FP7-INFRASTRUCTURES-2012-1

Grant Agreement no. 312845

Scoping Study for a pan-European Geological Data Infrastructure

D 4.4

Report on recommendations for implementation of the EGDI

Deliverable number	*D4.4*
Dissemination level	*Restricted*
Delivery date	*19 June 2014*
Status	*Final*
Author	*Sylvain Grellet, BRGM*

EGDI-Scope Project, WP4

Report on recommendations for implementation of the EGDI

Table of Contents

1 Overview of WP4 – Technical design

Work package 4 of the EGDI-scope project sets out the requirements for technical design, deployment and maintenance of a possible European Geological Data Infrastructure (EDGI), in order to fulfil the user requirements and required data provision, identified in WP2 and WP3 (and proceeding parallel to WP4). Some of the most important requirements for an EGDI infrastructure will be based on the principles and directives defined within the INSPIRE framework and other large initiatives dealing with geospatial information, and it will build on the experience of the design, implementation and operations of the different portals and other geological information systems developed within previous and on-going projects and initiatives.

This document is the report on recommendations for implementation of the EGDI. It consolidates previous deliverables (D 4.2, D 4.3) recommendations into a coherent scheme in strong relation with WP1 D 1.3 “Implementation Plan EGDI”.
The value-added of the current document is this coherent scheme; detailed technical description of the components being already available in D 4.2 and D 4.3, the reader is advised to refer to the above mentioned deliverables in order to get more technical information.

2 Summary of the rationale

The EGDI architecture will follow the INSPIRE Directive and other environmental information systems principles and best practices. A service oriented architecture will allow data to remain as close as possible to the producer and be exchanged efficiently and effectively. This distributed system will rely on information supplied by national data providers; mainly EGS members surveys.

It will be organised using the following three layers architecture as in other major programmes (Inspire, GEOSS, etc …):

- Access layer: containing the data services produced by the geological surveys at the national or regional level,
- Mediation layer: containing the common components that are required to register, view, access and process data,
- Client layer: the "visible" component of the "architecture", and containing the EGDI portal, thematic portals, or smartphone apps for instance. It uses services delivered by the mediation layer or by the access layer.

Semantic interoperability will be achieved by:

- Documenting each dataset and service using metadata
- Using whenever possible INSPIRE defined data models and extending them when additional information is required to satisfy specific use cases
- Using common controlled vocabularies: INSPIRE defined ones are a starting point but they don't always cover all needs. A coordinated governance of the content will allow extensions to be properly developed.

Technical interoperability will be enabled using commonly defined and openly documented web services standards. At least the following services categories will be deployed: Discovery service, View service, Download service and Spatial data service.

On top of this information backbone, human access interfaces (portals) will then be set up more easily.

Thematic projects (like e.g. Minerals4EU) will most of the time deploy their own Thematic Portal tailored towards meeting the requirements critical for the relevant end user groups but utilising the underlying EGDI technical infrastructure.
One single portal - the EGDI Portal - will be a part of the EGDI central node and will through simplistic metadata discovery and view functionality provide easy access to all information in the central node as well as in the distributed part of the underlying infrastructure. This includes data generated by thematic projects as well as more generic baseline data. Its Catalogue, Registry and Viewer will enable the end user to easily identify the piece of information/project output that suits his needs.

Being deployed according to the main principles described above, EGDI, the EU information system for geological information, must also be connected to other initiatives (GEOSS, EPOS, EU open data portal…) and domains (Marine, Water, Risks…).

3 Prioritisation of the components to be deployed

Previous deliverables identified components that will part of the EGDI layered architecture.
The information system cannot be set up all at once, and hence the development of components has to be prioritised.
This should be seen in relation to the prioritisation of datasets (WP3) and coordinated with the development of thematic portals in on-going projects to ensure proper data exchange between EGDI central node and data generated by such projects.
Data Services deployed can also be prioritised: WMS and WFS being the minimum acceptable.
The architecture will then grow step by step, adding new components and domains as new thematic projects will be set up.

One cannot anticipate neither the technical/thematic decisions that will be taken in thematic projects in the near future nor the opportunities to interact with projects and communities that are not under the pure scope of EGDI. Thus, various possible scenarios can lead to different prioritisations of the components development.

