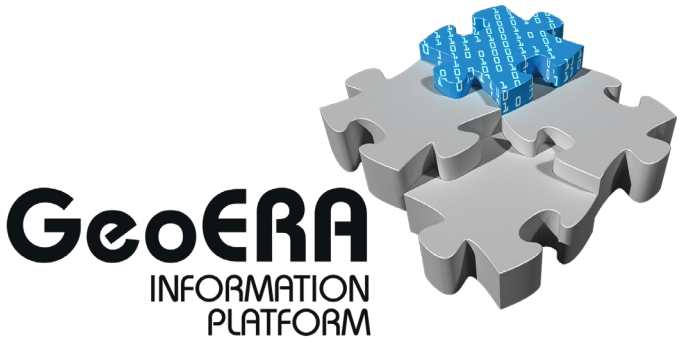


Deliverable 5.3

EGDI Platform - Architecture assessment and perspectives

Authors and affiliation:
Jean-Baptiste Roquencourt, BRGM
Sylvain Grellet, BRGM
Margarita Patricia Sanabria Pabón, IGME
László Sőrés, MBFSZ
Dana Čápová, CGS
James Passmore, BGS
Carlo Cipolloni, ISPRA
Andrej Vihtelič, GEO-ZS
Blaž Bahar, GEO-ZS
Martin Hansen, GEUS

E-mail of lead author:
jb.roquencourt@brgm.fr

Version: 22-10-2021

This report is part of a project that has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement number 731166.

Deliverable Data		
Deliverable number	D5.3	
Dissemination level	Public	
Deliverable name	EGDI Architecture assessment and perspectives	
Work package	WP5, Architecture	
Lead WP/Deliverable beneficiary	BRGM	
Deliverable status		
Submitted (Author(s))	22/10/2021	Jean-Baptiste Roquencourt
Verified (WP leader)	28/10/2021	Jean-Baptiste Roquencourt
Approved (Coordinator)	29/10/2021	Jørgen Tulstrup

GENERAL INTRODUCTION

The European Geological Data Infrastructure platform (EGDI) existed since 2016. It already provides functionalities appreciated by users.

Entering into GeoERA project, the GIP-P project was able to start from the existing EGDI platform. This helped us developing new functionalities without delay and answering the urgent requirements of the scientist in the GSPs. However, in science and particularly in Information Technology, the landscape is ever evolving. Usages of scientific data puts the emphasis on workflows, notebooks and AI. EGDI platform needs to be ready to address them, be it on the functional level as well as on the technical level.

This document reviews the state of the European Geological Data Infrastructure (EGDI) platform with regards to the recommendation made within GIP-P and propose a path with solution tested with heavy trafficsolutions to address the new data user requirements.

TABLE OF CONTENTS

Definitions

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

Data user: Advanced end users who are also using data services and as such use APIs, and data models

End user: User of the EGDI platform inclined to use WEB interface or traditional shape, csv files

FAIR: Findable Accessible Interoperable Reusable – principles that an IT platform should try to achieve for (meta)data dissemination (https://www.go-fair.org/fair-principles/)

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

GIP-P: GeoERA Information Platform Project.

GSP: GeoERA Scientific Project. The 14 scientific projects of the GeoERA programme.

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

Module: an application or software that is involved in serving product**.**

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

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

1.INTRODUCTION

This document aims to present the compliance between the vision of EGDI, its transcription in the envisioned WP5 architecture deliverables D5.1 and 5.2 and the actual EGDI architecture evolution during the GIP-P.

It has been written based on a long experience of designing architecture for distributed interoperable Information Systems, web GIS, web service and APIs up to 1,5 M hit a day.

Based on the software architecture principles which have reached a consensus amongst IT communities (DevOps, Craftmanship, etc), we review GIP-P outcome and what should be achieved after the GIP-P project.

Within this document, based on D5.1 and D5.2, we consolidate their vision and propose pragmatic solutions for the future of EGDI.

Therefore, we briefly introduce the context entering of EGDI into the GIP-P and the actual context of European project. Following is the review of the actual architecture delivered within GIP-P. As the achievement are described in other deliverables, we focus on the next step of the architecture by refining the target, and associated actions to finally propose an extended vision of the actual EGDI platform.

It was originally proposed in the GIP-project to write two updated versions of D5.1 and D5.2, but we find it more useful to combine the findings and recommendations into one document which is this one.

2.EGDI CONTEXT

2.1 EGDI Goals

Reference is made so several documents as the basis for design the most suitable architecture.

In EGDI as a component of EGS' strategy, the vision is given as the following:

*"EGDI is **the initiative of EuroGeoSurveys to facilitate sustainable access to geological data, information and knowledge** at the European level building on the NGSOs' national databases and roles thereby addressing societal challenges including future resource needs, risk mitigation, environmental protection, subsurface management, climate change and European welfare. This will be **achieved by sharing expertise, developing and implementing common standards, procedures and tools and by building, operating and maintaining a technical infrastructure and repositories**. This will be to the benefit of a wide range of users including decision makers, public authorities, researchers, industry, NGOs and the general public, either directly or through interaction with other infrastructures."*

From https://eng.geus.dk/media/13934/nr41_p95-98.pdf

"One of the main challenges for all these European initiatives, including EGDI, is to make them sustainable. […] Finally, it has turned out that it is difficult to convert national geological databases and make them interoperable a-cording to the requirements in the INSPIRE implementing rules. Many of these rules are very complicated, and many resources have been allocated to the database administrators at the national surveys in order to make their data compliant with the standards."

2.2 Multiple components and multiple teams

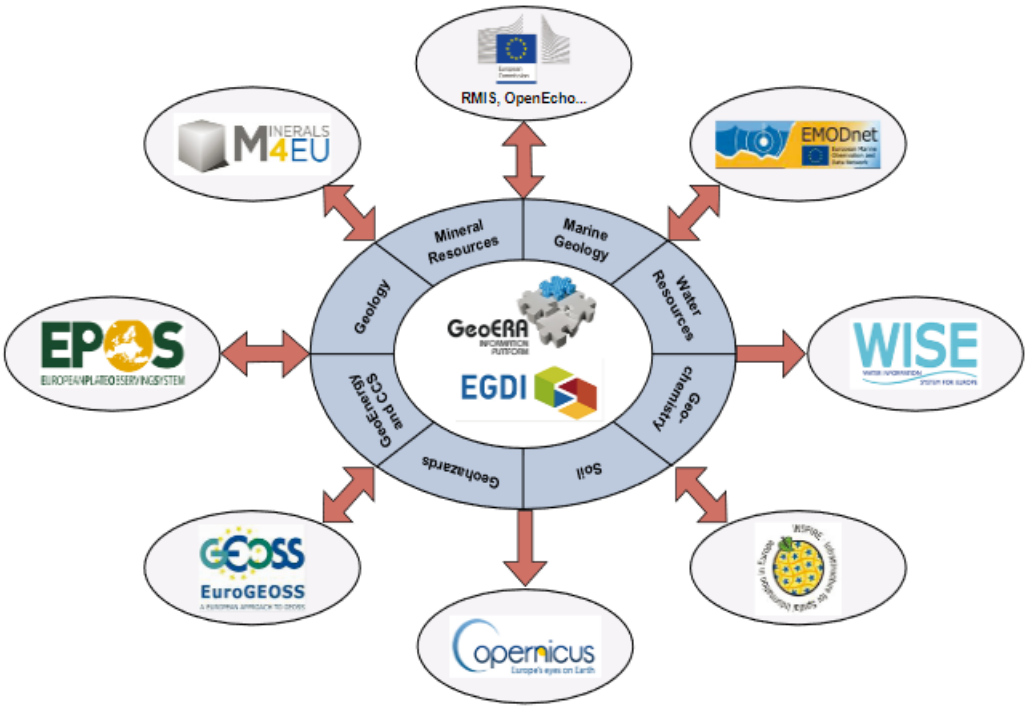
EGDI is composed of multiple components hosted by multiple teams:

- BRGM: European Geoscience Registry, URI Resolver, EGDI web portal, web statistics
- CGS: EGDI Metadata Catalogue, EGDI Metadata Harvesting
- GBA: 3D viewer and viewer for project vocabularies
- GeoZS: Min4EU Harvester, EGDI repository search application/Solr
- GEUS: EGDI web, EGDI admin, 3D database, 3D viewer, monitoring
- IGME: Search system (hosted at GEUS)
- ISPRA: eLearning Platform
- SGU: Component for showing timeseries (hosted at GEUS)

These components are also maintained and supported by multiple organizations: BGS, BRGM, CGS, GBA, GeoZS, GEUS, IGME, ISPRA, SGU, and TNO. With multiple skills: development, technical architecture, software architecture, interoperability and semantics.

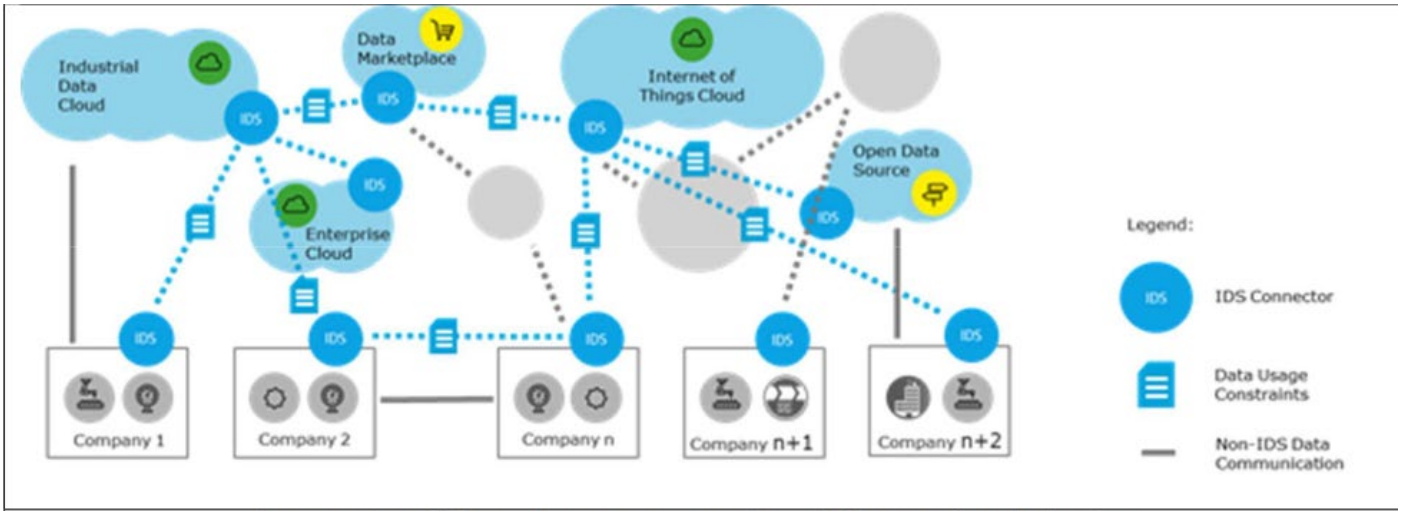
2.3 Related European project

EGDI is part of the European data science. As explained in D5.1, EGDI acts both as a data harvester/aggregator and provider for other platforms, such as EPOS. Being part of EPOS will also make it available for ENVRI-HUB by levering DCAT-AP and EOSC-HUB. A broader vision is given by Figure 1: EGDI interactions with International Initiatives.



