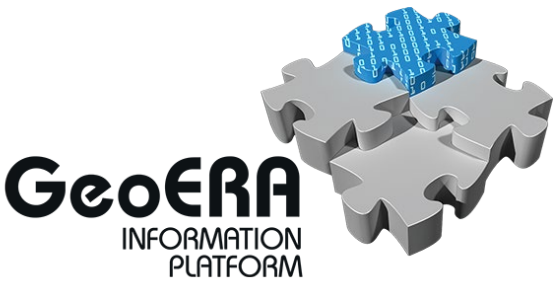


Establishing the European Geological Surveys Research Area to deliver a Geological Service for Europe

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

Martin Schiegl [GBA]
Patrick Bell [BGS]

Deliverable 5.1
Preliminary version

GIP blueprint: data and services architecture

E-mail of lead author:
f.tertre@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	D5.1 – version 1.1	
Dissemination level	Public	
Deliverable name	GIP blueprint: data and services architecture	
Work package	WP5, Architecture	
Lead WP/Deliverable beneficiary	BRGM	
Deliverable status		
Submitted (Author(s))	8/10/2019	Sylvain Grellet
Reviewer	10/10/2019	Marc Urvois
Approved (Coordinator)	21/10/2019	Jørgen Tulstrup

GENERAL INTRODUCTION

The interaction between the GeoERA Scientific Projects (GSPs) and the GeoERA Information Platform Project (GIP-P) starts with the sharing of data and services. During the previous European Project made in the three themes covered by GeoERA, some of these problematics have been studied, different approaches have been developed and the GIP-P will be built on these experiences. First of all the GIP-P must support the GSPs in delivering their results and it will do that by building on the European Geological Data Infrastructure (EGDI) and enhance that to meet current Best Practices (W3C, OGC…) in information sharing where possible.

In parallel of GeoERA, other initiatives exist that work on the same issues and the GIP-P has to consider i/ if it will capitalise these initiatives and built on their lessons learnt and ii/ if it will integrate with these external initiatives.

At an international level, standards exist to share data and to provide and use services: standards for technical interoperability and for semantic interoperability. The use of these international standards is part of the five stars scheme[1] which tend towards a better description and use of data.

Finally, this document will also present what will be the way to serve data and documents to the EGDI. It will present how to serve data with their metadata and how serve documents that will become part of the EGDI.

Deliverable D 5.1 (blueprint document) target is to frame the DNA of the GeoERA Information Platform Project (GIP-P). It will be updated iteratively based on GSPs' user requirement and GIP-P WPs feedback. Every other deliverable of the project shall be aware of its recommendations and take them into account if possible.

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 blueprint 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 in the framework of the EGDI enhancement using operational harvesting mechanisms and a recently updated MICKA metadata catalogue.

[1] https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 AN ADJUSTMENT OF PARADIGM -> UPDATED BEST PRACTICES

1.1 Former approach & its limits

The former projects developed during the last years to promote the diffusion of the data for the different thematic domains of geoscience (some of these projects are described in Annex C) use a set of techniques to fulfil the obligation of interoperability and harmonization between countries of Europe.

Regarding spatial data, the work made by the international groups in the OGC (Open Geospatial Consortium) has led to a bunch of standards (firstly, technical standards, then thematic standards) that can be implemented by different organisms to share their data.

In Europe, the Commission, through the INSPIRE directive, has edited some rules to follow for the diffusion of European environmental data, and, to help the users that must conform to these rules, has edited some technical guidelines which rely on the use of OGC standards.

Following the different stages proposed by INSPIRE, 4 levels can be found:

- Search (i.e. have a catalogue of metadata that describe the existing data)
- View (i.e. have the possibility to view the data of a provider)
- Download (i.e. have the possibility to download the data of a provider - most of the time, to perform some processing on it)
- Transform (i.e. have the possibility to transform the data from one state to another - change of Coordinate Representation System, thematic transformation…)

1.1.1 *Search*

This level allows to identify data or services through the use of catalogues. The data or services must be described by metadata.

Metadata can be explained in few ways:

- Data that provide information about other data.
- Metadata summarizes basic information about data, making finding and working with particular instances of data easier.
- Metadata can be created manually: to be more accurate, or automatically: but may contain more basic information.

Metadata must answer to the 6 W's (or 5W1H):

- What (thematic)
- When (temporal coverage)
- Where (spatial coverage)
- Who (data creator/provider/contact point)
- How (the data has been created, what is its history)
- Why (the data has been created)

Metadata must accompany all the data provided to the EGDI.

Metadata must be described according to the recommendation of INSPIRE (see https://inspire.ec.europa.eu/metadata/6541).

In past projects, the approach was to use OGC CS/W (Catalogue Service for the Web) with a definition of the records in ISO-19139.

1.1.2 *View*

The second level, as proposed by INSPIRE, allows to visualize data directly on a map (most of the time through Web GIS). These data can have been discovered by a search service or directly by its URL. This level of service allows to co-visualize data, to move, change scale, zoom in and out, to see legends and to query for points on the map.

Only visualization is available within this level of service, implying a limitation of the number of action a user can do with the data.

In past projects, the approach was to use OGC WMS (Web Map Service) to propose the visualization of data.

1.1.3 *Download*

The third level allows to download the data, partially or entirely, to be able to process or reuse them afterward.

In past projects, the approach was to use OGC WFS (Web Feature Service). The WFS can propose different level of features, the easiest one, Simple Feature (SF-0, SF-1, SF-2), is used to represent flat data (property/value). This level allows representing a single table of a database or a geospatial file and is easily usable by a non-expert user. The more complex one, Complex Feature, is used to represent complex data models with multiple concepts and links between these concepts. The complex features allow to implement data models like GeoSciML or EarthResourceML. A better representation of the domain is possible with Complex Features, to the detriment of the ease of use.

1.1.4 *Transformation*

The last level allows to transform spatial data set, most of the time to improve interoperability. It can be transformation of coordinate system (to have be able to use at the same time two data sets), but also thematic services.

1.1.5 *Limits*

The approach described in the first part of this chapter shows now its limits. All these techniques were a good strategy to share data in an interoperable way, but limit themselves to advanced user that can have an intimate knowledge of all the specifications in order to interact with the data services (CS/W, WFS, SOS …).

The example of use of metadata catalogue is symptomatic. The standards, CS/W to communicate with the catalogue, ISO-19115 and ISO-19119 to write metadata for spatial data and spatial data services, are not easily understandable and require a big effort to be, not mastered but, at least used in a proper way.
The use of thematic standards (like GeoSciML, EarthResourceML…) on Web Feature Services raise the same issue. These thematic standards can be quite complex and result in a lack of

understanding of what is represented if the user doesn't have a good knowledge in the thematic and of the API to be queried (here WFS).

1.2 W3C/OGC collaboration -> W3C Data on the Web / Spatial Data on the Web

For many years, the W3C has been leading a considerable collaborative effort supported by many actors from both private and public sectors. This effort consists in providing a set of technologies, standards and best practices that constitute the basis for representing, publishing and sharing the data over the Web in a standardized way. The set of technologies and standards proposed by the W3C, also known as the "Semantic Web" technologies (*cf.* the Semantic Web Layercake), are embodied by the four simple principles, known as the "Linked Data" principles, outlined by Tim Berners-Lee:

1. Use URIs to name (identify) things.
2. Use HTTP URIs so that these things can be looked up (resolvable, "dereferenceable").
3. Provide useful information about what a name identifies when it is looked up, using open standards such as RDF, SPARQL, etc.
4. Refer to other things using HTTP URI-based names when publishing data on the Web.

The so called the "Web of Data" is the result of the adoption of these principles by data providers over the world. It is not intended to be a substitution to the already existing "Web of Documents", but to complete it in order to achieve the original vision of the Web as proposed by Tim Berners Lee.

In additions to these principles, Tim Berners-Lee has also proposed a "5 star deployment scheme" for open data to operate a rating system for Linked Open Data. The goal of the scheme is to urge the data providers to follow the best practices that would enhance the visibility and usability of their data over the Web. More details about this scheme are in annex A.

1.2.1 *The Semantic Web technologies (layer cake)*

The semantic Web layer cake, proposed by the W3C, provides a complete scheme of specifications that describes the different layers for identifying, structuring, sharing, interrogating, documenting and reasoning on the data over the Web. We detail below the main technologies involved in the Linked Data paradigm.

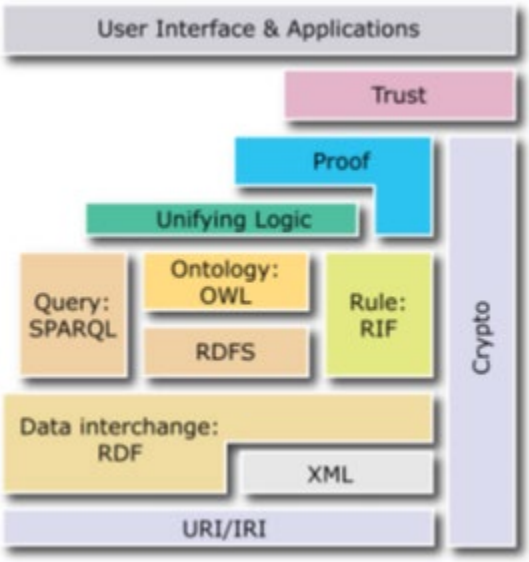


Figure 1 - Semantic Web layer cake diagram « https://www.w3.org/2007/03/layerCake.png »

1.2.2 *Why HTTP URIs? (W3C, 2014) (RFC3986, 2005)*

URI (Uniform resource identifier) is a compact sequence of characters that identifies an abstract or a physical resource. We mean by resource anything that can be denoted from the "universe of discourse", i.e. which can have an identity. This includes informational entities such as documents, media files, etc. or non-informational entities such as persons, concepts, physical objects, etc. The uniformity of URI, guaranteed by common syntaxes, provides a mechanism for unifying the context of the identifiers independently from their access mechanism. When well chosen, the syntax also provides a uniform semantic interpretation for the deferent objects and their types.

In addition to identification, URIs based on the well-established HTTP Web protocol provide a way to lookup the description of the identified entities over the Web.

The duplicity of utility of URIs constitutes the foundation of the Web of data. However, the first question one could probably ask is what should we expect as content when we lookup the URI, whether it is information in a form that we could understand or data that is in a form that we could process automatically. In fact, one of the strengths of the HTTP protocol is to provide a mechanism for specifying the wanted format of the content behind a URI. It is known as the content negotiation mechanism. The subjectivity of what is the best representation to retrieve is the main reason behind the existence of such mechanism.

This triggers a fundamental concept of differentiating the entity from its representation(s).

1.2.3 *Content negotiation*

Content negotiation is "the mechanism for selecting the appropriate representation when servicing a request. The representation of entities in any response can be negotiated (including

error responses)" (RFC2068, 1997). The content may vary on many dimensions: file format (e.g. in Figure 2), language, content coding, etc.