However, two aspects remain certain:

- The crucial need to quickly have an EGDI central node connected to currently running EU thematic projects (first implementation phase of EGDI),
- Within each main aspect ("Semantic interoperability" – Chapter 5, "Technical interoperability" – Chapter 6, "Upgrade EGDI nodes functionalities" – Chapter 7) some element should be deployed before others. Thus relative priorities can be defined.

As a consequence this document proposes in the four chapters below:

- A prioritisation to have a first version of the EGDI central node up and running and also connected to current EU projects (first implementation phase of EGDI),
- Then a prioritisation proposal within each main aspect. Each aspect has a specific chapter in which prioritisation proposals appear using a combination of bullet numbers and, if needed, letters (ex: 1.a).
  *For example: in Chapter 5 – "*Semantic interoperability*", metadata have first to be Inspire compliant (bullet number 1). Once this is done, compatibility with the CKAN registry could be discussed (bullet number 2)*

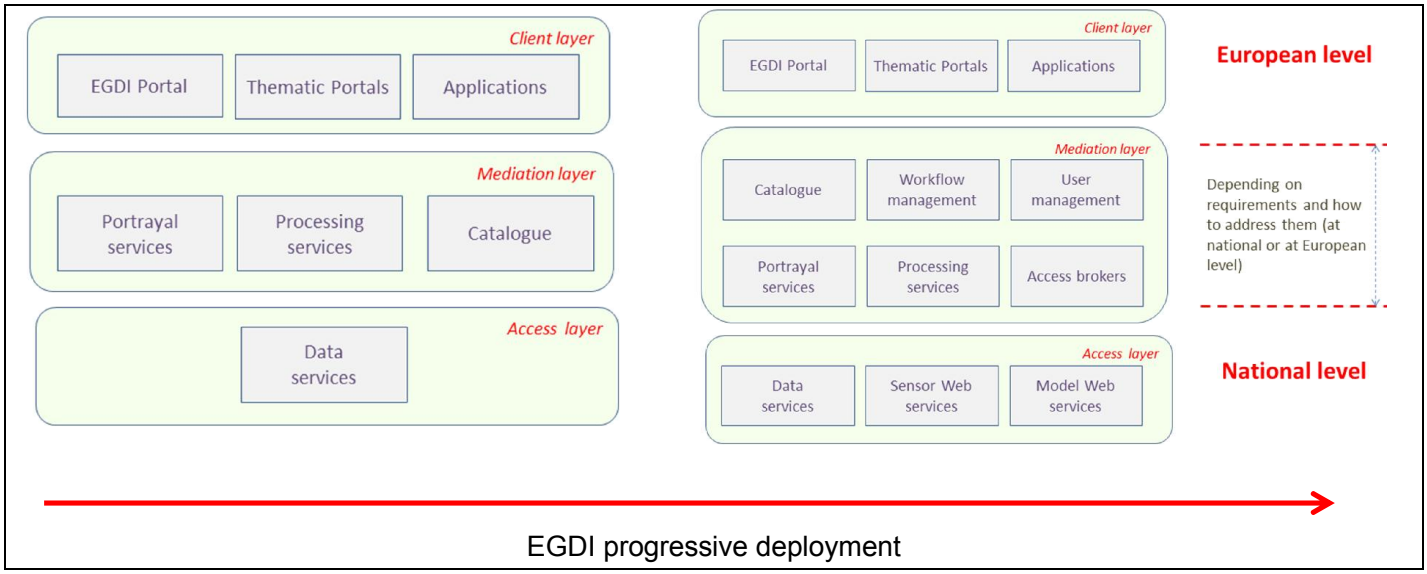

Figure 1 - From EGDI nutshell to full scale components

4 First implementation phase of EGDI

4.1 Central node

The information system currently being defined needs to gain visibility to ensure optimal support from NGSOs and generate momentum. To achieve this, a first version of the EGDI portal has to be set up already early in the first implementation phase.

This initial central node nutshell will build on reusing experience and, as much as possible, pre-existing bricks identified from various reference projects (ex: OneGeology-Europe, EuroGeoSource, Minerals4EU, ...).

A roadmap for the implementation of existing European datasets was proposed in D3.3. On a case by case basis, it will have to be determined which component/content is to be re-used from those projects.