During the GIP-P project, other platforms like AI4EU have made significant improvement to consume standard data and associated services. In the case of AI4EU[1], the link is established with Observation and Measurement data via SensorThing API. This is represented by the links to Internet of things Cloud in Figure 2: AI4EU architecture - links to data sources. If EGDI in the future will expose a SensorThing API, an AI4EU user will be able to use EGDI dataset for AI purposes.

[1] Interoperability design and implementation choices reference: https://www.ai4eu.eu/sites/default/files/inline-files/Deliverable_AI4EU_D2.7_M8_vfinal.pdf



International Data Spaces connecting different cloud platforms

3.EGDI ARCHITECTURE REVIEW

3.1 Use cases

There are at least 3 types of users in the EGDI context: EDGI administrators, data provider users, and public users (institutions, EOSC, European Commission, scientist, student, etc.)

So far, the EGDI users consulted were data providers (aka GSPs in the GIP context). They can be distributed over several main use cases as follows:

- Data users:
  - User searches documents, spatial data or metadata
  - User wants to view/download spatial data
  - User wants to view/download document
  - User searches metadata
  - User creates and updates metadata
- Data Providers
  - User uploads spatial data
  - User uploads documents and maps

The interesting part in these use cases is that they are data centric. Data can be of different quality, but if one wants to achieve a certain degree of Technical Readiness Level[2], one must make its Data FAIR, be it on the semantics (data models) or technical tools (services). These points are addressed by WP5.

What is missing here is the point of view of data users as opposed to data providers. Data users' needs and goals are well described in the last Horizon Europe call such as HORIZON-INFRA-2021-TECH-01-01 and HORIZON-INFRA-2021-SERV-01. Digital Twins and Workflow are expected in these calls, which implies machine to machine communication, and comes with new requirements. These points will be tackled in *Refining the target*.

3.2 GIP-P proposal

The GIP-P Proposal does not state anything else than FAIR data access through the use of a robust, modular and interoperable architecture.

"By creating an information platform that aligns and integrates with wider e-infrastructures across Europe and beyond (such as EPOS, EMODnet, Copernicus, European Open Science Cloud, GEOSS) we will open up data from the European geological surveys to be integrated with a wider range of earth science data. Our approach will be to build

[2]GIP-P D5.1 blue print on data and services : https://geoera.eu/wp-content/uploads/2019/10/D5.1.v1-GIP_blueprint-Data_and_services_architecture.pdf - page 18

the information platform in a modular way to produce core components that can be plugged in to other interfaces, initiatives and infrastructures. For example, functionality based on GeoSciML", GIP-P Proposal page 5

Our project will conform to data models and standards from INSPIRE, Open Geospatial Consortium (OGC) and the IUGS Commission for the Management and Application of Geoscience Information (CGI).", GIP-P Proposal page 6

"The platform will be based on a coherent architecture which will take into account experiences gained in previous EU funded data harmonisation projects and be built as an extension to the European Geological Data Infrastructure (EGDI).", GIP-P Proposal page 1

"The system will be developed upon outputs of different work packages: from requirements of WP2, from standards and interoperability in WP3, vocabularies from WP4 and it will be based on the architecture defined by WP5 and requirements from WP6 (data sets, data types and functionalities of the webGIS)", GIP-P Proposal WP7 page 22

"Develop the central components of the digital archive to support GSPs (reports, unstructual data, etc.)", GIP-P Proposal WP7 T4.1

3.3 GIP-P architecture target

In this chapter, we will quickly review some of the key elements on other GIP-P deliverables driving the architecture design and the achievement of the GIP project towards the use cases.

3.3.1 From D3.3 "Validation service specification and requirements"

Through D3.3[3], GIP-P was able to create a consistent list of raw data produced and used by the GSPs. For each of them D3.3 provide the corresponding "FAIR" data model/standard to used. Through assessment, a list of corresponding standard services to deliver those data was produced, both on the data model and on the services/API level.

Hence D3.3 was able to pave the way from raw data with private services to FAIR data Figure 1: Data model identified in D3.3 present the relevant data model, and FAIR services Figure 2: Standards and associated implementations, present the associated services and tools

[3] https://geoera.eu/wp-content/uploads/2020/01/D3.3-Standards-validation-procedures.pdf

Relevant data models identified and considered for D3.3:

Name of the document	Date / version
OGC GeoSciML	4.1 Rev 16-008
OGC GWML2	2.2 Rev 16-032r2
EarthResourceML	2.0 October 2013
INSPIRE AC (Atmospheric Conditions)	Revision 4618 This version corresponds to the content of the Implementing Rules
INSPIRE AF (Agricultural and aquaculture facilities)	
INSPIRE AM (Area Management)	
INSPIRE EF (Environmental Monitoring Facility)	(EU) No 1089/2010, No 102/2011, No 1253/2013 and the latest publicly available version of the data specifications of Annex I, II+III.
INSPIRE EL (Elevation)	
INSPIRE ER (Earth Resources)	
INSPIRE GE (Geology)	
INSPIRE LU (Land Use)	
INSPIRE MR (Mineral Resources)	
INSPIRE OF (Ocean Features)	
INSPIRE SO (Soil)	
INSPIRE NZ (Natural Risk Zones)	
EPOS BoreholeView	1.0.0
EPOS ModelView	1.0.0
ISO 19156 : Observations & Measurements	2.0 Rev 10-025r1 (OGC)
WaterML 2 - Part 1 / Timeseries	2.0.1 Rev 10-126r4
ISO 19115 / ISO 19139	
OGC Coverage Implementation Schema with Corrigendum (09-146r8)	Version 1.1.1 Published 2019-10-28

Figure 3: Data model identified in D3.3

Standards	Existing Technologies				
	MICKA//GeoNetworks	52°North Sensor Observation Service	Frost-Server	ArcServer/GeoServer/MapServer/TinyOWS	Rasdaman/Petascope
Catalogue Service	Y				
SOS		Y		Y	
SensorThing			Y		
WCPS					Y
WCS				Y	Y
WFS				Y	
WFS-T				Y	
WMS				Y	Y
WMTS				Y	
WPS				Y	

Figure 4: Standards and associated implementations

3.3.2 From D5.1 “GIP blueprint: data and service architecture of the overall system”

“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."

"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."*

Complementary to this 'state of the art' architecture, D5.1 proposed in every architecture pattern a complementary one as mentioned in its section 2. Target system:

- *"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)."*

During the course of the project this complementary pattern was one option almost all GSP preferred. They rather delivered copies of their geospatial data to the central EGDI database instead of setting up their own OGC services. This should be taken into account in the forthcoming projects both from an IT and human point of view (see Requirements 4 and 5).

3.3.3 From D5.2 "GeoERA Central System specification"

"*As identified in D 5.1, WFS 3 is the upcoming OGC standard API 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."

"1.7 Data publication alternative […] There is a specific effort to be carried on that aspect"

2 years after, things have moved and OGC API - Features - Part 1: Core is now officially published (OGC 17-069r3) and also endorsed as a validated INSPIRE download service[4].

3.3.4 From D7.2 "Finished testing the system and identifying problems"

[4] https://inspire.ec.europa.eu/good-practice/ogc-api-%E2%80%93-features-inspire-download-service

The purpose of testing is to evaluate the system toward the target and potentially KPIs. D7.2 [5]was able to enlighten the project on the strengths and weaknesses of EGDI. While the achievements are myriad, there are also weaknesses. For the sake of continuous improvements, we will only review the latter.

Considering the Web-GIS, these critical points where explained:

"[…] it will nearly also double the work needed to deploy new software and manage the setup in case of problems thereby harming agility.

Managing the setup is no easy task because of the many processes and modules involved."

"We can expect higher load as GeoERA matures with more user activity and more data products. In that case, it is advisable to introduce load-balancing gradually"

On a software architecture point of view, the difficulty to manage the set up shows the monolith approach of the centralDB architecture, and its drawbacks. This encourages us to find pragmatic solutions while pondering the benefits of the actual system and fostering agility and usability.

3.3.5 Deliverable synthesis

The deliverables anticipate the work to be done to provide a better support to the use cases: standard compliance, the associated software to be used, and the development of new components if needed.

3.4 EGDI GSP dataset access

Since 2016, the EGDI platform has been developed to meet most of the requirements identified through the use cases described previously.

It has already built-in capabilities to support standards either natively or by extensions, like WMS, CSW, etc., as listed in *Figure 3: Data model identified in D3.3* and *Figure 4: Standards and associated implementations*.

However, as show in *Figure 3: EGDI and datasets dissemination flow*, the actual publication process of a dataset, results in uploading and disseminating non-harmonized dataset. Indeed, uploaded datasets rarely comply with one of the data models of *Figure 3: Data model identified in D3.3.*

[5] https://geoera.eu/wp-content/uploads/2020/07/D.7.2-Finished-testing-the-system-and-identifying-problems.pdf