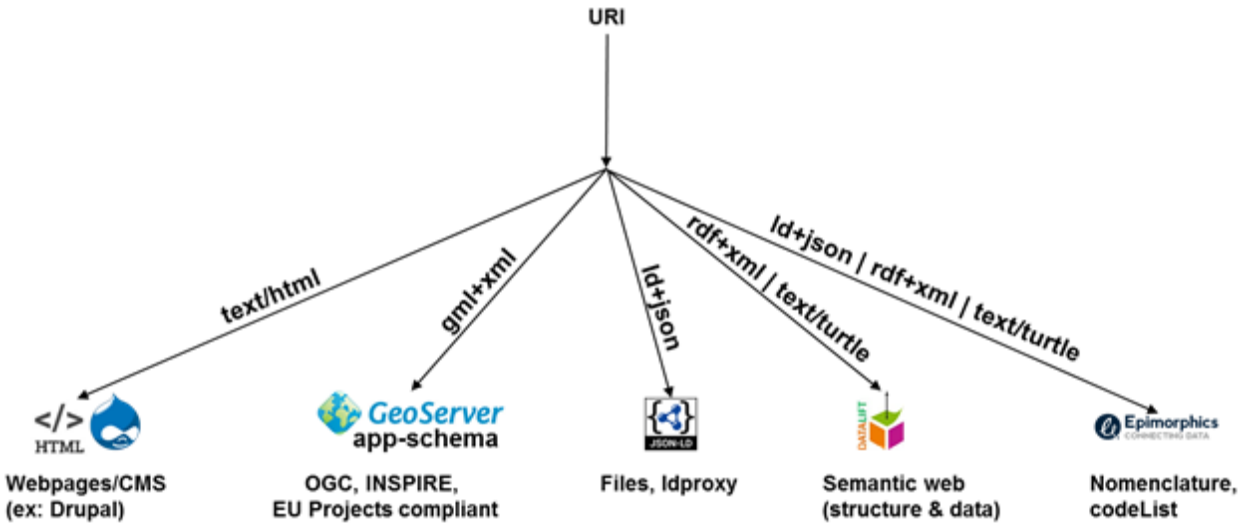


Figure 2 - Content negotiation

This mechanism can be mainly achieved through two distinct methods: server-driven and agent-driven negotiation.

The server-driven mechanism allows the selection of the requested representation thanks to a program located at the server level. This program, based on the content of particular header fields (Accept, Accept-Language, Accept-Encoding, etc.) of the request sent by the user agent client, or trying to guess the best response representation if the required header fields lacks of values, will provide the most relevant representation. It presents the advantage of the simplicity of sending the response to user client without including it in a back-and-forth negotiation delay. The main disadvantage of such a method is the impossibility to accurately determine what the best response for the client agent desires.

In contrast, in the agent-driven mechanism the user agent performs the selection of the best response after receiving a first response from the server. This first response contains a set of available representations included in the header fields. The user agent can then select automatically the desired representation of give the choice to the user to do it manually. This method is advantageous when the server presents large set content possibilities. However, it has the disadvantage of the delay caused by the need of two-request mechanism.

The combination of these two methods is called "Transparent Negotiation".

1.2.4 *RDF (Resource description framework)*

RDF is framework for representing information in the Web (W3C, 2014). It is a model that describes the data as an oriented tagged graph. The core structure is representing any set of statements by triples (consisting of a "subject" a "predicate" and an "object") (Figure 3). The subject is a resource identified by its URI. The predicate is a URI of a property identified described

in a predefined vocabulary. The object or may be a literal value resource (identified by its URI) wither it is from the same data source or an external one.

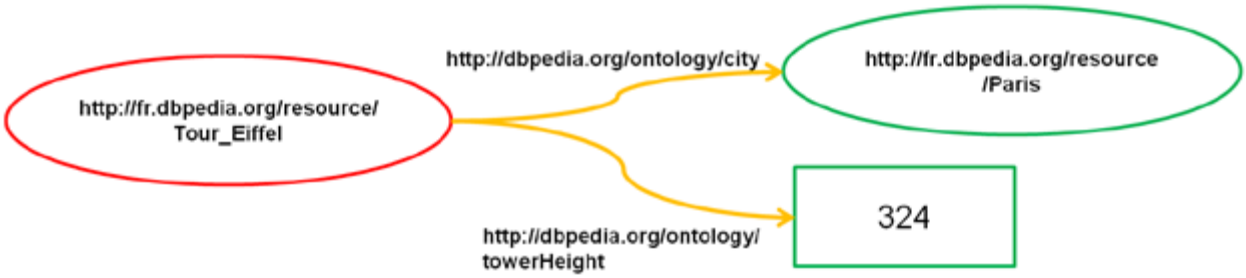


Figure 3 - RDF triples example

This graph-based data model can be serialized in many formats such as RDF/XML, Turtle, N-Triples, N-Quads, N3, JSON-LD, RDFa, etc.

The properties that ensure the structure of the this graph that are defined in vocabularies (ontologies), together with the "classes" of the resources defined also in ontologies, are the main components that provide semantics to the graph.
We commonly define an ontology as "a specification of conceptualization (Gruber, 1995)". In computer science, it is more specifically a referential that allows the creation of knowledge bases by providing a logical definition of the different concepts (terms, classes, etc.) of a universe of discourse and the different semantic relations (hierarchical or others) that can exist between them. To represent an ontology (or a vocabulary), many languages that differ in their formality, their level of expressiveness and their complexity are proposed, such as SKOS, RDFS and OWL.

1.2.5 *Sparql*

RDF data can be interrogated and manipulated thanks to SPARQL query language. It represents for RDF what is SQL is for relational data. This language is based on the use of triple patterns to select a set of data or rebuild a subgraph of data.

1.2.6 *W3C Data on the Web and Spatial Data on the Web working groups*

Data on the Web working group released its first public working draft document in February 2015 which became a W3C recommendation on January 2017 31st (*https://www.w3.org/TR/dwbp/*). This set of recommendations aims that 'Data should be discoverable and understandable by humans and machines'. They are summarized in annex B.
Building on this, a joint W3C/OGC working group defined the spatial data on the web best practices. They are also summarized in annex B.

In 2017, a OneGeology Linked Open Data Workshop already did the exercise of analyzing those recommendation. This blueprint document takes the first set of identified best practices as a starting point. They are summarized in the table below where 'BP' stands for data on the web best practice and 'SBP' for spatial data on the web best practice (the number corresponds to the one used in the W3C doc).

Best practice	Topic
BP1	Provide metadata
BP9, SBP1	**Persistant URI / Unique global id**
BP10	Use Persistant URI within dataset
BP14	Multiple formats
BP15/ SBP10	**Reuse vocabularies (code list registries, ontologies)**
BP17	Bulk download
BP18	Subset of large dataset
BP19	Content Negotiation
BP22	Explain missing data
SBP2	**Make it indexable by search engines**
SBP12	Expose data through APIs

Table 1 - OneGeology analysis of W3C recommendations

Some of them are deemed of the utmost importance (they are in bold in the table)

- "BP9, SBP1 : Persistent URI / Unique global id" which imply assigning persistent and resolvable URI to dataset, dataset entries, metadata, APIs
- "BP15/ SBP10: Reuse vocabularies (code list registries, ontologies)" which implies taking the 'open world assumption' and also expose it as codelist in registries and vocabularies using URIs
- "SBP2: Make it indexable by search engines" which amongst several solutions target JSON-LD

1.3 Recent trends in OGC specifications

Stemming from the work initiated by W3C/OGC collaboration, OGC specifications are progressively moving to a more 'webfriendly' approach.

As such WFS3 and SensorThings API are targeting a more 'RESTful' behavior.
WFS 3 group went further as the Core specifications first draft release is currently available as an OpenAPI specification. This draft specification is under public comment and the stable version is expected mid-2019. Implementations of the draft core are already available and being tested within many institution (as it can be seen on the WFS_FES GitHub).

SensorThings API is progressively moving to its version 1.1 and is already proposed to be a valid observation download service for INSPIRE (see: https://doi.org/10.3390/geosciences8060221).

Such a change in OGC specification from the XML/XSD (and KVP, POST, REST + SOAP) to a behavior considered modernized shows how deep the impact of the collaboration with W3C and the will to meet web developers expectations are.

There is still a way to go but slight adjustments will allow datasets that are currently 'hidden' behind current web services to be directly visible to the web of data, thus indexed, thus more reused.

However, it is way too early to consider the previous way of exposing geoscience dead and replace it completely with the new one being shaped. This also implies GeoERA Information Platform Project (GIP-P) needs to consider both in its deployment

2 TARGET SYSTEM

Depending on the needs arising from WP2 various system architectures can be foreseen.
Note that:

- in all those architectures, the final client is not always a 'central system' or a 'central system map viewer'. It can just be the outside world in which a machine, a scientific code could reuse the exposed datasets or just a search engine robot in charge of indexing new data,
- all the proposed architectures consider that the initial data provider may not have the IT capacity/know-how and propose an alternative for data publication (see: the blue box 'Shp -> WFS "Cloud" in the figures describing each architecture option).

2.1 Shared elements

All the foreseen architectures share common elements that don't appear in the diagrams below for readability sake:

- Shared data and services specifications available to all on a common place
- A codelists registry tool to share codelists
- Metadata (ISO 19115/139) for services, dataset
- Linked-data (URI on features, codelists) with URIs that actually resolves to something.
- Synchronization (up-to-date data): Pub/Sub VS Repeated harvesting
- Performances aspects: caching may be needed …

- And capacity building via sharing tools (and configuration) and practices (workshop, …)

2.2 Situation A: Direct access to Data provider system 'à la ONE-Geology'

In this situation the client can directly access data provider services.

One running example of such an approach is the One Geology portal (http://www.onegeology.org/use/portal.html).

This is one of the simplest architecture, it just requires the shared elements listed below.