For example:

- Harmonised maps available at OneGeology-Europe are an important value. So are to “reference” datasets/layers from other projects: Pan-Geo, ProMine, EMODNet,
- EuroGeoSource and Minerals4EU experience could quickly enhance the central node capacities with: a European repository (central database, …), Brokers (mediators) to harvest national data, a search engine on data but also on documents, etc ...

The same discussions will also have to be applied to search interface, the map viewer, the metadata viewer, …

4.2 Architecture / Link with other projects

Based on the above mentioned elements, the central node will de facto be deployed in a distributed system way, already applying the interoperability rules extensively described in the previous WP4 deliverables. Some exceptions to the distributed approach might occur due to the non-distributed nature of some datasets (e.g. the EMODnet substrate map). Furthermore, to sustain the results of some projects it may be required to implement central harvesting databases to increase performance, ensure reliability of services and enable advanced processing routines in the mediation layer.

In order for datasets stemming from thematic projects to be available through the EGDI portal, services to collect information will be deployed at the thematic portals level. Depending on the level of information needed, standards WMS/WFS already deployed to make data available on the thematic portals are good candidates for this.

Once retrieved, the EGDI metadata catalogue and services will allow the external world to discover, view and download information produced by thematic projects.

In a couple of actions, a first EGDI information system architecture will then be available.

5 Semantic interoperability

5.1 Metadata

Datasets and services metadata will be available for resources coming from thematic projects and European products.

1 First priority is to abide by INSPIRE Implementing Rules. This will allow as a positive side-effect interoperability with EPOS/CERIF metadata and DCAT-AP as both mappings are on their way.

2 The connection of EGDI metadata to the CKAN registry could be implemented at a second stage.

5.2 Data models

WP3 has already defined the priority for the datasets that must be taken into account:

- Geology,
- Mineral Resources,
- Water Resources,
- Geohazards: flooding, earthquakes, landslides, subsidence,
- Soil.

As far as data models generation methodology is concerned it is recommended to:

1 Target INSPIRE compatibility first
   a. This includes extensions of Inspire data models by thematic projects and for European products,
   b. But also distribution schemas (exchange scenarios like Portrayal Classes as in GeoSciML) for better reuse, for example in GIS tools. This aspect is often overlooked which reduces possible reuse of datasets by a broader user spectrum.

2 Structure data models development, maintenance and access.
   This concerns both data models (specifications) produced for Thematic projects and European products.
   Keeping in mind that those shall be:
   a. Centrally maintained to ensure the overall consistency of the system and to avoid parallel understandings that, in the long run, bias the entire information system (ex: having different understanding of the same concept definition). Clear roles have to be agreed on,
   b. Openly documented: to ensure a better understanding, thus reuse, via accessible documents available at the EGDI central node level,
   c. Easily accessible,
   d. Perennial: stored in an accessible versioning system such as subversion (with access control to editing parts) as Iso, OGC, CGI, Inspire data models are. Otherwise models and corresponding tools can become useless overtime.

3 Ensure the link to the Inspire Maintenance:
   Ideally via the Inspire thematic clusters currently being set up on Geology (GE), Soil (SO), Natural risk zones (NZ), Mineral resources (MR), Energy resources (ER).

4 Specify and deploy mechanisms (services) to validate xml retrieved from the lower levels in the architecture
   a. Those should be deployed at the EGDI central node and/or the thematic portal levels.
   b. Validation must also cover the dereferencing of the links from the xml files to the vocabularies. Schematron solutions will be the reference to do so.

5 Provide OGC SLD (Styled Layer Descriptor) and SE (Symbology Encoding) for Web Map Service.
   To ensure common portrayal rules

6 Ensure coherence with international standards via:
   a. The use of international standards (ex: GWML2, …). Those should be considered as additional exchange scenarios for the information system when Inspire specifications already exist on the same domain
   b. Linking with international standardisation organisations: OGC, CGI, …

5.3 Vocabulary

For the third of the semantic interoperability aspects, it is recommended to deploy the vocabulary parts as follows:

1 Deploy the EGDI registry
For codelists defined in thematic projects and Feature Catalogues related to data model extended by thematic projects
   a. Via a single consultation web page as a starting point,
   b. that have to be complemented applying the de-facto standard (SKOS/RDF) in a Linked Data Repository
2 Provide tools for the experts to define, comment vocabularies,
3 Coordinate with CGI when extending codelists,
4 Specify and deploy a vocabulary webservice.