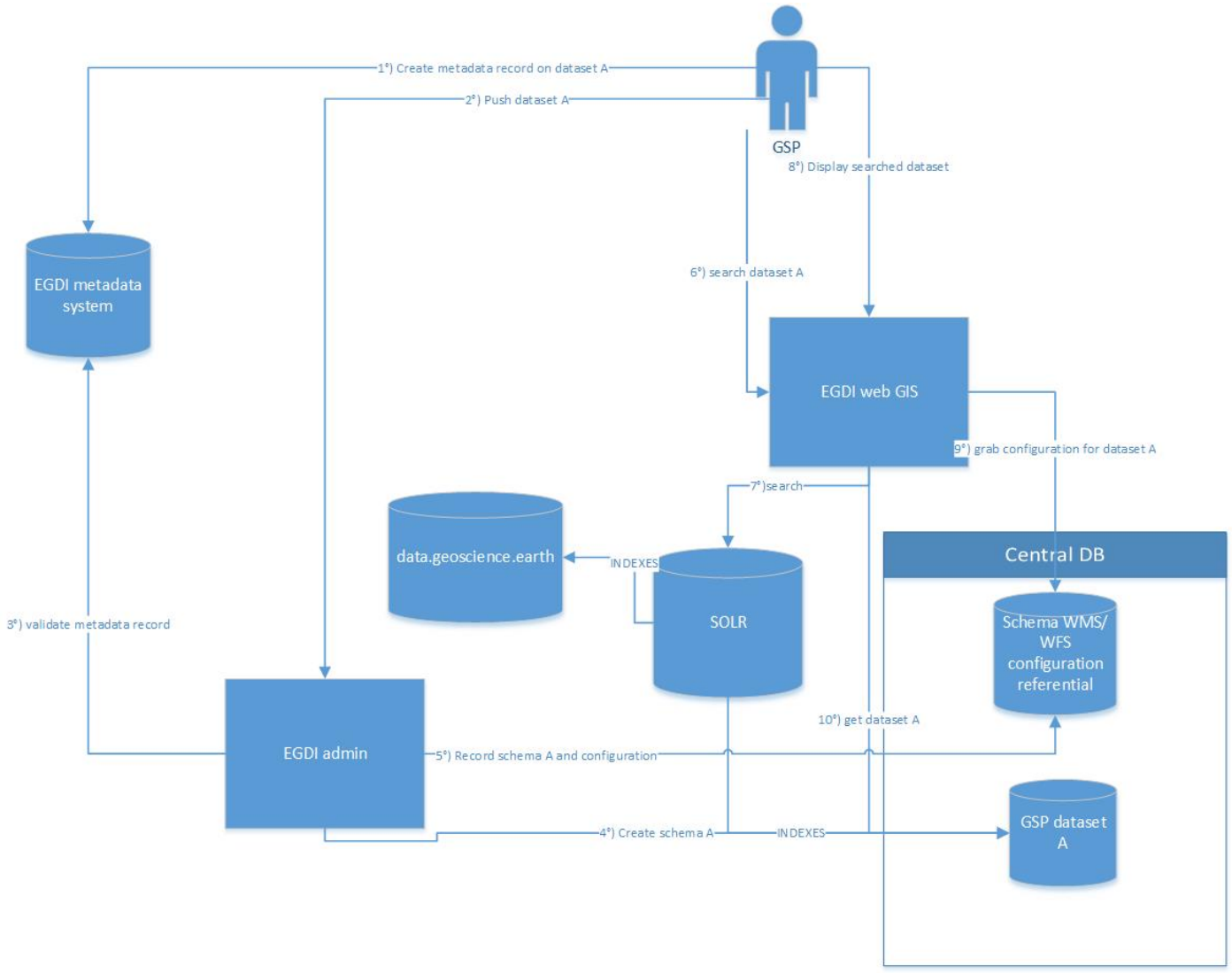


Figure 5: Simplified version of the actual EGDI and datasets dissemination flow

The dataset publication process can be synthesised as follows:

1) a data provider uploads data
    i) validates link to metadata
    ii) validates supported format
2) EGDI admin pushes dataset into schema A in the central DB
3) EGDI web-GIS uses Java + mapscript to store the WMS/WFS configuration (Those MapServer based WFS come without semantics)
4) EGDI also stores the definitions of the different layers, there thematization, optional filtering options and map definitions in the central database
5) A user search and the data available in EGDI web-GIS

4. EVALUATING FAIRNESS

As mentioned above there is a need to make EGDI as FAIR as possible to meet scientific user needs with regards to data. As experienced in ENVRI-FAIR, evaluating FAIRness of an Information System or Data Infrastructure, is not a trivial exercise. It first requires reaching a common vision/understanding of what reaching FAIR principles implies with regards the overall data architecture of EGDI and the IT practices this entails.

The consensus reached during GIP-P is introduced below and can be summarized as follows:

- Achieving FAIRness applies to the (meta)data (read metadata & data) entering the system (ex: in that case from the GSPs to EGDI) but ALSO to the corresponding (meta)data exposed by the system (ex: here from EGDI to the outer world). As such, the FAIRness evaluation has to be as holistic as possible as opposed to the evaluation of one component of the system (ex: a metadata catalogue or a WFS service).
- Following INSPIRE requirements is already a first step towards FAIRness. However, following INSPIRE to the fullest won't guarantee reaching 100% FAIRness. This has to be complemented by proper use of community standards (OGC, IUGS/CGI, W3C, RDA, ISO …), best practices (W3C, RDA, OGC) and APIs (OGC).

4.1 Findable

The first step to re(use) data is indeed finding them. Metadata and data shall be easily accessible and findable for both human and machines.

4.1.1 F1: metadata and data are assigned a globally unique and eternally persistent identifier

For Metadata records, it should be stressed that the target is to provide a construct that helps to resolve the metadata record for human or computer reading, taking into account the following constraints: hosting organization can change, metadata representation as well.

Currently INSPIRE Metadata Technical Guidelines V.2.0.1 are still based on ISO 19115:2003. This is clearly a limitation with regards this FAIR item. In ISO 19115:2003, the fileIdentifier (characterString) is used to identify the metadata record. This uuid triggered many disruptions in national SDI metadata architectures and portals. Indeed, when harvested across catalogues, it can conflict with others (same uuid, different context), it is not always respected (uuid replaced which creates duplicates in distributed systems where the same metadata record can exist in more than one catalogue)

Thanks to this experience, it is now replaced in 19115:2014 by a metadaIdentifier (MD_Identifier enriched with codeSpace) which provides the solution to achieving F1. There is a need to go beyond 19115:2003. As such there are exchanges with EC JRC to update INSPIRE Metadata Guidelines to this new version especially because Research Infrastructure are not bound to INSPIRE MD Guidelines which creates issues having to manage compliancy to 2 ISO versions of the standard.

While INSPIRE Metadata Guidelines are not updated accordingly, there is thus a need to go beyond those guidelines to reach F1.

With regards to data, INSPIRE metadata have that type of information named URI (Uniform Ressource Identifier) or via a DOI. There is no mandatory element as to assign URI to data instances under INSPIRE. However, there is a movement starting from W3C Spatial Data on the Web Best Practices to assign URIs to features encountered in INSPIRE spatial datasets

4.1.2 F2. data are described with rich metadata (defined by R1 below)

In INSPIRE, metadata are information describing data, to ease their inventory, their discovery and reuse. One can easily mention some 'richness' in their content.

4.1.3 F3. Metadata clearly and explicitly include the identifier of the data they describe

In INSPIRE Metadata Technical Guidelines V.2.0.1, TG Requirement 1.3 mandates that a unique identifier shall be given for each described dataset or dataset series.

4.1.4 F4. Metadata and data are registered or indexed in a searchable resource

INSPIRE mandates that each spatial dataset series falling in its scope is described by a metadata record. That these metadata record are kept up to date and, like data, published on Internet.

Again, on the data side, there is no mandatory action on the INSPIRE side. However, again thanks to W3C Spatial Data on the Web Best Practices, there is a dynamic to make Spatial Data Infrastructure (SDI) data indexable by search engines

4.2 Accessible

Once data found, the user needs to know how to access it, sometimes via authentication and authorization.

4.2.1 A1. Metadata and data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

INSPIRE directive guidelines recommends the deployment of data and metadata publication tools compliant to the Open Geospatial Consortium (OGC) specification such as CSW (metadata), WMS (maps), WCS (coverage), WFS and now OGC API – Features (spatial features), SOS and now OGC SensorThings API Part 1 (observations). Those protocols support operations allowing to retrieve an element by its identifier.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

INSPIRE directive recommends free access to geographic datasets. However, Spatial Data Infrastructures (SDI) allow secure access to datasets such as those falling under GDPR.

OGC specifications are designed so that they can easily be deployment in networks where an authentication and authorization layer exists.

4.2.2 A2. metadata are accessible, even when the data are no longer available

As far as we know, there is no clear guideline from INSPIRE on that point. This is more part of the strategy to be defined within each SDI. In such a case, metadata can be kept and still be accessible to everyone, tagging it with an “obsolete” status in the Identification section.

4.3 Interoperable

Usually data are used amongst other datasets within applications and treatment chains to be stored, analyzed, …. That is one of the elements that drove INSPIRE directive set up. Interoperability covers both technical and semantics aspects. Data and metadata can be considered interoperable when they have interconnected connections (for example as required by the INSPIRE rules).

4.3.1 I1. Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation

INSPIRE implementing rules on metadata are mapped to ISO standards on metadata. Datasets semantics is done either using INSPIRE data specification or internationally agreed ones (ex: IUGS CGI / OGC GeoSciML, Observations & Measurements).

All of the above have been for decades formally noted in UML. There is now a movement to have the domain knowledge acquired through those years transcripted into SemanticWeb notations. Such dynamic (linked to W3C: Data on the Web Best Practices) include metadata in GeoDCAT_AP, Observations in W3C:SOSA, light weight GeoSciML ontologies, …

4.3.2 I2. Metadata and data use vocabularies that follow FAIR principles

INSPIRE directive mandates the use of controlled vocabulary for both metadata (ex: keyword) and data (ex : INSPIRE register federation).

Taking the term “vocabulary” to its broader sense (vocabulary = controlled vocabularies but also ontologies), elements mentioned under I1 are also valid.

4.3.3 I3. Metadata and data include qualified references to other metadata and data

This aspect is not really covered by ISO 19115 or INSPIRE Metadata Technical Guidelines (but can be completed in some cases by INSPIRE Data Specifications Technical Guidelines).

However, such references are partially made possible by proper application of F1.

Provided that metadata and data are assigned a globally unique and eternally persistent identifier, it is then possible to link to them reusing those identifiers (ex: URIs) be it between metadata records or between features according to a data model (ex: between a Well and an Aquifer).

The hardest part to achieve will be to qualify that association as, most of the time, the underlying data models don’t allow to provide information regarding the quality of the association.

4.4 Reusable

FAIR ultimate goal is to optimize data reuse. Especially in the light of recent incentives from the European Commission such as in “Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information” (https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32019L1024). To reach such a goal, metadata and data shall be properly described so that they can be integrated and combined into different contexts and usages.

4.4.1 R1. metadata and data have a plurality of accurate and relevant attributes

R1.1. metadata and data are released with a clear and accessible data usage license

The INSPIRE directive mandates that condition applying to access and reuse of data are defined by a license and appear in Metadata in the corresponding information element.

R1.2. metadata and data are associated with their provenance

INSPIRE metadata allow answer the following questions:

- By whom and where data was produced,
- When (creation, publication, update)
- How and
- Why. Technically it is not an INSPIRE triggered element, but ISO 19115 provide the necessary information element to do so

R1.3. metadata and data meet domain-relevant community standards

With regards metadata, ISO 19115 is since 2003 the reference for geographic dataset publication. SDI almost all now expose their datasets following this international standard that is compliant with INSPIRE metadata requirement. More recently since W3C and OGC experts started working together, DCAT and is various application profiles (AP) are also progressively implemented by SDI (ex: GeoDCAT-AP)

Data are made available either using INSPIRE data specification or internationally agreed data specifications (ex: IUGS CGI / OGC GeoSciML, Observations & Measurements).

5.REFINING THE TARGET

We have demonstrated so far that the use cases already implemented are mostly data provider centric. They are fit for purpose in the context of each GSP.

We have also demonstrated in 4.2 that the deliverables anticipate the needs for:

- Interoperability,
- Standards,
- Scalability.

This is consistent with the usage of IT developments made by scientist nowadays. Indeed, scientists also develop software. First, they use graphical user interfaces for discovering data. Secondly, they use APIs to filter and download data in order to perform computation for science.

For example, to assess/predict the level of the water table of an underground reservoir, they use:

- the actual level of water in the reservoir,
- the air temperature,
- the water level based on meteorology prevision,
- etc.