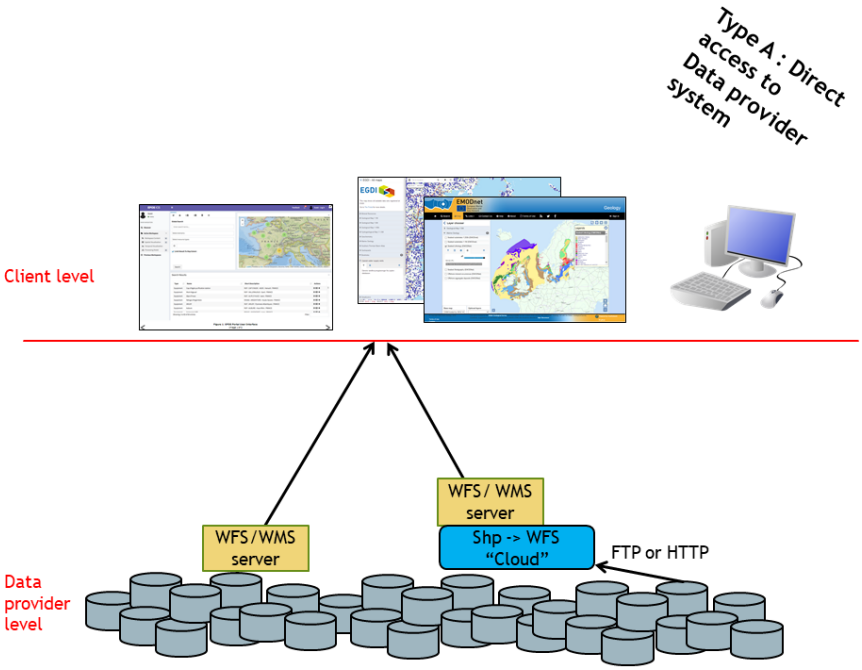


Figure 4 - Situation A: Direct access to Data provider system

2.3 Situation B: SimpleFeature / Index approach 'à la EPOS TCS Geological Information and Modelling'

In some other cases there is a need to collect summary information from the data providers in order to consolidate the information in a single place to:

- provide a unique EU data endpoint for a given feature type
- and re-expose it according to various representations (XML, JSON-LD, triple store, …).

In this case, the central system only handles Simple Feature when it harvests or expose content.

One running example of such an approach is the central node deployed by EPOS Thematic Core Services Geological Information and Modelling (TCS GIM).

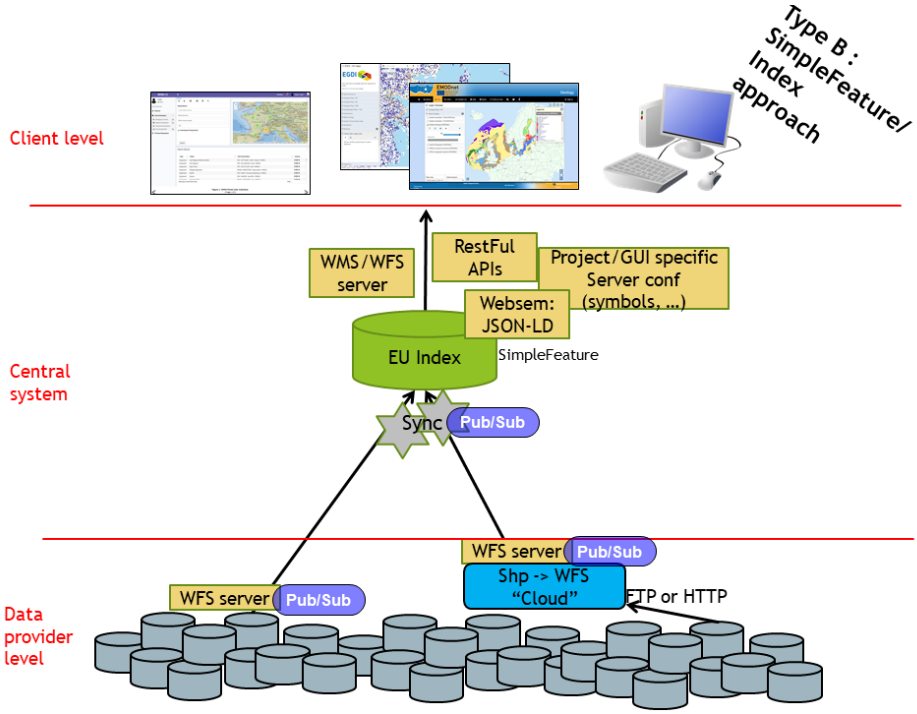


Figure 5 - Situation B: SimpleFeature/Index approach

In EPOS TCS GIM the EU index content points in turn to more detailed (complex) flows available from the data provide but does not require them to be harvested.

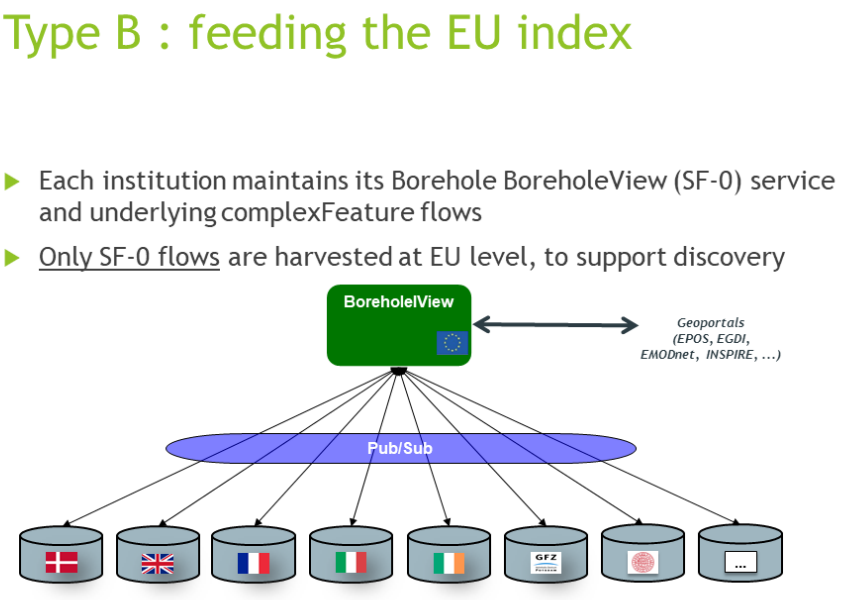


Figure 6 - Situation B: feeding the EU index

Lessons learnt from this exercise in EPOS TCS GIM (for example on borehole features) are that:

- Exposing summary Information (SimpleFeature SF-0)[2] is easier/faster than complex Feature which enable to collect content from more data providers.
- Harvesting SimpleFeature is more reasonable than complex Feature. WFS on complexFeature is not meant/suited to synchronize millions of instances (~ DB dump)!
- It has been considered by the community (EPOS TCS GIM and OGC GeoSciML SWG) that this SimpleFeature representation acted as a kind of vCard for a borehole instance (~ vCard to exchange contacts between e-mail clients). Harvesting a 'Borehole vCard' just for discovery makes more sense than a comprehensive and complex borehole description.

BRGM tested the complete implementation on linking from the Borehole Index entry (Boreholeview in the above schema) to complexFeature flow using GWML2:GW_GeologyLog, INSPIRE:EnvironmentalMonitoringFacility, WaterML2. The linked data approach was validated and presented to OGC GeoSciML SWG, GroundWaterML2 SWG and both Hydrology and Geoscience Domain working groups.

2.4 Situation C: When the central system produces new / restructure information 'à la Minerals E4U'

Eventually, in some specific situations there is a need to dissociate a harvesting system and a diffusion one which produces new/enriched content.
In this case, the harvesting system often consumes complex feature flows, provides Quality Analysis / Quality Check on it and then its data is pushed to a diffusion system in charge of generating new content.

One running example of such architecture is Minerals4EU.

This architecture could also be relevant even if there is no need for producing new or restructuring information in situations where performances issues arise and are not solved by distributed system approaches.

[2] See Chapter « 1.1.3 Download »

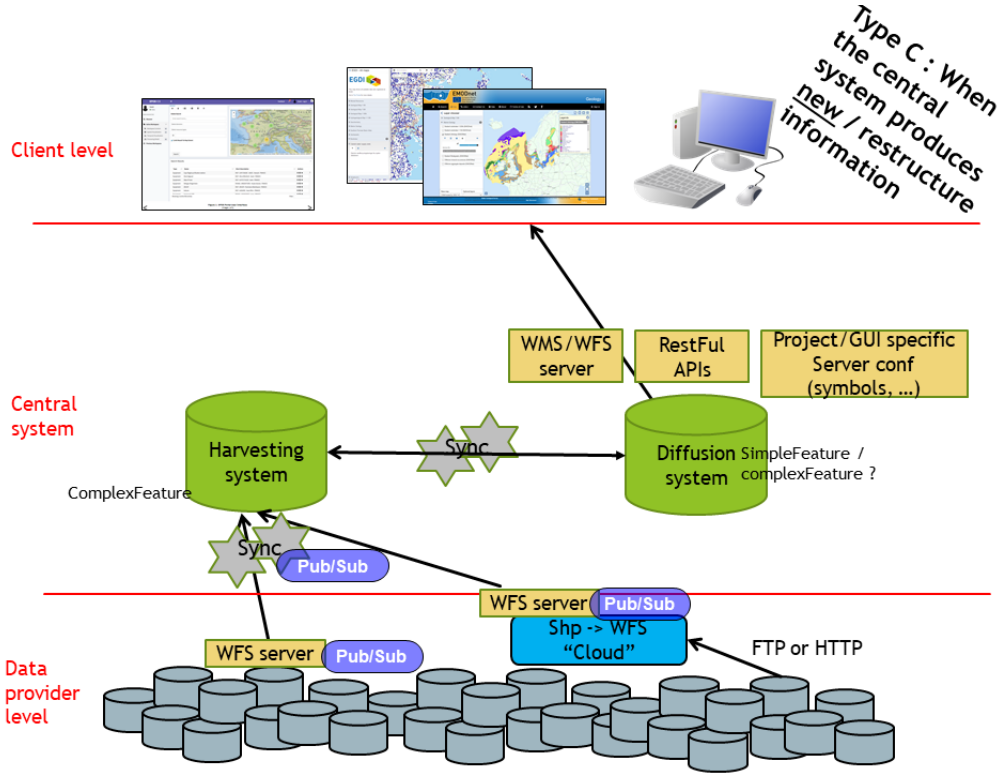


Figure 7 - Situation C: When the central system produces new/restructure information

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

During the last years, multiple projects on the thematics of GeoERA have been done. Beyond the change of paradigm, some functionalities already exist and may need only to be adapted or completed to fit the requirement of GeoERA. These former projects are described in annex C and the functionalities that can be retrieve from them are presented in the following table.

To evaluate the maturity of the functionalities, we use the Technology readiness levels (TRL), a method of estimating technology maturity of products. Initiated by NASA in the 70's, it was then further canonized by the ISO 16290:2013 standard and later adopted by EC in 2014 as a reference for Horizon 2020 program. This is the scale used here after.

On that scale, we will judge levels 1 to 4 as laboratory results, 5 and 6 as products that already have proven their use, 7 as products with proven experience in operational but not with high constraints, 8 and 9 as near final products.