6 Technical interoperability

6.1 Core services

As already defined in the other deliverables, three core services must be deployed.
They all share the same priority level:

1 Discovery Service (CSW), View service (WMS), Download service (WFS).

6.2 Enlarging the service panel

Various services will be added to the information system. Those are also described in details in deliverable D 4.3.
Depending on the thematic projects coming, a handful of different services could be deployed. Some can already be anticipated, others are listed at the end of this chapter more as a check list rather than a prioritisation proposal.

1 Quality services for data validation and conformity testing (QA/QC)
   a. Specification
   Those specifications should be openly documented.

   b. Deployment
   Extension of the validation service described in chapter 5 “Semantic interoperability”; those should be deployed at the EGDI central node and/or the thematic portal levels.
   Schematron solutions will be the reference to do so.

2 Access Brokers (Mediators)
   a. Specification
   To avoid deploying new ad-hoc solution for each new project.

Those specifications should be openly documented.

b. Deployment

3 ATOM feeds
For pre-defined dataset download as recommended in Inspire.

4 Support the OpenSearch standard
With geospatial and time extensions as it is proposed in the possible profile for the CS-W 3.0 part 4.

5 Identification layer
Inspire and other EU projects will, in the meantime, provide better recommendations on this.

The richer services check-list encompasses:

- WMS-T,
- Other download services that could be added to Inspire (currently under discussion) : Sensor Observation Service, Coverage, …,
- Gazetteer,
- As identified in D 4.2 (Services identified in European projects): WPS, WCPS, 3D web services, Model web services.

New service type that does have pre-existing specifications will at some point be developed in projects running under EGDI umbrella. For example, this could happen to web services serving 3D data.
Then it will be highly recommended not to restrict the activity to the development of the service only but also to have open service specifications made available. This is already advised for QA/QC services and Access Brokers (Mediators) above.
This will foster reuse of those new possibilities by the community.

7 Upgrade EGDI nodes functionalities

7.1 Portal

In addition to the recommendations for the implementation EGDI central node first version (see Chapter 4.1 “Central node”) the following components might be added:

1 Enhanced search possibilities
Deploying enhanced search engines (SolR, elastic search, …)

2 The possibility to process data
Including statistics, charts, graphs, …

3 User Management

4 Workflow management interface

5 Provide applications for smartphone or other

6 Deploy a repository for tools developed under EGDI umbrella
So that they can be reused from a project to another.

7.2 Architecture

The architecture might also be strengthened by:

1 Monitoring it
To ensure it is aligned with user usage and respects Inspire requirements on Quality of Service.

2 Investigating cloud based architecture benefits

8 Methodology to include new thematic projects

A methodology to include new thematic projects will be developed.
It will be applied by the EGDI Technical secretariat according to governance rules defined within EGDI.

This will enable the Technical secretariat to

1 Check what is available in EGDI and can be reused by the new projects,

2 Provide advice and coordination on
   a. Semantic interoperability: To ensure the overall coherence of the information system when it comes down to applying and extending Data models and Vocabularies. And to push those evolutions to the EGDI registry so that they are openly available to the broader community.
   b. Technical interoperability: To identify the architecture that best suits project needs amongst the possible architecture solutions identified in D 4.3 (Engineering Viewpoint). And to link the new project to the EGDI central node.
   c. Portal functionalities and reuse of pre-existing tools

3 Follow EGDI technical rules for improving EGDI Repository with results from the new project.
   For example
   a. Reference layers delivered as WMS/WFS and registered in the metadata catalogue,
   b. Deployment of new services

The use of the Reference Model for Open Distributed Processing applied in D 4.3 is recommended to avoid mixing all issues into one level, especially when addressing many thematic domains.

Below is an example of what could be done for Minerals4EU

Viewpoint	
Business	To provide EU with Mineral Resources information (here, is the purpose of the project)
Information	Metadata: to use INSPIRE metadata regulation
Service	Data model: to use INSPIRE Specification + extension (for Mining Wastes)
Engineering	Vocabularies: to use INSPIRE code-lists extended by CGI vocabularies
Technology	Information from documents (described by some metadata)

Table 1- RM-ODP example for Minerals4EU (summary)

9 Not only an infrastructure but an information system

Over the course of the project is has become more and more obvious that the need was not only for a Data Infrastructure but for an Information System. A real change of paradigm is needed from one shot project oriented activities to a perennial information system in which every project contributes to the overall endeavour.