The scientists use multiple services as data source. In this respect, they would rather use known and standard APIs rather than home-made ones which would drastically increase the process to discover/ingest new data. Moreover, standards empower the scientists with existing software libraries. This is far easier than developing software libraries of their own: which would be prone to waste of effort, duplicated work and technical debt. On this specific topic if we have a deeper look at the previous mentioned Horizon Europe call:

- HORIZON-INFRA-2021-TECH-01-01
  - *"[…]pre-operational prototype of an interdisciplinary Digital Twin[…]"*
  - *"[…]a robust framework enabling Researchers to ensure the quality, reliability, verifiability of the data[…]"*
  - *"[…]enabling research communities to continuously learn and update themselves from data and information originating from different sources[…]"*
  - *"[…]Work under this topic should reach a sufficiently high TRL level (6-7)[…]"*
- HORIZON-INFRA-2021-SERV-01:
  - *"[…]wider, simplified, and more efficient access to the best research infrastructures[…]"*
  - *"[…]better management, including implementing FAIR data principle, of the continuous flow of data collected or produced by research infrastructures[…]"*
  - *"[…]to further develop the remote or virtual provision of services may also be supported[…]"*

Following this topic, some scientists are more and more using Virtual Research Environment (VRE) powered by tools like Jupyter notebooks, Taberna, Zepplin, etc. It must be noticed that EOSC providers already offering VREs [6]and that other H2020 projects are providing Jupyter notebooks (ENVRI-FAIR, EPOS, VRE4EIC, AI4EU etc.).

These environments need standardized and interoperable APIs and semantics, and adequate tools that provide performances too match user access and complex requests, security (validate user identity through federation like eduGain, etc.).

In GeoERA, the thematic projects seem to be far from ready to use interoperable standards and best practices for data sharing. Despite the detailed attributes/codelist mapping works carried out by WP2 and WP3, the interviews conducted with almost all of them lead us to understand that they are not ready to implement the chosen standards.

The EGDI platform will have to answer to these needs by the mean of adequate process, communication, training and efficient tools.

5.1 Issues

Issue 1: How to address the lack of skills and motivation?

As described here above from GIP-P Proposal to the use cases gathered during the project, there is a real convergence towards a better support for standardization/interoperability. This effort, although already started, requires efforts and more development as emphasized by the fact that few data providers provide dataset compliant with any standard. One of the main issues is the lack of knowledge and willingness to implement interoperability standard from the data providers. Implementing those is often considered as an extra constraint with little value added by them.

Issue 2: How to make the usage of interoperability expertise as efficient has possible?

With the help of WP2 and WP3 deliverables, WP5 has delivered test beds using the HIKE datasets. This work consisted in an assessment of the potential standard to use and adapt datasets to the standard. This had required discussions and interactions between the data provider and interoperability experts on a case by case basis.

Issue 3: How to sustain this effort in the long run?

Harmonization requires both interoperability experts interacting with scientist to make the corresponding transformation (aka mapping) from non-harmonized data to harmonized data. We can imagine the huge amount of time to deal with each individual dataset. In addition, this would be important to archive this mapping for scientific reproducibility.

[6] Jupyter notebook as VRE examples : https://www.eosc-hub.eu/training-material/egi-jupyter-notebooks-examples

Issue 4: How to use the adequate tools to provide a better standard support and compliance (INSPIRE, FAIR, etc.) at the very core of EGDI

Once the datasets are made interoperable, it would be a waste to disseminate them with a loss of information. WP3 identified and recommended the adequate tools for each kind of standard data. However, EGDI is presently using mapscript only, because its flexibility, but it may be limited in terms of FAIR compliance and ease of integration of evolutions in interoperability.

Issue 5: How to provide better tools for a better standardization, FAIRness, without breaking the existing EGDI platform?

By introducing fit-for-purpose tools through prototyping, the EGDI platform may become too complex to maintain if kept as a monolith. As explained above, only leveraging mapscript begins to show its limits in term of standard compliancy. Although Mapscript covers WMS/OGC API MAP, it does not fully comply to WFS/OGC API Feature and does not provide SensorThings API support.

Issue 6: How to ensure that the technical platform keeps up with ever growing demand and integrate new needs (domain & IT)?

We can cite the last Horizon Europe call such as HORIZON-INFRA-2021-TECH-01-01 and HORIZON-INFRA-2021-SERV-01, where Digital Twins and Workflow are expected. Such data science implies FAIRness and level of performance of the data access. Curation and standardization are also to facilitate the usage of the data. Data science requires always growing data size and access performance.

Issue 7: How to evaluate EGDI FAIRness

The overall consensus reached is described in the previous section. Evaluating FAIRness of all EGDI (= all metadata, data, services) was not realistically achievable within the course of this project.

However, the action started in this context triggered the required momentum and consensus with regards practices necessary for a thorough FAIRness evaluation of EGDI in a subsequent project. A first assessment round on a reduced number of datasets is also important to exemplify the consensus thus make it more concrete by everyone.

5.2 Potential new requirement

From the issues described above, we suggest an extended list of requirements:

- Requirement 1: Engage expert in standardizing datasets → *this addresses issue 1*
- Requirement 2: Create or use an interface to study and map raw datasets against standards → *this addresses issues 2 and 3*
- Requirement 3: Reuse previously standardization exercises → *this addresses issue 3*
- Requirement 4: Extend EGDI with the adequate tools → *this addresses issues 4, 5, 6*
- Requirement 5: Trigger motivation into using interoperability standards and best practices → *this addresses the 'motivation' side of issue 1*
- Requirement 6: Evaluate EGDI FAIRness → *this addresses issue 7*

6.ACTIONS

6.1 Requirement 1: Engage experts with data providers

It has to be noticed that this work has already been carried out in GIP-P:

- in WP7 on HOVER dataset, in alignment to WP3 recommendations.
- in WP5 on HIKE dataset, in alignment to WP3 recommendations.
- It has started in WP8 on the MUSE datasets.

This 'peer interoperability set up' has proved really useful as it allows to spread knowledge and ease communication with regards to interoperability with people who actually create and reuse data.

6.2 Requirement 2: Facilitate mapping activity

Teams in WP5 has proposed a prototype of a semantic mapper. A mocked-up interface has been produced to illustrate the easiness of mapping. However, tools already exist like Hale Connect[7] (server with ad hoc license) or Hale Studio[8] (desktop only, and open source). There will be a need to discuss which tools should be used in the future.

6.3 Requirement 3: Reuse mapping or preconfigured data structure

The mapping previously created has to be saved to be reproducible. We want to capitalize on previous standardization expertise. This would allow users with zero knowledge to reuse previous mapping. Hence, experts would only be mobilized for new or extended mapping. Over time, it is expected to engage in a virtuous cycle that will:

- Create more engagement of users,
- Spread knowledge,
- Maintain the expert intervention to a minimum.

Another way to enforce a 'reuse' philosophy is to provide users with preconfigured data structures already compliant to interoperability standards at the start of a project that will deliver data to the EGDI platform.

This will help users starting their project and get progressively acquainted with the semantics of the data standards in their domain and, when need be, add their extra need for information.

Otherwise relying only on mappings have two implications:

[7] https://www.wetransform.to/products/haleconnect/

[8] https://www.wetransform.to/products/halestudio/

- Domain experts work with the semantics and IT patterns they create for a given project. This costs time and money and can be faulty as this is not their core expertise,
- And when integrating each new project, those specific semantics/IT patterns have to be mapped to interoperable standards. Which costs time and money a second time and can lead to faulty interpretations as IT/data experts are not those who created the datasets and conducted the domain project in the first place.

There is a need to move from a 'curation' point of view to a more holistic view of the system.

6.4 Requirement 4: Extends EGDI

WP5 work led to two new services implementation that make full use of EGDI extension capabilities: FROST and GeoServer. In this case FROST provides Reusability of data for example to AI4EU users.

WP5 intended to capitalize this effort in a semantic mapper that gives the capability to advanced user to create mapping from a project dataset attributes to standard-related attributes.

FROST and GeoServer are known to be scalable (be it horizontally or vertically) and moreover they are known to address data centric uses cases. These services are potential new extensions for EGDI.

6.5 Requirement 5: Trigger motivation

Another angle of approach is to consider domain experts are not motivated because they don't see the value added of the investment into using interoperability standards and best practices.

There is a crucial need to ensure that the overall EGDI system is up to those standards.

Indeed if on both Server (EGDI as a provider of data) and client side (ex: QGIS, Jupyter notebook, Python libraries, …) using interoperable datasets proved to be streamlined from discovery to reuse in the users' environment, that would help them make the necessary effort.

There could be a mid-term goal to 'provide a fully interoperable ecosystem' for domain scientists to work with which would provide harmonized dataset by leveraging adequate communication, process and tools. Leaving the 'one shot exercise' done in mapping HIKE, to such a coordinated effort, people will see the return on investment and may be more willing to actually abide by the practices data and IT specialists ask them to implement.

Actually, there is a longer-term goal to have domain scientists apply those practices even without knowing they are doing so. Which implies an active investment in tools so that those practices are already applied within them.

6.6 Requirement 6: Evaluate EGDI FAIRness

Leveraging on the experience from participant in ENVRI-FAIR it was decided to reuse the CSIRO 5 stars Data Rating Tool (http://5stardata.csiro.au/) and then translate the result into FAIR principles. The CSIRO proposed a pragmatic and easy approach to evaluate the FAIR principles mapping them to standards and practices from well-established communities when it comes down to data sharing (W3C, RDA, OGC). It has greatly facilitated the discussion amongst interoperability experts.

The approach was holistic in the sense that we evaluated FAIRness of data, metadata, and services. To reach a global appreciation of compliance to FAIRness, every data, metadata and services should be assessed. This was not possible in the context of GeoERA, timewise and also because not all datasets were delivered. Hence, we focused on evaluating datasets known by the interoperability experts.

This exercise was really beneficial to the project to rise the FAIR principle awareness and reaching a common consensus. Moreover, this exercise involved experts from organization within the EGDI consortium (BGS, BRGM, IGME), which is an outcome that should not be overlooked in terms of sustainability of EGDI.

This resulted in two spreadsheets with consensus amongst the expert: CSIRO 5 stars assessment and its translation to FAIR principles. Those are available in section 11 “FAIRness assessment exercise”.

With regards URIs, has mentioned in other WP5 deliverables, having a base URI based on https://data.geoscience.earth proves its value added for many FAIR principles as it allows already assigning URIs to metadata, datasets, data instances, vocabularies (controlled ones & ontologies), APIs, etc… and to link them between one another. GeoERA WP4 and other EGS related projects are already making an important use of it. This could be propagated to other EGDI components.

There have also been important discussions with regards some component of EGDI such as EGDI Metadata catalogue and how to carry out a more ‘holistic’ approach.

These are also available in section 11 “FAIRness assessment exercise”.

Some conclusion can be taken from that test evaluation exercise that need to guide the further upcoming discussions/work on EGDI and FAIR.

- The first and more important one is that the FAIRness assessment is done by an interoperability expert group. Experts are people able to consider FAIR principles side with the best practices from communities (ex: W3C, RDA, OGC); they are aware of the work going on within FAIR related groups and able to actually understand what those principles mean from the IT perspective. From our experience in other project, these skills are mandatory to grasp the complexity of FAIRness and avoid bias.
- The second is directly linked to the previous. Interoperability experts doing the assessment shall build on the experience gained in equivalent projects (ex: ENVIR-FAIR), not reinvent their own wheel and accept to adjust their usual practices in the light of emerging (internationally agreed) ones.
- The third one is that one shouldn't be rating each EGDI component individually for FAIRness (or the whole EGDI platform for FAIRness), the rating should rather be done on the data and metadata on their own.