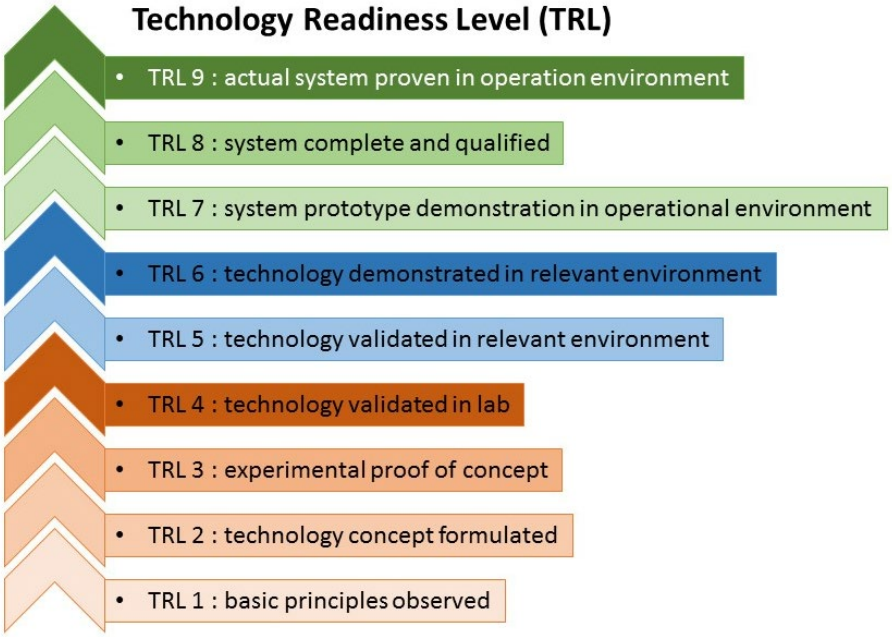


Figure 8 - Technology Readiness Level (TRL)

Functionality	Project(s)	Maturity	TRL
Data models for Mineral Resources (ERML/INSPIRE MR)	Minerals4EU (& derivative) ProSUM (for Mining Waste)	Mature	TRL 6

Data models for Geological information (GeoSciML/INSPIRE GE)	One Geology global, Europe, EGDI	Mature	TRL 7
Data models for Hydrogeological information (GWML2/INSPIRE GE-Hydrogeology, EF, AM)	At some NGSO level over the world (ex: BRGM, NR-CAN USGS, GNS...)	Mature	TRL 7
Data models for GeoEnergy information (INSPIRE ER)	ERA Net Geothermy	Proof of concept	TRL 4
Database for handling Mineral Resources data (conforming to ERML data model)	Minerals4EU (& derivative) ProSUM (for Mining Waste)	Mature *use PostgreSQL/PostGIS*	TRL 7
Data models for Boreholes and associated information conceptual and logical (EPOS Borehole Model, GWML2, INSPIRE EF)	EPOS GIM	Mature	TRL 6
Data Model for discovery of 3D/4D Models	EPOS GIM	Mature	TRL 6
Web Feature Server *for complex features*	Minerals4EU (& derivative) EPOS TCS GIM	Mature *for MR, a Deegree configuration already exists otherwise, GeoServer is now more efficient with complex features.* *For Indexes (Borehole, GeologicUnit, MineralResources…) GeoServer application schema configuration are shared by EPOS GIM*	TRL 6
Data Index services	EPOS GIM	Mature	TRL 7
Implementation cookbook	Minerals4EU	Outdated	TRL 5
Harvesting system	Minerals4EU (& derivative)	Mature	TRL 7

	EPOS GIM (SimpleFeature and Pub/Sub oriented using Apache Kafka)		
Web Map Server	Minerals4EU (& derivative) EPOS GIM EGDI	Mature	TRL 9
Map Viewer	Minerals4EU (& derivative) EGDI OneGeology	Varied	TRL 5 to TRL 9
Knowledge base (documents/references storage)	Minerals4EU (& derivative)	Mature	TRL 6
Search (in data and/or documents/references)	Minerals4EU (& derivative)	Mature	TRL 6
Knowledge base (ontology based)	MICA	Prototype	TRL 5
Content Management System	Minerals4EU (& derivative)	Mature	TRL 9
Metadata catalogue	Minerals4EU (& derivative) EGDI EPOS GIM	Micka: Mature	TRL 7
		GeoNetwork (implemented in EPOS GIM): Mature	TRL 9
Borehole and associated data complex features	EPOS GIM	In maturation for the topic of Boreholes	TRL 4
CodeList Registry Tool	EPOS GIM Minerals4EU	Mature	TRL 7
URI Scheme that resolves to EU data specification, API endpoints, CodeList Registry doing content and semantic negotiation (partially for the later)	EPOS GIM and Minerals4U	Mature for EPOS GIM, partially mature for Minerals4EU both under https://data.geoscience.earth/	TRL 6

DataProvider URI scheme that resolves to features/object instances	EPOS GIM	Mature depending on the data provider	
Datasets and services metadata according to DCAT and EPOS ICS-C requirements (EPOS_DCAT_AP)	EPOS GIM	Mature	TRL 5

Table 2 - Functionalities from former projects

4 REACHING THE TARGET - USE OF DATA STANDARDS

Based on the various GeoERA domain projects, several standards and standardization dynamics can be pre-identified. Some are stemming from European communities (e.g.: extending around INSPIRE data specifications), some are driven by international communities broader than EU only and some benefit from a real 'symbiosis' of both dynamics.

The list below will need to be polished based on the precise needs arising from WP2.

It has to be specified that, not always, a data standard covers de facto all the needs arising from domain experts and that those extra needs/usage might trigger update to the corresponding standard to allow it to cover a domain and also to be widely used by a community.
Some GeoERA Information Platform Project (GIP-P) members are already involved in the various standardization bodies of interest for the project. They will be able to extend the identified standards to meet domain projects needs and propose the corresponding evolutions to the relevant standardization bodies.

- GroundWater data exchange can be achieved using either
  - Various INSPIRE data specifications: Geology - hydrogeology package (GE), Environmental Monitoring Facilities (EF), Area management/restriction/regulation zones and reporting units (AM) etc…
  - and / or OGC:16-032r2 'OGC WaterML 2: Part 4 – GroundWaterML 2 (GWML2)' which is already implemented around the globe and the subject of scientific publications (for example, see https://doi.org/10.1007/s10040-018-1747-9,). During its specification process the EU community was represented (BRGM, JRC LfU-Geological Survey of Bavaria); as a result elements were added to the standard so that it maps to and it compatible to INSPIRE Geology - hydrogeology package (GE). A proposal was also pushed to the INSPIRE maintenance in 2018 to have GroundWaterML2 considered as a valid encoding for INSPIRE data compliant exchange (same rationale as for OGC 16-008:GeoSciML4.1).

- Mineral Resources is already well covered thanks to a continuous international dynamic (mainly of EU projects). This dynamic ensure a continuous evolutions of both the International Standard (EarthResourceML) and its EU counterpart (INSPIRE Mineral Resources theme).

- Geo-Energy is a relatively new topic when it comes to setting up international interoperable data exchanges
  As, to our knowledge, there is no international standard for such exchange, INSPIRE data theme on Energy Resources (ER) is the natural candidate. A previous EU ERA-Net project already explored that path and confirmed INSPIRE ER theme can serve as a basis for such exchanges sometimes including some extensions (see https://doi.org/10.1080/17538947.2015.1073378)

- 3D/4D geological models is also new on those aspects

It has been discussed several times in the OGC Geoscience Domain Working Group (see https://external.opengeospatial.org/twiki_public/GeoScienceDWG/WebHome) and launching a Model Interoperability Experiment was evoked. The consensus was that the most important need was not to standardize the exchange of the 3D/4D model itself but more the description of the existence of a given model in a given area (also that the term metadata was 'misleading').

Some EU projects have explored two different paths:

- extending classical geographic information metadata (ISO 19115/19139) in GeoMol,
- and providing a UML model to describe the 3D/4D (not its 3D/4D elements but mainly how it was produced) in EPOS WP15 following the discussions started in INSPIRE Geology specification - Geophysics extension.

- Description of Observations, Interpretations (and other terms around this topic) is covered by the OGC/ISO standard Observations & Measurements (ISO 19156). By its nature, this standard is widely reused and not always in the geosciences domain.
  Guidelines have been produced within INSPIRE for its reuse (see https://inspire.ec.europa.eu/id/document/tg/d2.9-o%26m-swe).
  It is also implemented in the Semantic Web (om-lite and sam-lite) and W3C communities (e.g.: SSN/SOSA)

The proposed modelling philosophy is the following.
Building on pre-existing EU and international geoscience data standardization dynamics, and to simplify the uptake by data providers, data models will be defined using ISO 191XX series of standards be it when no preceding standard exist and/or when standards need to be extended.

Complementing this approach and when deemed useful for reuse at the GeoERA Information Platform Level (or by any external system), ontology will be generated building on the discussion held at the OGC

- see https://external.opengeospatial.org/twiki_public/GeoScienceDWG/WebHome - OGC TC Orleans - UML to OWL ad hoc meeting (23/03/2018)
- and OGC 18-097 'Environmental Linked Features Interoperability Experiment' Engineering Report (in the Pending documents at the time of writing that deliverable),

This will enable

- on the one-hand to collect data from data providers not asking them to learn a change of IT notation (UML to OWL)
- and on the other hand data re-exposition at GeoERA Information Platform level according to various representation/serialization

5 REACHING THE TARGET - USE OF SERVICES STANDARDS

As presented in chapter one, best practices for Geoscientific information exchange have drastically evolved over the last years (see especially chapter about 'Recent trends in OGC specifications').

In order for the EGDI to be compatible with both trends it is proposed that data providers expose their data, in most cases, using OGC services only. Identifying standards that are well balanced between maturity and also simplicity of access it is proposed that data providers share

- their metadata using OGC CSW,
- their features using WMS, and application schema compliant WFS 2,
- their observations using SensorThings API part 1,
- their spatial coverage data using WMS and, if possible WCS,
- and assign URIs that resolve to both metadata, features and observations and consuming URIs of the codelists exposed by the Information platform registry tool.

The central system will take care of

- hosting the central URI resolver. It is proposed to build on the URI space already used by EPOS GIM and by Minerals4EU codelists (https://data.geoscience.earth/)
- hosting the codelist registry tool. It is proposed to build on the URI space already used by EPOS GIM and by Minerals4EU codelists (https://data.geoscience.earth/ncl/)
- hosting the shared data specifications. It is proposed to build on the URI space already used by EPOS GIM (https://data.geoscience.earth/def/)
- hosting the Information Platform metadata catalogue (MICKA)
- proposing a data publication alternative for the data providers that don't have the IT capacity/know-how
- proposing validation services to enable help data providers expose their information according to the jointly identified specifications
- deploying a harvesting system where needed/preferable depending on the architecture,
- managing data front-end,
- deploying a central spatial database where needed/preferable depending on the architecture chosen.

It will also take care of exposing the content available within EGDI according to the same component as the one required from data providers and also according to new data exchange practices, potentially

- metadata content using JSON-LD and DCAT_AP,
- features using JSON-LD,
- features using WFS3 (in combination with JSON-LD if feasible),
- spatial coverage data using WMS and, if possible WCS,
- observations using SensorThings API in combination with JSON-LD if feasible,
- the EGDI content in a SPARQL endpoint.

