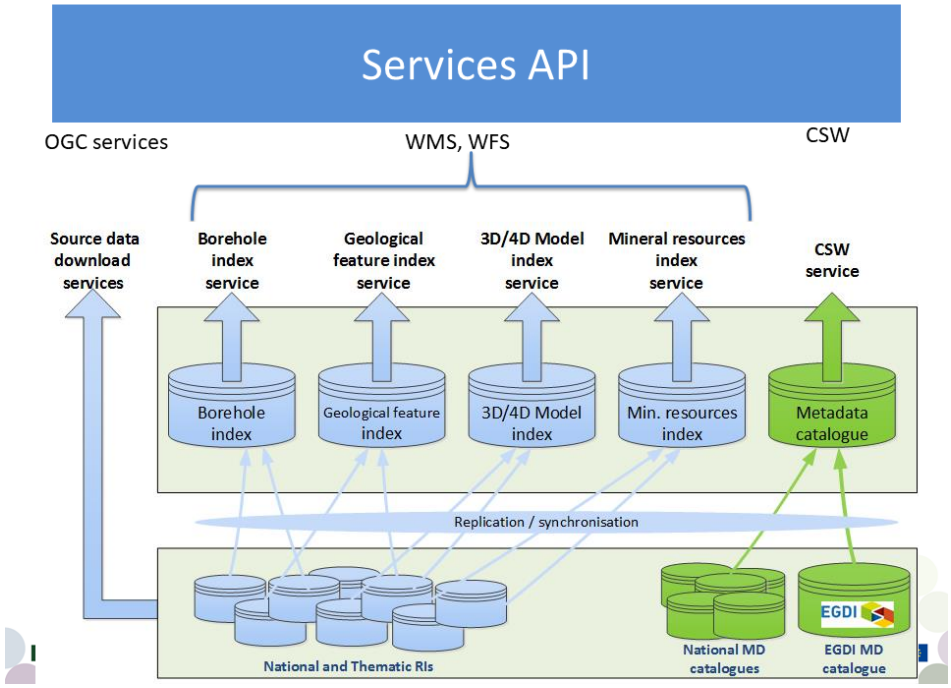

Figure 28 - EPOS Services API

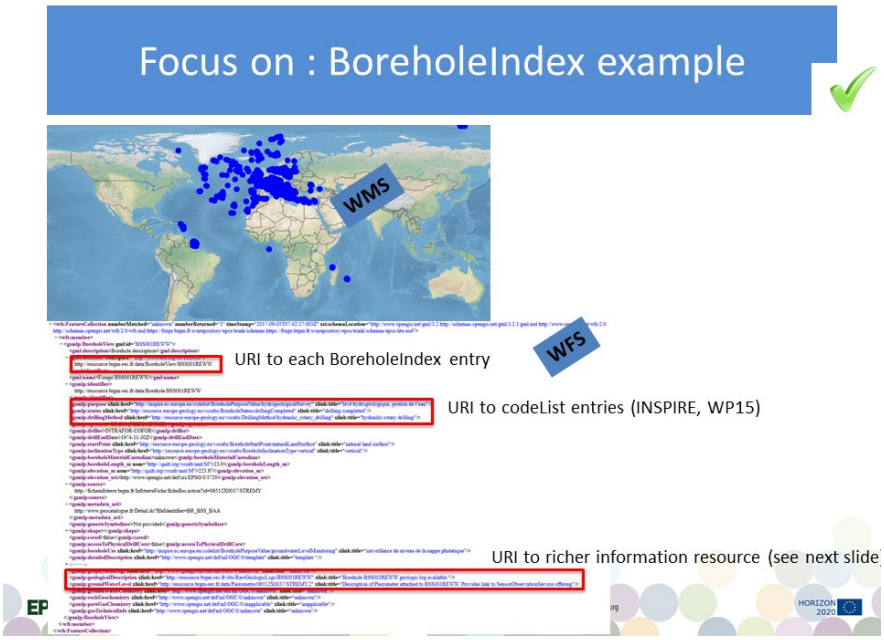

Figure 29 - EPOS BoreholeIndex example

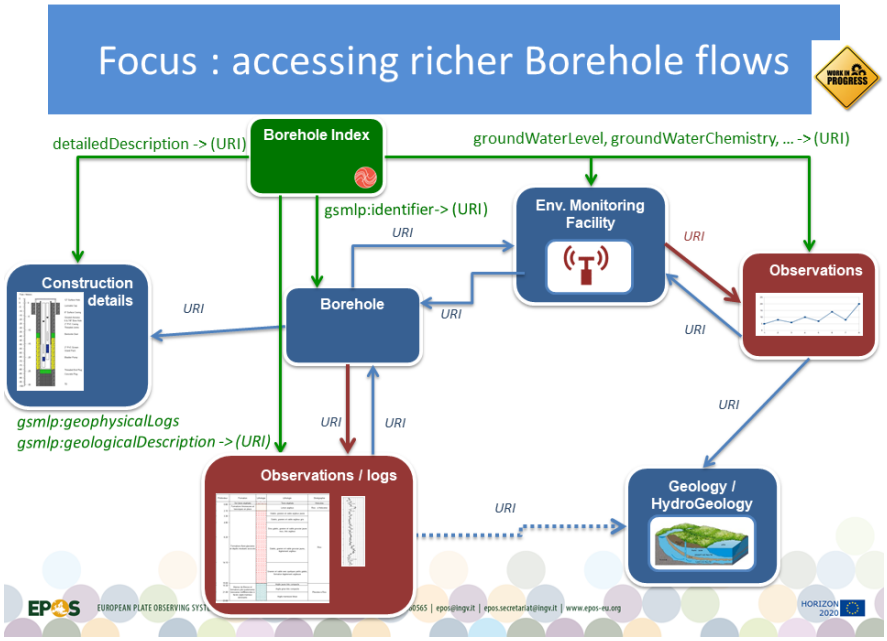

Figure 30 - EPOS accessing richer Borehole flows

From a software point of view

- CSW is provided by a GeoNetwork.
- WMS is made available using MapServer 7.2.

- WFS is provided using GeoServer 2.14.1 with app-schema and SolR extension. SolR extension interaction with app-schema extension was funded by BRGM in order to enable exposing with decent performances the EU Borehole Index.
- Back end database is PostgreSQL/GIS 9.x at the time the system was deployed.

The use of EPOS GIM the infrastructure with a QGIS desktop as been presented at EGU in 2018 within the EPOS booth (the video is online on QGIS GMLAS GitHub).

3 CONCLUSION

This deliverable content will be consolidated over the course of the project.

The overall central system described here completes the information given in the D5.1 on the blueprint and aims to precise the components that may compose the Information Platform. It will be updated iteratively based on GSPs' user requirement and GIP-P WPs feedback.