The Information System content goes beyond a pure data infrastructure as it sets up:

- Domain groups: to define domain needs, enhance data structures, define needs for value added data, quality (QA, QC), …
- IT groups: to coordinate best practices, help domain groups to structure their information needs, define data workflow and data update rules, support deployment of new data collection solutions at Member States level, …
- An IT backbone around its central node: architecture, collection/dissemination databases, viewer, portals, web services, …

Moreover, there is a bigger interest in organising an Information System instead of a data infrastructure as:

- An information system uses data infrastructures, but is not limited to …,
- An infrastructure depicts mainly hardware issues in most minds whereas EGDI will create synergies between initiatives and projects,
- It provides more flexibility in the way it is deployed,
- It broadens the scope, by targeting also groups involving domain experts to define data models (e.g. Mineral waste extension to INSPIRE), quality (QA, QC), add value to data…

The Information System could support the European Commission, collecting information relating to European directives / policies and thus streamline the information flow from national to European level:

- Mining Waste directive,
- Raw Material Initiative,
- Groundwater directive (link with WISE), …

It will become de-facto the reference geoscience information pipeline towards EU Commission and provide link with other initiatives:

- INSPIRE: strongly implementing INSPIRE rules (extending them when necessary)
- SEIS: becoming SEIS Geoscience information pillar
- GEOSS: being GEOSS European counter for Geological data
- EPOS: strengthening the link with research communities

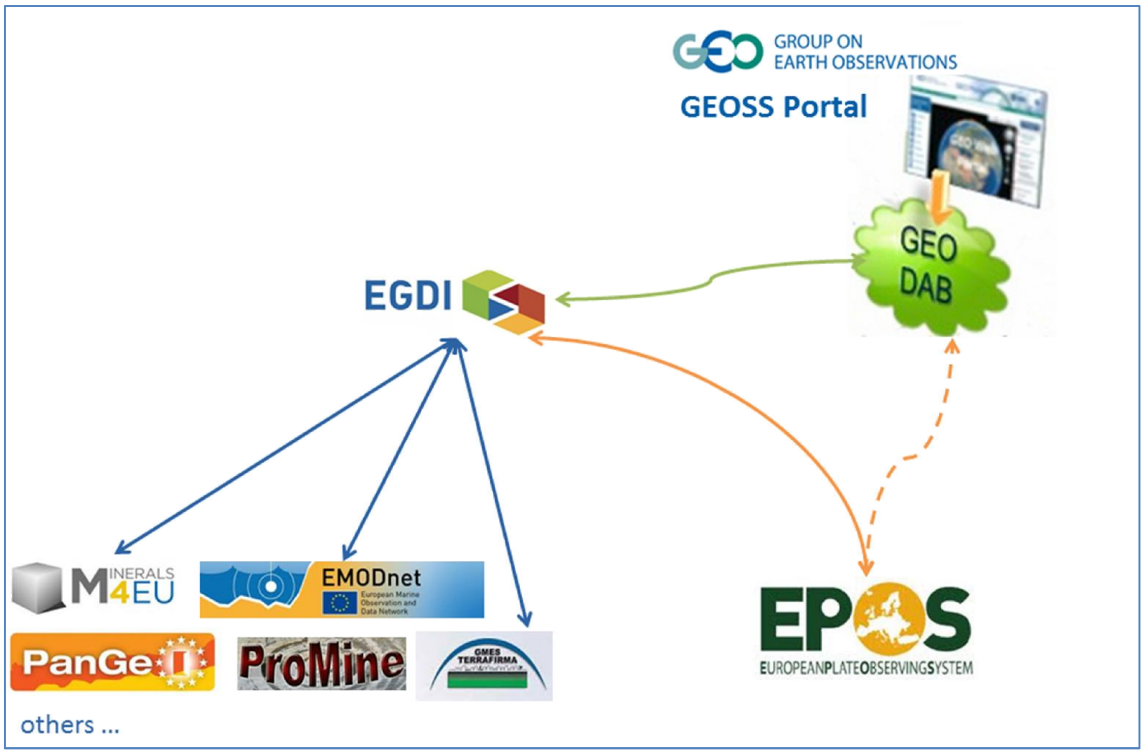

Figure 2 – Example of links with other initiatives (ex: GEOSS and EPOS)