6 HOW TO PROVIDE DATA TO THE INFORMATION PLATFORM?

The different types of data a provider can send to the Information Platform can be classified from two main criteria: Spatiality and structuration.

	Structured	**Semi-structured**	**Non-structured**
Spatial	PostgreSQL DB with PostGIS extension Shapefile or GeoPackage	Excel with X/Y	Document referring to a geographical location
Non-spatial	Standard DB Excel following a data model	Excel CSV	Document (a document can be anything)

Table 3 - Example of data repository, depending of their type

Depending of their types, data will not be exposed by the data provider to the Information Platform using the same services.

6.1 Situation 1 - data provider has the IT capacity

In this situation, the data provider has enough IT capacity and skills to be able to directly serve the data with services, and in the best case, the data will be harmonized and will follow a data model defined according to GIP-P specifications.

This data provider must provide services to the Information Platform. For the delivery of simple map/rasters, the data provider must provide WMS that will be directly displayed in the WebGIS of EGDI. This basic solution must be limited to the data that cannot be provided with another solution as this only allow visualization and no other processing of the data.
Vector data must be provided using WFS. These WFS must follow an application-schema (depending of the thematic of the data, the data model to follow will be proposed by WP3 of the GIP-P about Standards and Interoperability issues). The data must use code-lists according to the thematic (also proposed by WP3).
Grid data must be provided using WCS.

The data provider has its own metadata catalogue (national metadata catalogue in the case of Geological survey, project metadata catalogue for project consortium). When selected metadata within are denominated by a keyword, they are harvested to the MICKA metadata catalogue. Harvesting rules may differ between data providers or catalogues. This aspect is further defined in D 5.2.

Unstructured data (documents, references, or other type of knowledge) must stay at the provider location and be accessible through HTTP. The data provider will have to create a metadata record in the MICKA and link to the document.

6.2 Situation 2 - data provider does not have the IT capacity

In this situation, the data provider does not have the IT capacity or skills to be able to service the data through services. The only way for him to serve data is through files (Excel, Shape, GeoTiff…) to a system provided by GIP-P. This aspect is further described in D 5.2 under 'Data publication alternative' section.

The minimum level the data provider must try to reach is using code-lists for the thematic of the data he wants to provide. For that, he might need to make a mapping from his own vocabularies to the agreed vocabularies used by the thematic community. He has also to have his data format as close as possible from the data models in used by the thematic community. Both code-lists and data models will be proposed by the WP3 of the Information Platform about Standards and Interoperability issues.

The data provider might not have its own catalogue for spatial data. Metadata can be inserted and edited directly on the catalogue for spatial data of the Information Platform.

Unstructured data (documents, references, or other type of knowledge) might be uploaded to the Information Platform or linked to another perennial platform and the metadata record related to these data can be inserted and edited directly in the dedicated tool of the Information Platform.

7 CONCLUSION

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

The overall blueprint described here frames the DNA of the GeoERA Information Platform Project (GIP-P). It will be updated iteratively based on GSPs' user requirement and GIP-P WPs feedback. Every other deliverable of the project shall comply with the blueprint.

In the project, the central node of the platform plays an important role. Its main components have been briefly described to clarify their position in the overall system.

In order to provide a clearer description of it, a specific deliverable D5.2 'GeoERA Central System specification' has been written. It complements the current blueprint.

Finally it is important to be aware of the long-term accessibility to the results of the GeoERA geoscientific projects (GSPs). Even though many of them might be able to set up services during the projects' 3 year duration they may not be able to maintain these services after the end of GeoERA. On the other hand EGDI has been established to maintain results from previous and on-going projects (like GeoERA) and EuroGeoSurveys is the organization behind this. So for many GeoERA GSPs a long-term solution for sustaining the results will be to deliver the data (spatial and non-spatial as well as structured and non-structured) to be stored centrally at EGDI.)

ANNEXES

Annex A - 5-star scheme

The 5-star scheme or 5-star Open Data is a rating system developed by Tim Berners-Lee, the inventor of the Web and Linked Data initiator. This rating system has been established to encourage data owners (especially government data owners) to follow the virtuous road to the good linked open data.
In GeoERA, following this scheme and having at least a minimum number of stars will be the minimum to fulfil the requirements of the Information Platform.

Before applying the 5-star scheme, we need first to remember the four principles of Linked data (also outlined by Tim Berners-Lee):

1. Use URIs to name (identify) things.
2. Use HTTP URIs so that these things can be looked up (interpreted, ”dereferenced”).
3. Provide useful information about what a name identifies when it’s looked up, using open standards such as RDF, SPARQL, etc.
4. Refer to other things using HTTP URI-based names when publishing data on the Web.

In the context of GeoERA, the principles apply, but must be adapted to the specificities of the data manipulated.

For example, for the third principle, the standards used will be of two types. At a technical level, for spatial data, it is recommended to use of Open Geospatial Consortium standards like WMS, WFS, WCS or SOS. At a semantic level, depending on the thematic, it is more than recommended to use the data specification/data models developed at international or European level for this thematic (e.g. EarthResourceML for mineral resources, GroundWaterML for ground water...).

The rating of the 5-stars scheme begins at one star and data gets stars when proprietary formats are removed and links are added.

★ *- Make your data available on the web (whatever format)*
but with an open license, to be Open Data

The first step of the Linked Data road is to follow the Open Data initiative and to open the data by making it available on the web (preferably using HTTP) in any format (it can be Excel format, GIS files or any files format that can be read).

From a consumer point of view, the benefits of this first star are immediate, he can look at the data, search it, store it (locally), use it in another system, change the data, and share the data with other. On the other hand, it can be difficult for the consumer to read the data which can be locked-up in a document, and he may need to write a custom data scraper to get the data out of the document.

For the publisher, this first star is easy to reach. The data is simple to publish (just take it and make it available somewhere). And a direct benefit is that the publisher will not have any more to explain to consumer that they can use his data and where to find it.

★★ - Make your data available as machine-readable structured data (e.g. excel instead of image scan of a table)

The second step to earn a new star is to process the data to make it directly usable by a machine. With the example of a table, the image scan of the table in a report allows a human to use the data (someone would say that machine with visual recognition can also do the job), but a table directly available in form of structured data will be directly processable by a machine.

For the consumer, he will be able to directly process it with (proprietary or not) software. He will be able to perform calculation on it, visualize it, and aggregate it or any other operations. The consumer will also be able to convert this data to any other structured format. The problem is still that it needs a proprietary software to get the data out of the document.

For the publisher, it is still simple to publish this data.

★★★ - Use nonproprietary format (e.g. CSV instead of Excel xls)

To earn the third star, the data must not be in a proprietary format.

For the consumer, he can now directly manipulate the data in any way he likes. The consumer doesn't need any more a proprietary software or proprietary libraries.

The publisher may have to convert the data from the original (and maybe proprietary) format to an open format. Nevertheless, it is still simple to publish this data (it is still a downloadable file).

★★★★ - Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

---

The fourth star is maybe harder to obtain. This is the star that really moves the data to the Linked data world. The data must use open standards from W3C, such as RDF and SPARQL, to identify things and use Uniform Resource Identifier (URI) to identify the data.

For the consumer, the benefits are multiple. He can link the data from any place, he can bookmark it to retrieve it, and he can reuse parts of the data (as these parts are also identified by URI). As some parts of the data may have an already known structure, the consumer can reuse existing tools or libraries to read these part. Finally, he can combine data with other data. However, the structure of the data can be more complicated to understand than tabular (CSV) or tree (XML, JSON) data.

For the publisher, as the data is now cut in small part, it is easier to have fine granular control over it and to optimise their access. Other data publishers can link into his data. On the other hand, the publisher has to invest some time slicing and dicing his data. He will need to assign URIs to data items and think about how to represent it.

★★★★★ - Link your data to other people's data to provide context

---

Earning the last star might be automatic, as the publisher will reuse some existing data from other people to complete his own data, he will create links between the data.

For the consumer, the links between the data will help to discover new things, and to enrich the knowledge about the domain. Unfortunately, some links with other data can be broken (as some broken links in standard HTTP that lead to 404 error pages).

For the publisher, his data will be discoverable, and as it will be linked from outside, it will gain in visibility. Also his data will gain in value with the link to other people's data. The publisher will have a new task, to enrich his data with links to other data on the Web, and he will have to repair broken or incorrect links from time to time.

Annex B - W3C data and spatial data on the web best practices summary

This annex only aims at providing an overview of the best practices from the two W3C working groups
Details description can be found here:

- Data on the web best practice: https://www.w3.org/TR/dwbp/#bp-summary
- Spatial data on the web best practice: https://www.w3.org/TR/sdw-bp/#bp-summary

Best Practice 1: Provide metadata

Best Practice 2: Provide descriptive metadata

Best Practice 3: Provide structural metadata

Best Practice 4: Provide data license information

Best Practice 5: Provide data provenance information

Best Practice 6: Provide data quality information

Best Practice 7: Provide a version indicator

Best Practice 8: Provide version history

Best Practice 9: Use persistent URIs as identifiers of datasets

Best Practice 10: Use persistent URIs as identifiers within datasets

Best Practice 11: Assign URIs to dataset versions and series

Best Practice 12: Use machine-readable standardized data formats

Best Practice 13: Use locale-neutral data representations

Best Practice 14: Provide data in multiple formats

Best Practice 15: Reuse vocabularies, preferably standardized ones

Best Practice 16: Choose the right formalization level

Best Practice 17: Provide bulk download

Best Practice 18: Provide Subsets for Large Datasets

Best Practice 19: Use content negotiation for serving data available in multiple formats

Best Practice 20: Provide real-time access

Best Practice 21: Provide data up to date

Best Practice 22: Provide an explanation for data that is not available

Best Practice 23: Make data available through an API

Best Practice 24: Use Web Standards as the foundation of APIs

Best Practice 25: Provide complete documentation for your API

Best Practice 26: Avoid Breaking Changes to Your AP

Best Practice 27: Preserve identifiers

Best Practice 28: Assess dataset coverage

Best Practice 29: Gather feedback from data consumers

Best Practice 30: Make feedback available

Best Practice 31: Enrich data by generating new data

Best Practice 32: Provide Complementary Presentations

Best Practice 33: Provide Feedback to the Original Publisher

Best Practice 34: Follow Licensing Terms

Best Practice 35: Cite the Original Publication

Data on the web best practice

Best Practice 1: Use globally unique persistent HTTP URIs for Spatial Things

Best Practice 2: Make your spatial data indexable by search engines

Best Practice 3: Link resources together to create the Web of data

Best Practice 4: Use spatial data encodings that match your target audience

Best Practice 5: Provide geometries on the Web in a usable way

Best Practice 6: Provide geometries at the right level of accuracy, precision, and size

Best Practice 7: Choose coordinate reference systems to suit your user's applications

Best Practice 8: State how coordinate values are encoded

Best Practice 9: Describe relative positioning

Best Practice 10: Use appropriate relation types to link Spatial Things

Best Practice 11: Provide information on the changing nature of spatial things

Best Practice 12: Expose spatial data through 'convenience APIs'

Best Practice 13: Include spatial metadata in dataset metadata

Best Practice 14: Describe the positional accuracy of spatial data

Spatial data on the web best practice

Annex C - Former projects analysis

Former projects for Geo-energy

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

Former projects for Groundwater

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

Former projects for Raw Material

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.