7.EXTENDED EGDI ARCHITECTURE

7.1 Actual EGDI Component

The actual EGDI architecture has been built with extension capabilities. This extension, modules or components can be deployed either locally to the EGDI main component or remotely, as shown on the following diagram Figure 4: Actual EGDI components.

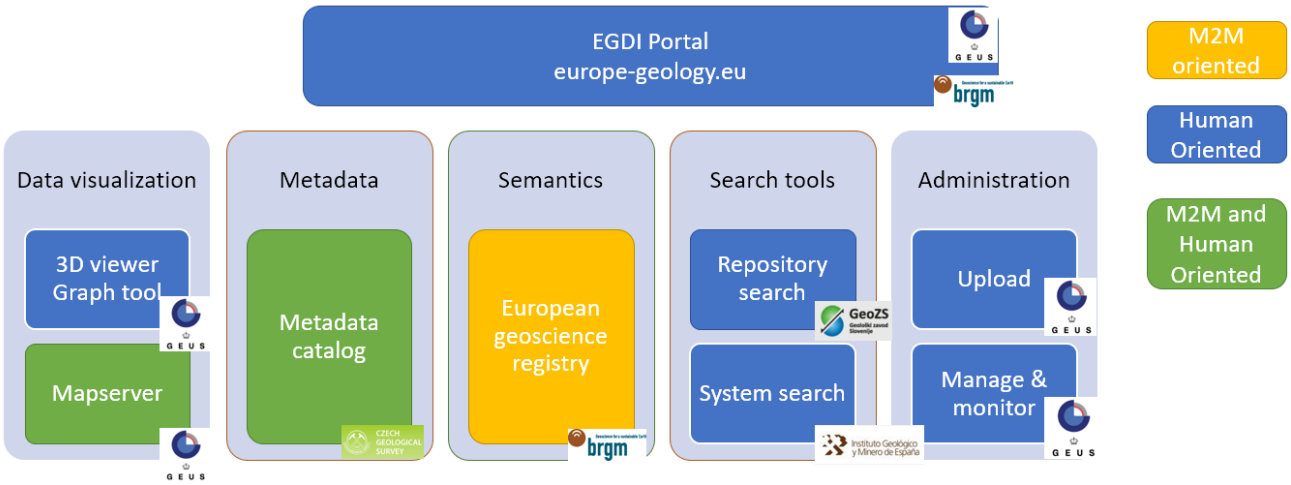


Figure 6: Actual EGDI components

We can see the components are mostly human oriented. Although this is sufficient in the context of GIP-P, we have to consider the strengths but also the weaknesses for the future. The EGDI platform has been designed to be extended, and we have made full use of this capability in WP5.

7.2 Extending the architecture

An overview of suggestions for an extended architecture is given by Figure 7: EGDI extended architecture - component view. It introduces potential new components in the data access and semantics:

- Data access with GeoServer and FROST.
- Semantics with the semantic mapper.

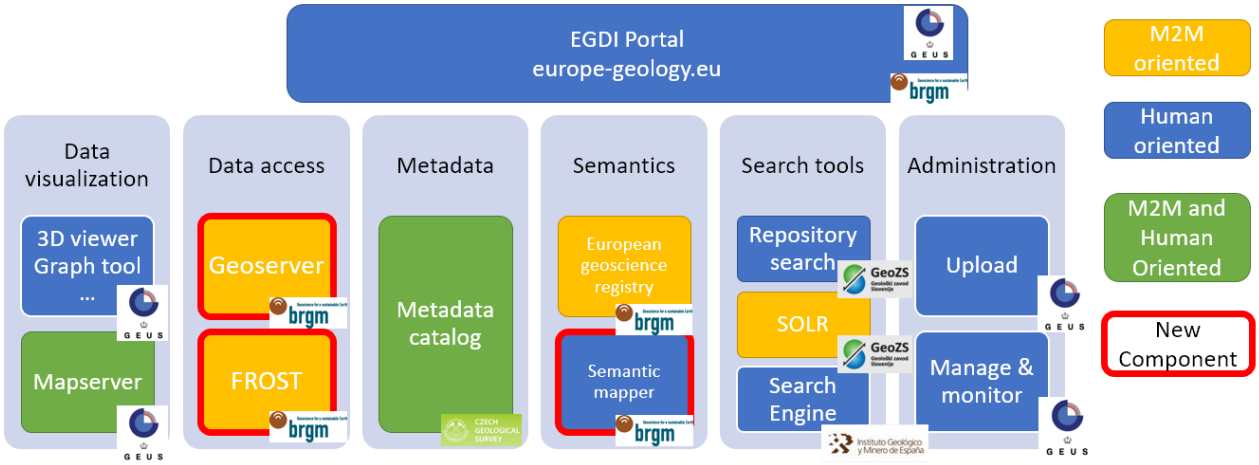
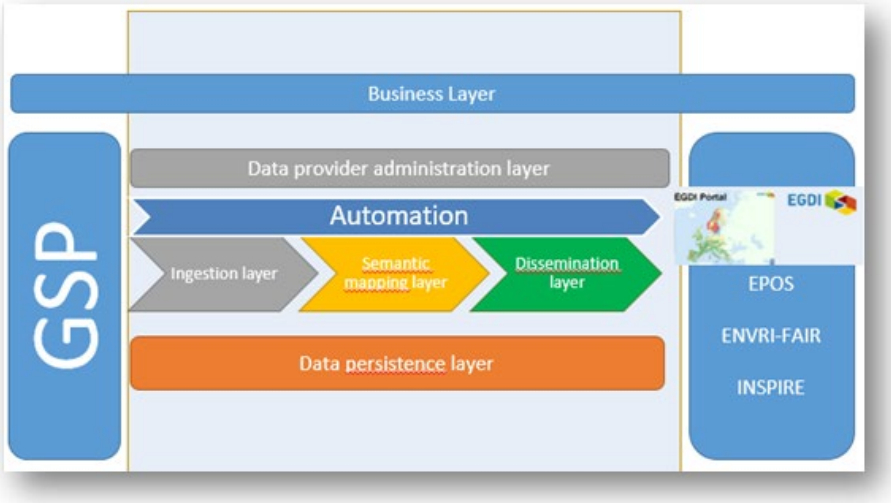


Figure 7: EGDI, suggestions for an extended architecture - component view

At a higher level the EGDI platform would address:

- Ingestion layer
- Semantic layer
- Dissemination layer: FAIR
- Data persistence layer: document, datasets, etc.
- Automation: automatic scaling of the services



7.2.1 Updating Data delivery process

7.2.1.1 Ideally

- User create semantically interoperable dataset

- User describes the metadata associated to the dataset
- User uploads the dataset
- Admin publishes the dataset
- Dataset is made available with:
    - Geospatial: mapserver/geoserver and downloadable on GeoPackage / GeoTIFF / NetCDF format
    - Observation and measurement: Maybe Frost/geoserver
    - Document: document repository
- Metadata is updated accordingly

7.2.1.2 Pragmatic approach

- Optional steps, depending on the capability of the data provider or willingness/time available:
    - User creates semantically interoperable dataset
    - Interoperability expert engage discussion with the user and leverage semantic layer to enhance interoperability
- User describes the metadata associated to the dataset
- User uploads the dataset
- Interoperability expert engage discussion with the user and leverage semantic layer to enhance interoperability
- Data providers publish the dataset
    - Dataset is made available with:
    - Geospatial: mapserver/geoserver
    - Available for download on the format they have been delivered on (GeoPackages, …)
    - Observation and measurement: Maybe Frost/geoserver
- Document: document repository

7.2.2 Impact on the actual architecture

The diagram Figure 6: EGDI data flow extended for better performance and FAIRness illustrates the dataflow within the platform EGDI could be if we extend it to match the above requirements.

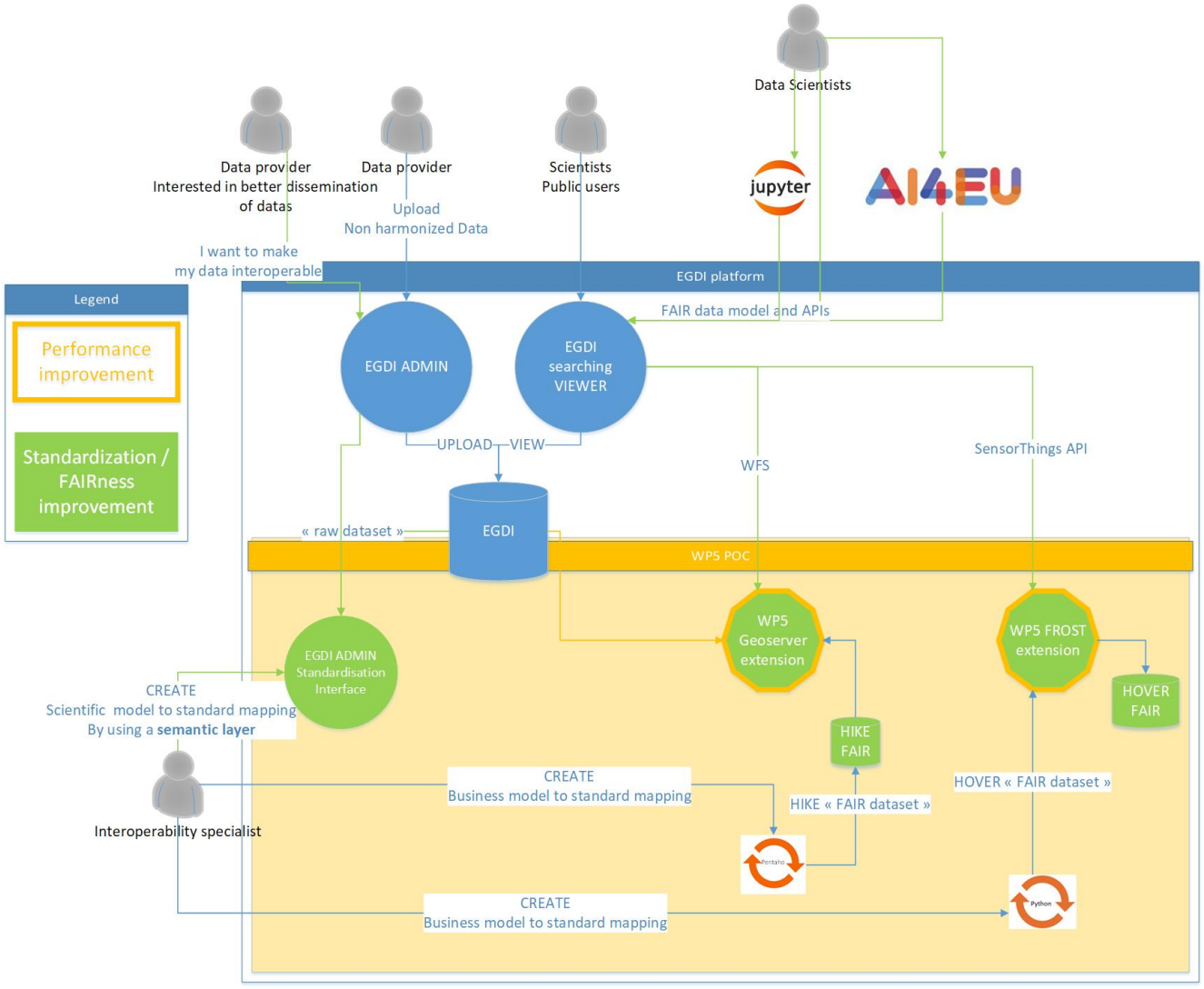


Figure 8: EGDI data flow extended for better performance and FAIRness

From the EGDI admin, with a new possible component (Semantic Mapper) that can be embedded, we add the following capabilities:

- Engage experts and data providers → *this addresses requirement 1*
- Facilitate mapping: easy interface to map attributes to a given standard data model, Figure 7→ *this addresses requirement 2*

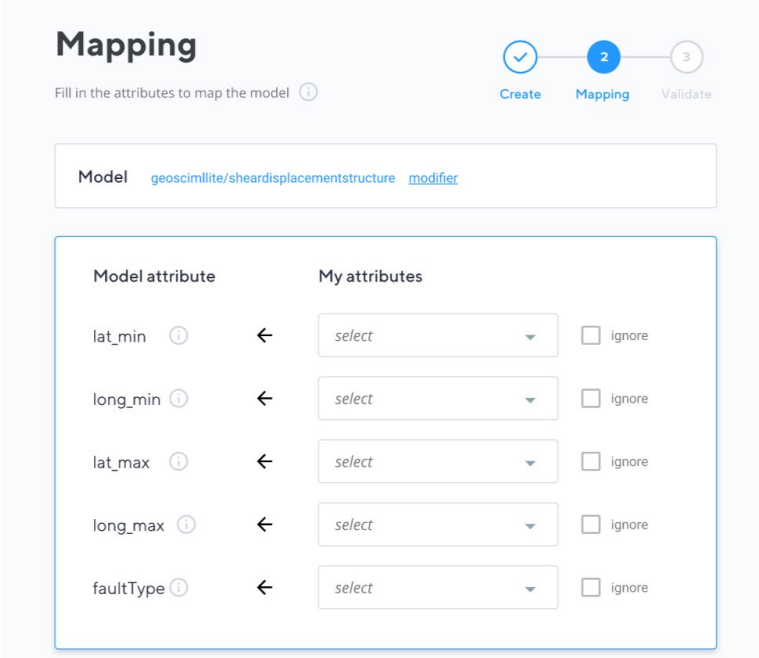


Figure 9: example of interface for mapping non harmonized dataset attributes "My attributes" to standard attributes "Model attribute" where the standard model is geoscimllite/sheardisplacementstructure

- Reuse mapping → *this addresses requirement 3*
- Extend EGDI without breaking the existing components → *this addresses requirement 4*

This new component would be able to trigger an ETL (Extract Transform and Load) component, represented by Pentaho and Python in Figure 2, which meets the following requirements:

- Facilitate mapping → *this addresses requirement 2*
- Reuse mapping → *this addresses requirement 3*
- Extend EGDI without breaking the existing components → *this addresses requirement 4*

Finally trigger the adapted tools for the mapped dataset:

- Extend EGDI without breaking the existing components → *this addresses requirement 4*

Potential other extensions:

- API on top of EGDI ADMIN to facilitate the extension
- API for the repository search (potentially with the OPENAIRE approach – OAI-PMH)
- Having a WFS (and its successor OGC API – Features) endpoint with the standardized data models for the entire Europe.

- Underlying Kubernetes infrastructure to host the extension to address scalability needs, be it vertical or horizontal scalability[9].

8.CONCLUSION

WP5 has delivered 3 prototypes which should be considered if the EGDI platform want to go beyond web GIS by answering to data centric and machine to machine oriented uses cases. Integrating these prototypes or equivalent solutions in the EGDI platform would help reaching a higher FAIRness level, thus fulfilling incentive from both EU INSPIRE directive and Open-Science policy.

These prototypes were made leveraging existing tools with no changes, with the idea that the EGDI consortium will be able to focus on building ad-hoc services with added value. It has to be noticed that the achievement is not only the prototypes themselves indeed the capability to implement these solutions reinforced the fact that EGDI architecture can be extended.

Throughout the FAIR assessment we were able to create an expert group that reached consensus on the FAIR principles. Therefore, this enhances the EGDI consortium skill panel: it can now leverage the knowledge of multiple teams composed of FAIR experts, software architects, developers, and as such this decoupling is a serious asset of the EGDI platform.

This deliverable gives a potential path for improvements of the EGDI platform following GIP-P. Although it focuses on the issues it should be clearly stated that these issues are not undermining the platform as of now. We suggest that they are considered in the forthcoming IT era where everything revolves around data, and their associated services. These services must foster innovation and facilitate decision making.

[9] Autoscaling Kubernetes: https://medium.com/nerd-for-tech/autoscaling-in-kubernetes-hpa-vpa-ab61a2177950

ANNEX A. **FAIRness assessment exercice**

A first outcome of the analysis exercise is available below. Another spreadsheet has yet to be reviewed by the group.

1. CSIRO 5 stars reviewed by peers

	x: implemented, a:anticipated	Datasets				
	CSIRO 5 Star evaluation	**DARLINGe.ActiveWell**	**hover_wp3_35c_indlayer2**	**hover FROST**	**hover/ntt_epsg_3034**	**hike_faultdb**
publication and indexing						
2. Published - is the data accessible to users other than the creator or owner?						
	No				**x**	
	By individual arrangement					
	File download		**a**			**x**
	Institutional or community repository					
	Bespoke web service (informal API)					
	Bespoke web service (OpenAPI/Swagger)					
	Standard web service API (e.g. OGC)	**x**		**x**		**x**
3. Citeable - denoted using a formal identifier						
	Not citeable					
	Local identifier					
	Web address (URL - not guaranteed stable)	**x**	**a**	**x**	**x**	**x**
	Persistent web identifier (URI)					
4. Described - tagged with metadata						
	No metadata					
	Abstract and keywords					
	Basic metadata (e.g. Dublin Core)					
	Specialized metadata (e.g. Darwin Core, ISO 19115/19139, schema.org scientific data profile)	**x**	**a**	**a**	**x**	**x**
	Rich metadata using multiple standard RDF vocabularies (e.g. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)	**x**	**a**	**a**	**x**	
5. Findable - indexed in a discovery system						
	no					

	local or internal system only					
	community wide or jurisdictional system	x	a	a	x	x
	highly ranked in general purpose index (Google, Bing etc)					
linked and useable						
6. Loadable - represented using a common or community-endorsed (i.e. standard) format						
	bespoke format (text, binary)					
	one standard format, denoted by a MIME-type		a	x	x	
	multiple standard formats	x				x
7. Useable - structured using a discoverable, community-endorsed (standard?) schema or data model						
	no formal schema				x	
	explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar		a			x
	community-shared schema or data model , available from a standard location	x		x		
8. Comprehensible - supported with unambiguous definitions for all internal elements						
	local field codes or labels				x	
	labels with full text explanations					
	community standard labels (e.g. CF Conventions, UCUM units)					
	some fields linked to externally managed definitions		a	a		x
	all fields linked to standard, externally managed definitions	x				
9. Linked - to other data and definitions using public identifiers (e.g. URIs)						
	no links		x		x	
	in-bound links from a catalogue or landing-page	x				
	out-bound links to related data and definitions			x		x
10. Licensed - conditions for re-use are available and clearly expressed						
	no license				x	
	license described in text	x	x	x		
	link to a standard license (e.g. Creative Commons)					x

maintenance and provenance						
11. Curated - commitment to ensuring the data is available long term						
	once-off dump, no ongoing commitment					
	best effort, project website	x	x	x	?	x
	public or institutional repository (e.g. CKAN, GitHub)					
	certified repository					
12. Updated - part of a regular data collection program or series, with clear maintenance arrangements and update schedule						
	one-time dataset	x	x	x	x	x
	part of series - occasional/irregular update					
	part of series - regular scheduled updates					
13. Assessable - accompanied by, or linked to, a data-quality assessment and description of the origin and workflow that produced the data						
	no quality or lineage information					
	text lineage statement	x	x	x	x	x
	formal provenance trace (e.g. PROV-O)					
14. Trusted - accompanied by, or linked to, information about how the data has been used, by whom, and how many times						
	no information about usage					
	usage statistics available					
	Clearly endorsed by reputable organization or framework	x	x	x	x	x
Project, organisational, institutional						
15. Complexity of the project						
	low					
	medium	x	x	x	x	x
	high					
16. Cross-organisational project?						
	1 organisation					
	2-4 organisations					
	5 or more organisations	x	x	x	x	x

2. CSIRO 5 stars reviewed translate into FAIR principles

MD stands for MetaData.

x: implemented, a:anticipated	MD	DATA	MD	DATA	MD	DATA

CSIRO 5 Star evaluation translated to FAIR Principles	hover/ntt_epsg_3034	hover/ntt_epsg_3034	HOVER WP3 D3.5c	HOVER WP3 D3.5c	hover FROST	hover FROST
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.						
F1: (Meta)data are assigned globally unique and persistent identifiers	**x**		**x**		**a**	
F2: Data are described with rich metadata	**x**	**x**			**a**	
F3: Metadata clearly and explicitly include the identifier of the data they describe	**x**	**x**			**a**	
F4: (Meta)data are registered or indexed in a searchable resource	**x**		**x**		**a**	
Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.						
A1: (Meta)data are retrievable by their identifier using a standardised communication protocol	**x**	*WMS only, data are not retrievable by any method*	**x**		**a**	**x**
A1.1: The protocol is open, free and universally implementable	**x**			**a**	**a**	**x**
A1.2: The protocol allows for an authentication and authorisation where necessary	**x**			**?**		**x**
A2: Metadata should be accessible even when the data is no longer available	**x**	**x**	**x**	**x**	**a**	
The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.						
I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	**x**		**x**		**a**	**x**

I2: (Meta)data use vocabularies that follow the FAIR principles	**x**		**x**		**a**	**a**
I3: (Meta)data include qualified references to other (meta)data					**x**	**x**
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.						
R1: (Meta)data are richly described with a plurality of accurate and relevant attributes					**a**	**x**
R1.1: (Meta)data are released with a clear and accessible data usage license	**x**	**?**		**x**		
R1.2: (Meta)data are associated with detailed provenance	**x**				**a**	
R1.3: (Meta)data meet domain-relevant community standards	**x**		**a**		**a**	**x**

3. CSIRO 5 stars not yet reviewed by peers

- Thermal and natural mineral waters in Europe : https://egdi.geology.cz/record/basic/610c05bc-3328-4767-8d59-66840a010833
- Gridded estimates of travel times for nitrate in the unsaturated zone in six European countries : https://egdi.geology.cz/record/basic/600ee753-5b98-4805-af06-23100a010833
- WP3-HydroGeoToxicity (HGT): Arsenic and Fluorine : https://egdi.geology.cz/record/basic/60fa8919-d94c-4bb0-a03a-6c6d0a010833
- Basal Reference Concentration (BRC): Arsenic and Fluorine: https://egdi.geology.cz/record/basic/60fa83f1-44bc-4fe0-9744-600b0a010833
- Potential Groundwater Recharge: https://egdi.geology.cz/record/basic/606dbcd8-7ff4-4a85-8b82-24710a010833
- Effective Precipitation bias corrected: https://egdi.geology.cz/record/basic/606dbaf0-3f08-4397-a2d4-1dee0a010833
- TACTIC_WP4_BHs_TimeSeries_Recharge: https://egdi.geology.cz/record/basic/60e5d17a-8c60-4c65-98e6-057f0a010833