The ProMine project was one of the first to will having a European coverage for Raw Materials. This project has opened the road to harmonized data feed in a central database covering Europe. The ProMine data was also served via the ProMine portal as Inspire compliant web services.

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.

Made as the same period than ProMine, EuroGeoSource project was the first project for Raw Materials to implement a harvesting system to retrieve the data of national provider to feed a central database that was able to answer user queries. The system has open the road of harmonized services provided by national GSO harvested to a central place.

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.

In the Minerals4EU project, the choice has been made to strictly follow INSPIRE recommendations: the data stays at the provider level, and each partner (provider) must furnish a service to access the data. For that reason, a toolstack has been composed with all the tools required to provide the data and some cookbooks have been created to explain the users how to set up a diffusion system on their own data.

For the furniture of data, INSPIRE has been followed. Datasets are described in a metadata catalogue (MICKA – OGC CS/W catalogue managed by CGS), both the datasets supplied by the different providers and the datasets that can be useful for the thematic of the project. The data of each provider are available through an INSPIRE download service (OGC WFS) using the Raw Material data specification implementation (INSPIRE MR / EarthResource ML v2) slightly enhanced for the needs of the project. Then the data are exposed through the web and can be harvested by the Central System of Minerals4EU (which plays almost the role of a caching system).

The toolstack is composed of the following elements:

- The service implementation cookbook, that helps the providers to install the toolstack;
- Some open source tools (advanced text editor, desktop GIS, version control client);
- The database software with its spatial extension (PostgreSQL with PostGIS);
- The application server (Tomcat);
- The Web Feature Server (Deegree 3);
- An ETL (Extract, Transform, Load) software used for transferring the data from the provider database to the diffusion database (GeoKettle);
- And some other tools and examples to help the providers to understand the whole stack.

FP7 EURARE

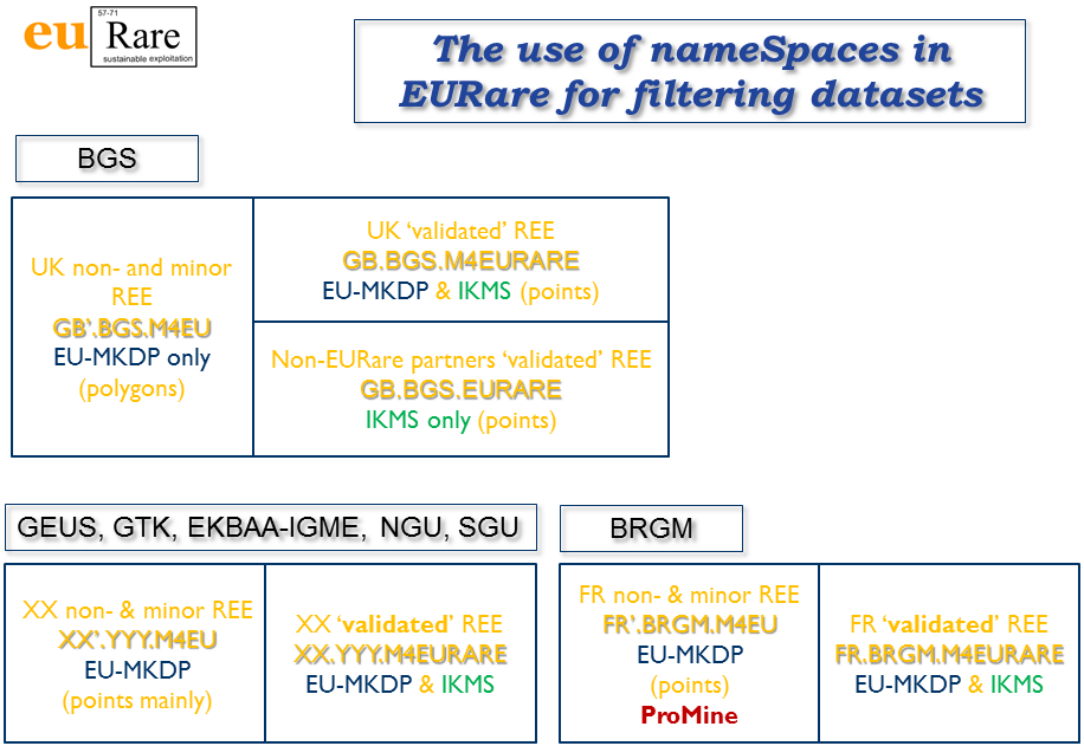
The main goal of the EURARE project was to set the basis for the development of a European Rare Earth Element (REE) industry. Establishment of an REE value chain in Europe would safeguard the uninterrupted supply of REE raw materials and products crucial for sectors of the EU economy (including automotive,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

The EURARE project shares the same system than Minerals4EU project, the two differences are that there is no Minerals Yearbook and the partners of the project are different and do not cover the whole Europe. Due to this latest specificity, the only notable point compared to Minerals4EU is that the data for the non-partner countries are served only by a single partner (BGS). As the two systems use the same database, the data coming from BGS and served for the countries which are not EURare Partners are differentiated with the use of a specific namespace.

In the datasets served by the partners, the projects they refer to is indicated in the namespace. Three cases are distinguished, data for both Minerals4EU and EURare project (i.e. data about REE served by partners of both Minerals4EU and EURare project), data for Minerals4EU project only (i.e. data on all commodities - incl. REE for some partners - served by Minerals4EU partners)

and data for EURare project only (i.e. data about REE served by BGS on behalf of countries non partners of the project).



Annex figure 1 - The use of namespaces for filtering datasets between EURare (IKMS) and Minerals4EU (EU-MKDP)

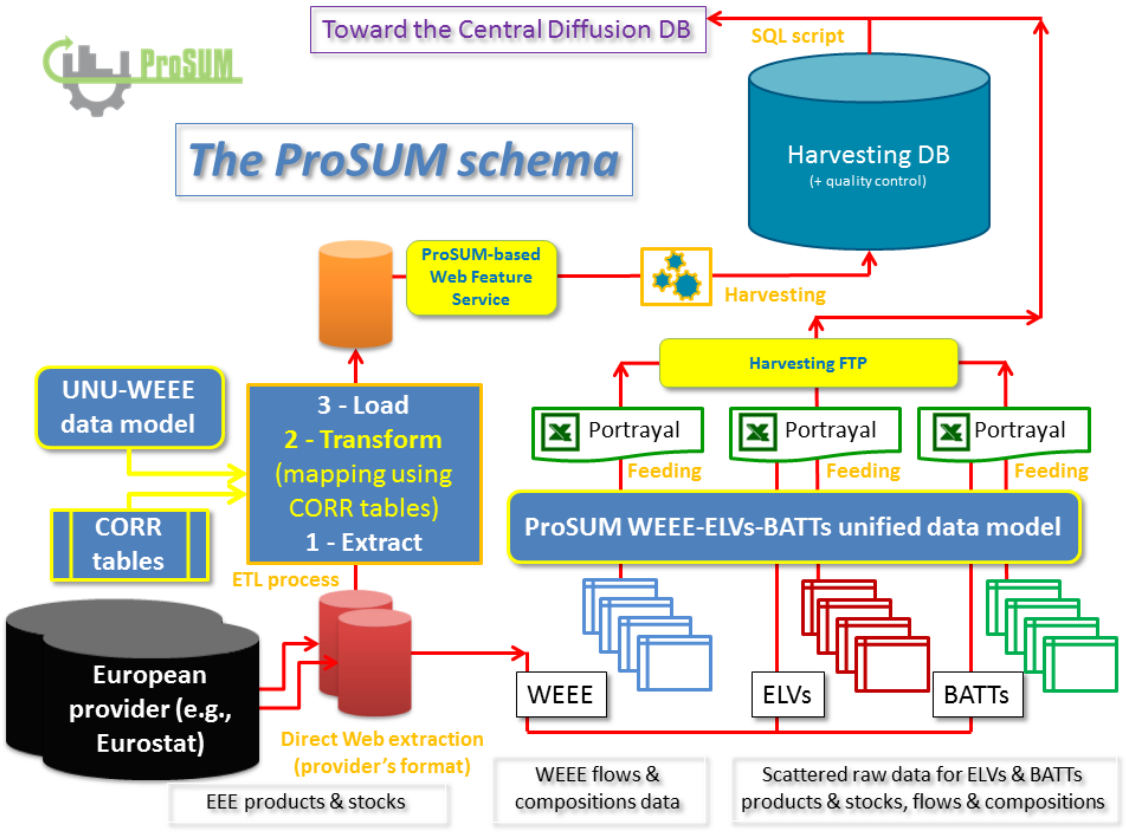
H2020 ProSUM

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.

The ProSUM project reuses the toolstack developed in Minerals4EU and extends the data model used (Inspire MR) to add the Mining Waste. The toolstack has been updated accordingly to these changes during the project.

The other (and main) part of the project is about the Urban Mine relative to End of Life Vehicles (ELV), Waste Electrical and Electronic Equipment (WEEE) and spent batteries (BATT). Before the ProSUM project, no data model was existing to represent the information on arisings, stocks, flows and treatment of these wastes. During the project, a *unified* data model has been developed that can handle the data for the three waste groups.

Compared to the Geological Survey representing the Mining Waste data providers, the provider of the Urban Mine data are not user of interoperable system to share data (they are not covered by the INSPIRE directly and don't have any obligation to provide their data). Furthermore, their data are more statistical than geographical. For these reasons, the ProSUM consortium has decided to create some Excel templates that could be used by the data provider in an easy way. These Excel templates can afterwards be integrated in a consolidated database where more computation can be done. An easy solution has been used during the project to transfer the data from the provider to the integrator using File sharing application, but a more long term solution has been studied with the use of an FTP for the provider and an automated solution to integrate the data in the database.



Annex figure 2 - The ProSUM data provider Schema

H2020 SCRREEN

The goal of the SCRREEN project is to establish an EU Expert Network that covers the whole value chain for present and future critical raw materials; to analyse 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 organise all of the data generated in other projects, associations, initiatives etc, and develop a knowledge data portal.

The system used in SCRREEN project is based on the one for Minerals4EU, like for EURARE project, the focus is specific (critical raw materials), and the data provider in the project do not cover the whole Europe. No other specificities can be raised at the data and services level.

EMODnet

The purpose of the project is to compile the harmonized offshore geological data including sea-floor geology, seabed substrates, rates of coastline migration, geological events and probabilities and mineral resources of the European seas and display them on a single internet portal. The system is created in such a way that it is possible to access the catalogs of data held by each project partner and the possibility of access to more detailed data. This will enable long-term sustainability of project results, as individual partners will maintain their own data. The project involves 34 partners.

Marine research is truly multidisciplinary as evidenced by e.g. the EMODnet I, II, and III projects that have been running since 2009. EMODnet Geology has succeeded in bringing together harmonised offshore data including sea-floor geology, seabed substrates, rates of coastline migration, geological events and probabilities and mineral resources.

Now in its third phase, EMODnet Geology Portal consolidates the existing data products with higher resolution and more contents. New services are being built, so users can investigate and search for borehole data, seismic survey data, and multibeam survey data using interactive maps and tools.

Other initiatives

European Plate Observing System (EPOS)

EPOS is a long-term plan to facilitate integrated use of data, data products, and facilities from distributed research infrastructures for solid Earth science in Europe. It will provide integrated access to solid Earth data to enable the Earth Science community to access data and products generated by different communities with different data formats and processing procedures in order to engage in cross-disciplinary investigations to advance the overall understanding of complex multi-scale geo-scientific questions.

It is a research infrastructure part of the EU ESFRI roadmap. The EPOS ERIC (European Research Infrastructure Consortium) was officially signed in November 2018 along with the engagement of many EU Ministries of Research to contribute to the ERIC and other already on the tracks to joining. Which implies that EPOS is not a one-off exercise that will finish at the end of the current H2020 EPOS-IP (https://www.epos-ip.org/) but is, indeed, a sustainable infrastructure

Annex figure 3 - EPOS in a nutshell

One of the scientific themes (Thematic Core Service (TCS)) that EPOS deals with is Geological Information and Modelling (GIM) and services will be central to providing such data to EPOS.

The objectives of the TCS GIM are:

- To design and implement an efficient and sustainable access to geological multi-scale data assets. This is done through the integration of distributed infrastructure components (nodes) governed by the EPOS Geology domain (geological surveys and research organizations communities).
- To provide a shared infrastructure to secure availability of services.
- To promote and implement standards for geological information and 3D models (INSPIRE, IUGS/CGI, OGC, W3C, ISO).
- To ensure integration with EPOS central hub (EPOS ICS-C)

In the first instance, these objectives will primarily be met by relying on existing work carried out by several EU projects (mainly EGDI, Minerals4EU and EPOS-IP itself) as depicted in one of images below.
They will deliver geological multi-scale data (e.g. borehole data, sample and analysis data, geophysical data), geological maps, subsurface (e.g. temperature, aquifers) and geo-hazard (e.g. landslides, surface faulting) data; borehole visualization, including visualization of logs, sampling/coring intervals, analyses; Geological 3D-4D models, including structural geology models to EPOS ICS-C

In addition, member organizations of the EGDI consortium (British Geological Survey, Bureau de Recherches Géologiques et Minières and the Geological Survey of Denmark and Greenland)

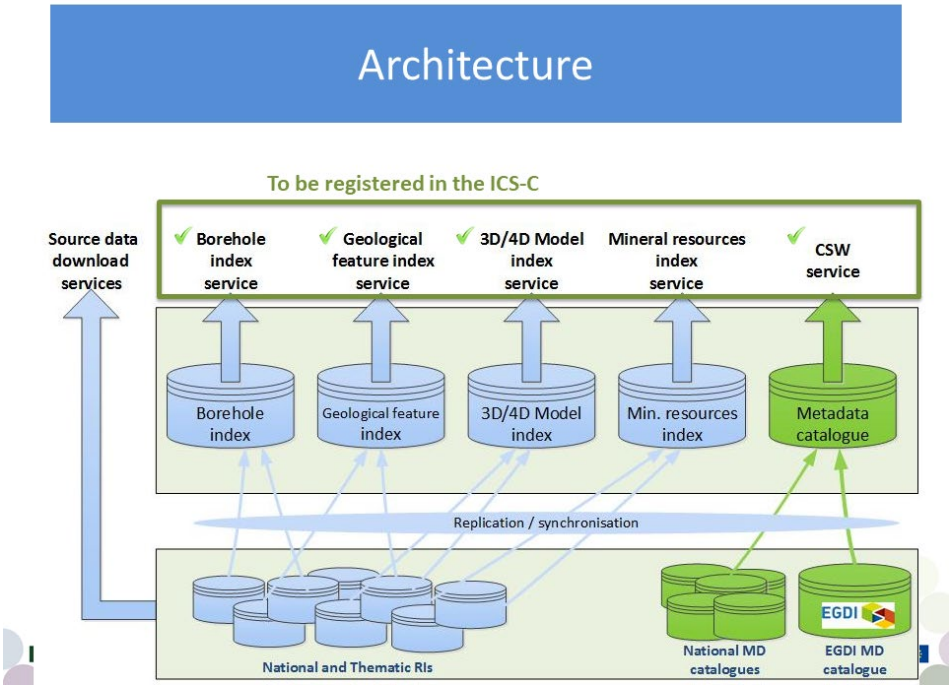
responsible for operating the EGDI will also be responsible for hosting the EPOS ICS-C once it becomes operational at the end of 2019.
These three organizations are in a position to ensure harmonization and integration between these initiatives in terms of both their architecture and technological approach so as to ensure consistency and efficiency in the delivery of integrated European geological data services to all stakeholders.

From a data and service point of view it has been decided to:

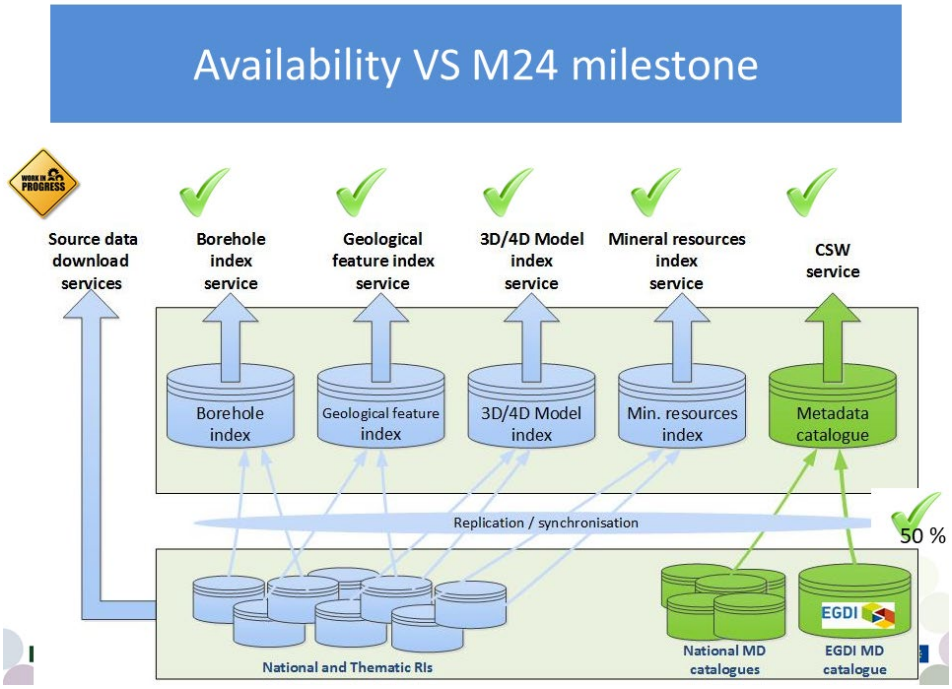
- only expose data index to EPOS ICS-C,
- define the data structure of those indices extending whenever possible pre-existing standards (ex : GeoSciML Lite) for the specific situation of 3D/4D models a ModelView was specified mimicking GeoSciML Lite and EarthResourceML approach,
- implement them using as much as possible Linked Data principles,
- collect them / harvest them from data providers to EPOS TCS GIM central node using application-schema compliant WFS 2.0
- have the data index entries point using URIs to more complex flows that will reside at data providers level

The only exception to this is the metadata catalogue.
As EPOS ICS-C asks each TCS to expose their datasets and services metadata in a specific flavor of DCAT-AP (EPOS_DCAT_AP) which specification are not yet finalized it was decided to set up a dedicated TCS GIM metadata catalogue (along with its CSW endpoint) which harvested partially EGDI catalogue using CSW.

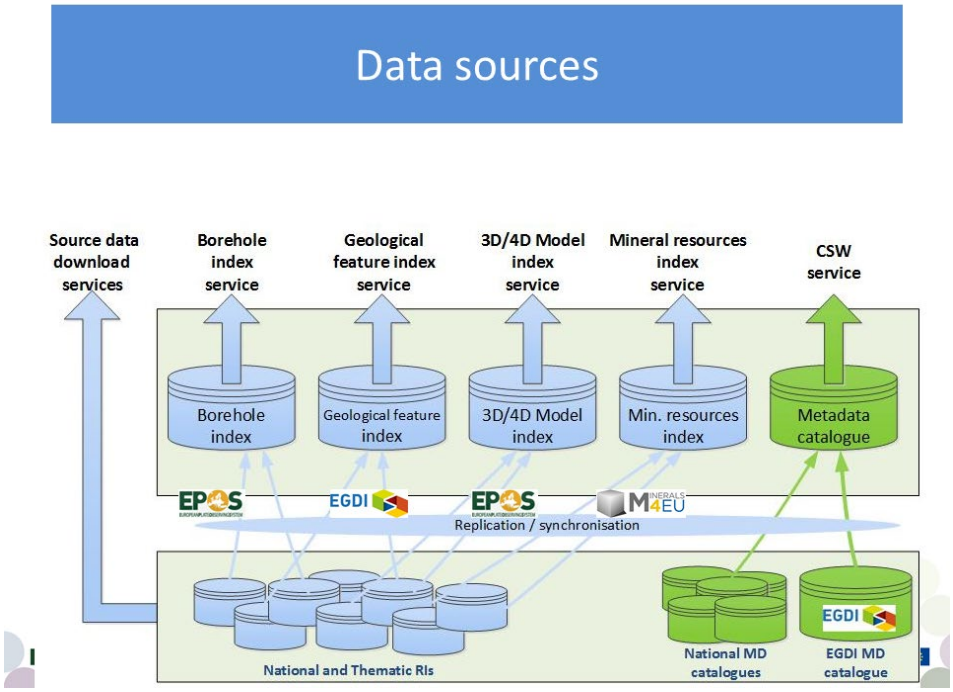
The images below are extracted from an EPOS consortium meeting in March 2018 and an INSPIRE conference presentation in 2018. They provide a good overview of EPOS TCS GIM architecture.