		HOVER	TACTIC

	CSIRO 5 Star evaluation not yet reviewed by peers	Thermal and natural mineral waters in Europe	Gridded estimates of travel times for nitrate in the unsaturated zone in six European countries	WP3-HydroGeoToxicity (HGT): Arsenic and Fluorine	Basal Reference Concentration (BRC): Arsenic and Fluorine	Potential Groundwater Recharge	Effective Precipitation bias corrected	TACTIC_WP4_BHs_TimeSeries_Recharge
publication and indexing								
2. Published - is the data accessible to users other than the creator or owner?								
	No							
	By individual arrangement							
	File download			X	X			
	Institutional or community repository							
	Bespoke web service (informal API)							
	Bespoke web service (OpenAPI/Swagger)							
	Standard web service API (e.g. OGC)	x	X			X	X	X
3. Citeable - denoted using a formal identifier								
	Not citeable							
	Local identifier							
	Web address (URL - not guaranteed stable)	X	X	X	X	X	X	X
	Persistent web identifier (URI)							
4. Described - tagged with metadata								

	No metadata							
	Abstract and keywords							
	Basic metadata (e.g. Dublin Core)							
	Specialized metadata (e.g. Darwin Core, ISO 19115/19139, schema.org scientific data profile)							
	Rich metadata using multiple standard RDF vocabularies (e.g. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)	X	X	X	X	X	X	X
5. Findable - indexed in a discovery system								
	no							
	local or internal system only							
	community wide or jurisdictional system							
	highly ranked in general purpose index (Google, Bing etc)	X	X	X	X	X	X	X
linked and useable								
6. Loadable - represented using a common or community-endorsed (i.e. standard) format								
	bespoke format (text, binary)							
	one standard format, denoted by a MIME-type			X	X	X	X	
	multiple standard formats	X	X					X
7. Useable - structured using a discoverable, community-endorsed (standard?) schema or data model								
	no formal schema			X	X	X	X	
	explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar	X	X					X
	community-shared schema or data model , available from a standard location							

8. Comprehensible - supported with unambiguous definitions for all internal elements								
	local field codes or labels							
	labels with full text explanations		X	X	X	X	X	X
	community standard labels (e.g. CF Conventions, UCUM units)							
	some fields linked to externally managed definitions	X						
	all fields linked to standard, externally managed definitions							
9. Linked - to other data and definitions using public identifiers (e.g. URIs)								
	no links	X	X	X	X	X	X	X
	in-bound links from a catalogue or landing-page							
	out-bound links to related data and definitions							
10. Licensed - conditions for re-use are available and clearly expressed								
	no license			X	X			
	license described in text	X	X					
	link to a standard license (e.g. Creative Commons)					X	X	X
maintenance and provenance								
11. Curated - commitment to ensuring the data is available long term								
	once-off dump, no ongoing commitment							
	best effort, project website							
	public or institutional repository (e.g. CKAN, GitHub)	X	X	X	X	X	X	X
	certified repository							

12. Updated - part of a regular data collection program or series, with clear maintenance arrangements and update schedule								
	one-time dataset					X	X	X
	part of series - occasional/irregular update	X	X	X	X			
	part of series - regular scheduled updates							
13. Assessable - accompanied by, or linked to, a data-quality assessment and description of the origin and workflow that produced the data								
	no quality or lineage information			X	X			
	text lineage statement	X	X			X	X	X
	formal provenance trace (e.g. PROV-O)							
14. Trusted - accompanied by, or linked to, information about how the data has been used, by whom, and how many times								
	no information about usage	X	X	X	X	X	X	X
	usage statistics available							
	Clearly endorsed by reputable organization or framework							
Project, organisational, institutional								
15. Complexity of the project								
	low							
	medium							
	high	X	x	X	X	X	X	X
16. Cross-organisational project?								
	1 organisation							
	2-4 organisations							
	X	X	x	X	X	X	X	X
rating		**3,17**	**3,17**	**2,33**	**2,33**	**2,67**	**2,67**	**3,02**

4. EGDI Metadata catalogue

Those elements are a 1st evaluation of EGDI Metadata catalogue that was done when trying to come up with a consensus.

It has to be put in the light of the consensus laid down in section 5 “Evaluating FAIRness” and the need for a holistic approach.

It is important to keep in mind that one shouldn't be rating the EGDI Catalogue alone just like one shouldn't be rating the EGDI platform for FAIRness, rather the data and metadata on their own.

4.1. F1: metadata and data are assigned a globally unique and eternally persistent identifier

Each metadata record has a unique file identifier - typically a UUID. It can be retrieved from the catalogue (example: https://egdi.geology.cz/record/xml/5cf8cda1-e5fc-4b8d-b4fb-49f70a010852).

Limitation identified in section 5 with regards F1 have to be mentioned here

- currently practices are based on ISO 19115:2003 for the reason mentioned regarding INSPIRE Metadata Guidelines V2. There is a need to go beyond this
- when harvested in another system, the uuid could be overridden by another one
- .cz : if the hosting of EGDI metadata catalogue moves to another geological survey this would break the persistence of the metadata record.
- .xml : the ‘identifier’ to the record changes with the requested serialization. For example, it becomes https://egdi.geology.cz/record/basic/5cf8cda1-e5fc-4b8d-b4fb-49f70a010852. What would it become of the record if one asks for a GeoDCAT_AP or DCAT_AP representation of the same Metadata record? There is now easy for an external system to know what to use in a serialization/metadata model neutral way.

Basing Metadata record identifiers on https://data.geoscience.earth and taking into account new approaches (ex: from ISO 19115:2014 or a mode neutral model) will help solve this.

4.2. F2. data are described with rich metadata (defined by R1 below)

The metadata content is compliant with INSPIRE and extended according to the EGDI requirements to provide rich and usable content

4.3. F3. Metadata clearly and explicitly include the identifier of the data they describe

A unique data identifier is mandatory for datasets in the metadata according to INSPIRE. The URI form is recommended.

However:

- when the datasets is an EU one, the URI used is not based on https://data.geoscience.earth/
- whilst EGDI Metadata catalog does clearly show a dataset identifier field in its HTML rendering, I does not seem properly correct to say that the content of this field is actually the identifier of the dataset. It has been seen in the FAIR evaluation spreadsheet (see sections above) that often the identifier is the link to the

project web site. The fact that this has happened on more than one occasion suggests that this is defaulted somehow by of the metadata creation template.

4.4. F4. Metadata and data are registered or indexed in a searchable resource

- The catalogue is publicly accessible as a standardized CSW 2.0.2 ISO AP 1.0 service, so any CSW client may access it (e.g. QGIS)
- Some records are harvested to the European INSPIRE portal: https://inspire-geoportal.ec.europa.eu/download_details.html?view=downloadDetails&resourceId=%2FINSPIRE-16542303-763e-11e4-8b38-52540004b857_20210325-102002%2Fservices%2F1%2FPullResults%2F161-180%2Fdatasets%2F16&expandedSection=metadata

- Some records are harvested to the European data portal: https://data.europa.eu/data/datasets/f2435e00-5e00-1243-b989-52caa6446ca8?locale=en

- The catalogue is experimentally connected to the Google data search console and the resources can be retrieved through this tool: https://datasetsearch.research.google.com/search?query=egdi radon

- EGDI metadata may be queried simply by google search, e.g.: https://www.google.com/search?q=egdi+radon or https://www.google.com/search?q=f2435e00-5e00-1243-b989-52caa6446ca8

4.5. A1. Metadata and data are retrievable by their identifier using a standardized communications protocol

All metadata are accessible with https protocol (GET, POST or SOAP) in this representations

- html form: https://egdi.geology.cz/record/basic/f2435e00-5e00-1243-b989-52caa6446ca8
- ISO 19139 XML: https://egdi.geology.cz/record/xml/5cf8cda1-e5fc-4b8d-b4fb-49f70a010852
- GeoDCAT-AP: https://egdi.geology.cz/csw?service=CSW&request=GetRecordById&id=f2435e00-5e00-1243-b989-52caa6446ca8&outputschema=http://www.w3.org/ns/dcat%23

4.6. A1.2 the protocol allows for an authentication and authorization procedure, where necessary

Basic authentication is allowed, there is also the possibility to access non-public metadata.

4.7. A2. metadata are accessible, even when the data are no longer available

Metadata may be stored in the catalogue also for the deleted resources.

4.8. I1. Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- ISO 19139 XML
- Geo-DCAT
- HTML RDFa
- HTML with Google JSON-LD included

4.9. I2. Metadata and data use vocabularies that follow FAIR principles

These vocabularies are used:

- INSPIRE Registry: https://inspire.ec.europa.eu/registry
- GEMET thesaurus: http://www.eionet.europa.eu/gemet/
- EUROPEAN Country vocab.: https://publications.europa.eu/resource/authority/country
- OPENGIS EPSG: http://www.opengis.net/def/crs/EPSG
- In GeoDCAT-AP format these additional vocabularies are mapped to:
- INSPIRE codelists: http://inspire.ec.europa.eu/metadata-codelist
- FOAF: http://xmlns.com/foaf/0.1
- vcard: http://www.w3.org/2006/vcard
- IANA media types: https://www.iana.org/assignments/media-types

4.10. I3. Metadata and data include qualified references to other metadata and data

These relationships between metadata records may be mapped:

- Parent - children (superset - subset) 1:N
- Service or application operates on some dataset metadata N:M
- Linkage - Dataset is created/derived from some other datasets (M:N)

4.11. R1.1. metadata and data are released with a clear and accessible data usage license

Access and use conditions and Limitations on public access are mandatory metadata elements.

4.12. metadata and data are associated with their provenance

Provenance (lineage) as text or a structured description (mapping sources) is part of the metadata record.

4.13. metadata and data meet domain-relevant community standards

These standards are used:

- ISO 19115/19119/19139
- Open Geospatial Consortium CSW 2.0.2 ISO AP 1.0
- INSPIRE metadata profile: https://inspire.ec.europa.eu/metadata/6541
- GeoDCAT-AP: https://inspire.ec.europa.eu/good-practice/geodcat-ap

5. 'Holistic' Approach exercise

The use cases described below represent reasonably high (but not exactly rated) level of FAIRness allowing efficient communication between a client (human or machine) and the EGDI system. It is assumed that the client has sufficient knowledge about OGC web services and the INSPIRE Geology data model but very few about EGDI.