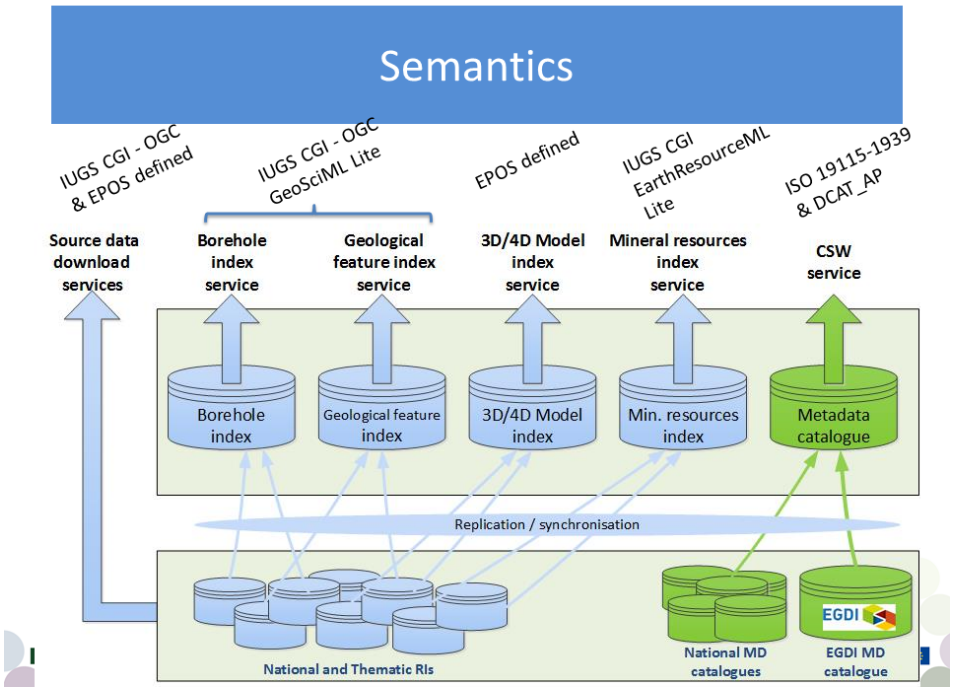
Annex figure 4 - EPOS architecture



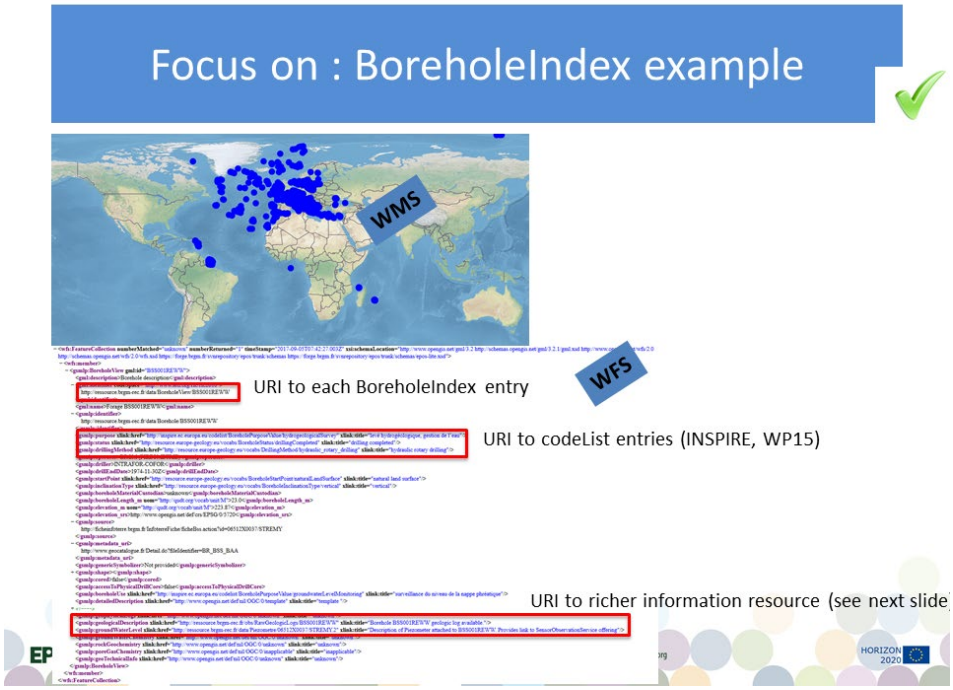
Annex figure 5 - EPOS Availability VS M24 milestone



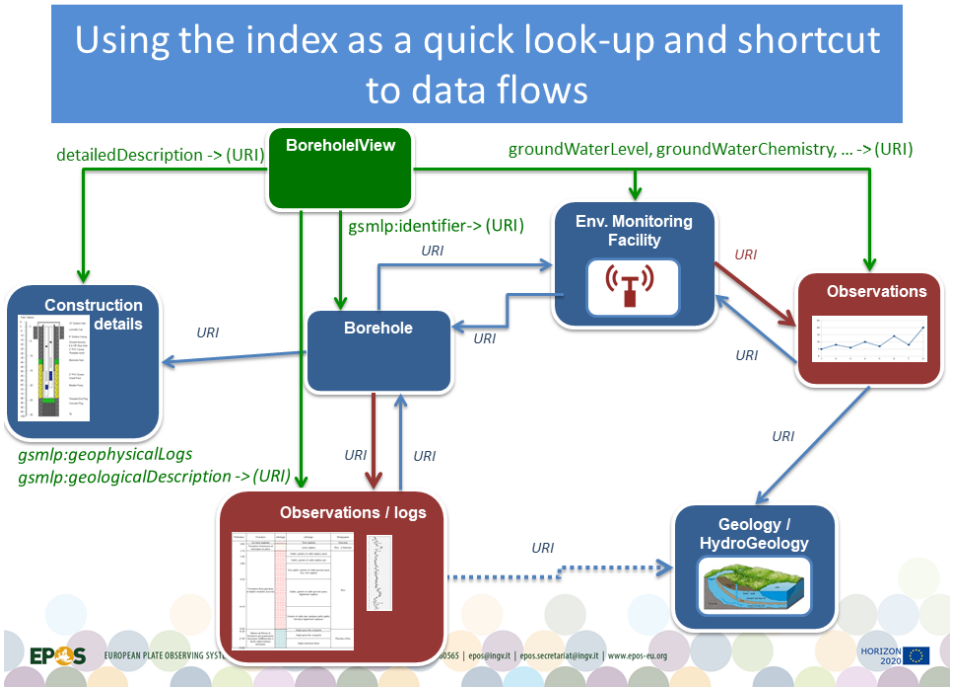
Annex figure 6 - EPOS data sources



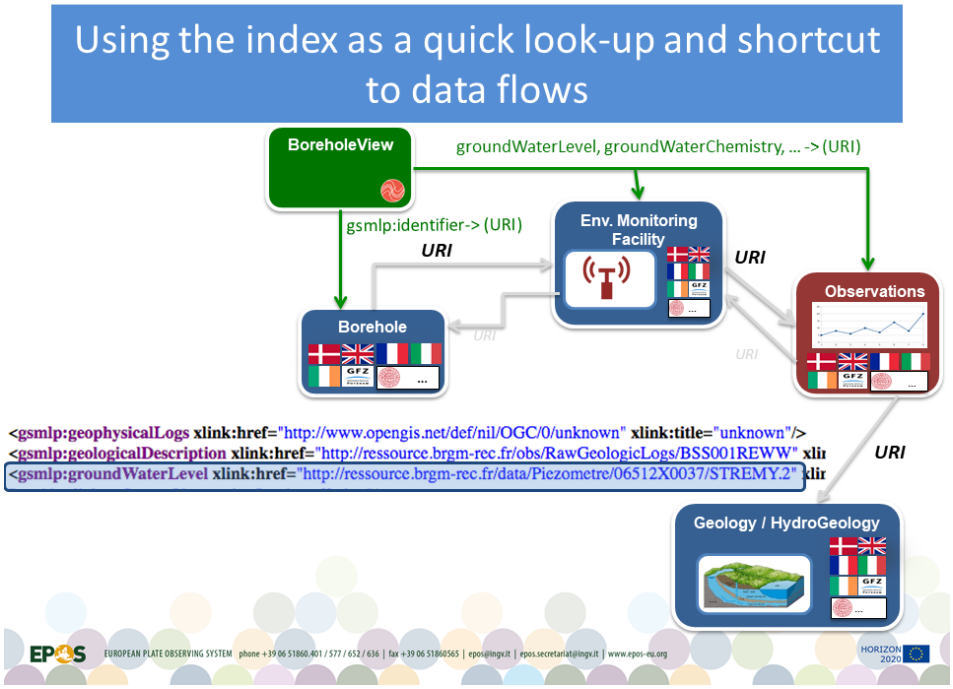
Annex figure 7 - EPOS Semantics



Annex figure 8 - EPOS BoreholeIndex example



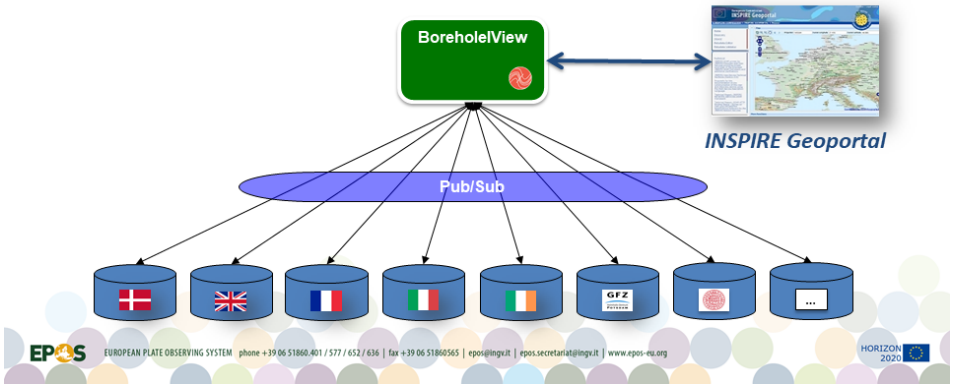
Annex figure 9 - EPOS Using the index



Annex figure 10 - EPOS Using the index (2)

How do we feed the index ?

- Each institution maintains its Borehole summary info service
- Only those summary info are harvested at EU level
- The index is up-to-date thanks to a pub/sub approach



Annex figure 11 - EPOS Feeding the index

ENVRIFAIR

Another initiative to which GIP should be closely aligned is ENVRIFAIR. The goal of ENVRIFAIR is to integrate EPOS and other participating European Research Infrastructures to build a set of FAIR (the same principles that underpin GIP) data services which enhance the efficiency and productivity of researchers, support innovation and enable data and knowledge-based decisions. ENVRIFAIR will integrate closely with the European Open Science Cloud. The European Open Science Cloud vision is “to give Europe a global lead in scientific data infrastructures and to ensure that European scientists reap the full benefits of data-driven science”. It is therefore important that GIP aligns closely.

The kick-off meeting for ENVRIFAIR will be held on 14-15 January 2019. More information about how GIP should aim to interact with ENVRIFAIR and EOSC will be included in the prototype version of this document due in 12 months time.