Those UseCases are an attempt to demonstrate some of the benefits of the holistic approach when working on the FAIRness assessment exercise

5.1. UseCase 1- Exploiting INSPIRE interoperability principles

A client wants to find and retrieve data about Oligocene Faults in Europe. The starting point is the consolidated INSPIRE model. In the Geology schema faults are Spatial Features of type ShearDisplacementStructure, and geological age is stored in the properties called olderNamedAge and youngerNamedAge. In a possible simplified scenario communication between client and server goes like this:

1. client initiates Metadata Keyword search with INSPIRE theme **Geology** and FeatureType **ShearDisplacementStructure**
    a. system: "HIKE fault data base" metadata record returned
2. client issues a WFS GetCapability request based on Resource Locator information
    a. system: WFS Capability document returned
3. client is looking for keyword "ge:**ShearDisplacementStructure**" in the FeatureType list and identifies the "**hike_detail**" layer.
4. client performs a WFS DescribeFeatureType query on "**hike_detail**"
    a. system: responds with schema "**hike_detailType**"
5. client tries to find "ge:**olderNamedAge"** in the list of elements but it is not found.
6. client performs a semantic search in the Eupean Geoscience Registry (EGR) and finds "**old_unit_uri**" as registered synonym for ge:**olderNamedAge** in the hike_detail layer description.
7. client performs a WFS query on the hike_detail layer with a property/value filter using property name = old_unit_uri and
   value = http://inspire.ec.europa.eu/codelist/GeochronologicEraValue/oligocene
    a. sytem: responds with the result FeatureCollection
8. Client receives the XML and processes the result set.

5.2. UseCase 2- Exploiting Linked Data and OGC web services

1. Client performs a semantic search on EGR to find "**ShearDisplacementStructure**".
    a. system responds with the "**hike_deltail**" EGDI layer description
2. Client is looking for a synonym of ge:**olderNamedAge** and finds "**old_unit_uri**" in the list of elements
3. client performs a WFS query on the hike_detail layer with a property / value filter using property name = old_unit_uri and
   value = http://inspire.ec.europa.eu/codelist/GeochronologicEraValue/oligocene
    a. sytem: responds with the result FeatureCollection
4. Client receives the XML and processes the result set.

In both use cases success requires existing and functional links between Metadata and Services. The figure below shows the graph of interconnected elements with links visited by the client.

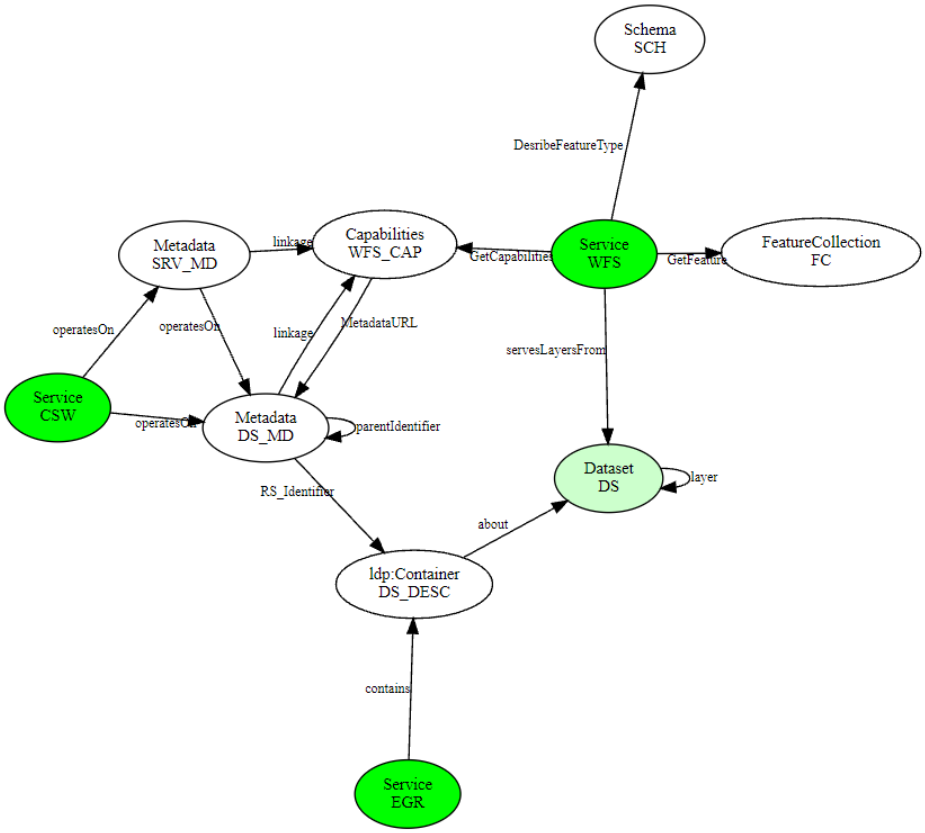


Figure 10: Interconnections between Services and Metadata. Abbreviations: SRV_MD – service metadata, DS_MD – dataset metadata, WFS_CAP – WFS Capability document, DS_DESC dataset description with permanent URI, EGR – European Geoscience Registry

ANNEX B. Geoserver and Pentaho service

This is the main prototype to show end to end mechanism from non-harmonized dataset to harmonized dataset and according dissemination.

1. Workflow

The worklow is detailed in Figure 11: Geoserver and mapping with Pentaho prototype workflow, and consists of the following steps:

- From a gpkg file
- Import it in a user interface
- Once uploaded
- Launch the ETL job(pentaho)
- Collect data through WFS: curl, QGIS, etc.

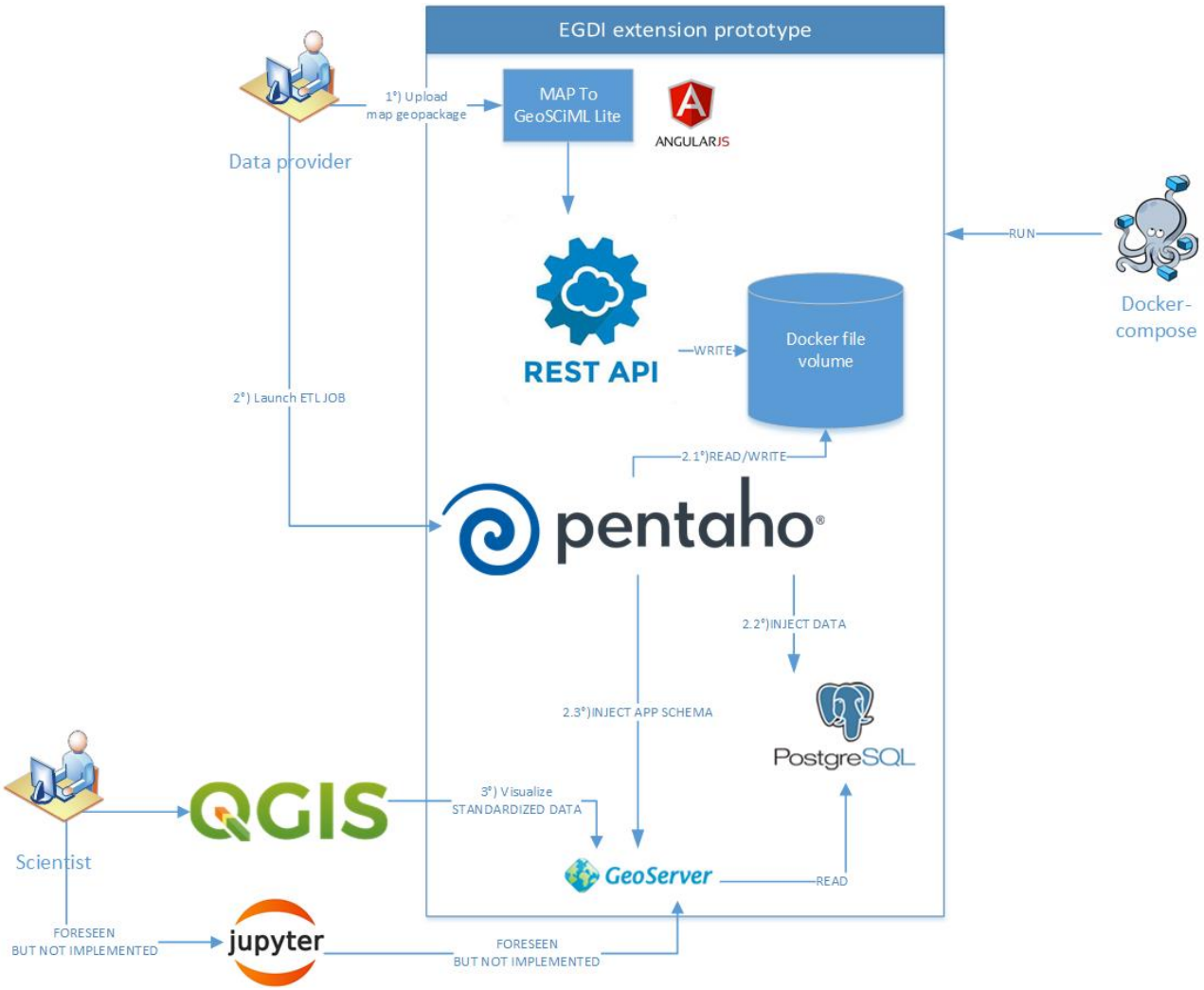


Figure 11: Geoserver and mapping with Pentaho prototype workflow

2. Getting started

For configuring your develop environment, add the following line in the **hosts** file :

```
127.0.0.1    geoera.brgm-dev.fr
```

2.1. Docker-compose

Contains all the configuration files and the dependencies.

To start : in the **root** directory run :

```
docker-compose up -d --build
```

IMPORTANT go to the online pentaho gui and run the *make_servive.kfb* job in the */pentaho_jobs/* directory.

The docker images used are :

- **apache** : used are proxy for angular front and maven back
- **node :** serve the angular front-end code using ng serve
- **maven :** serve the spring back-end code using mvn spring-boot:run
- **posgtreSQL :** with the postGIS plugin
- **geoserver :** with the app-schema plugin
- **pentaho :** launch the job in a script (not a pentaho server)
- **artemis :** JMS Server used to manage the workflow

2.2. back

Contains the spring back-end code

2.3. front

Contains the angular front-end code

2.4. geoserver

Contains the workspaces used by geoserver.

2.5. pentaho

Contains the transformation used by pentaho.

2.6. URLS for the applications :

- http://geoera.brgm-dev.fr : the front-end to upload file
- http://geoera.brgm-dev.fr/back : the back-end to manage the file operations
- http://geoera.brgm-dev.fr/geoserver/ : the geoserver for the WFS

2.7. CURL scenario

List of the curl command to run the scenario.

- upload gpkg
- send mapping to back

```
curl 'http://geoera.brgm-dev.fr/back/geoera/api/etl/' -H 'User-Agent: Mozilla/5.0 (X11;
Linux x86_64; rv:68.0) Gecko/20100101 Firefox/68.0' -H 'Accept: application/json,
```

```
text/plain, */*' -H 'Accept-Language: en-US,en;q=0.5' --compressed -H 'Referer:
http://geoera.brgm-dev.fr/fieldList' -H 'Content-Type: application/json' -H 'DNT: 1' -H
'Connection: keep-alive' -H 'Pragma: no-cache' -H 'Cache-Control: no-cache' --data
'{"filename":"NL-TNO-
v20190710.gpkg","mickaId":"AA132465","fieldMapping":{},"viewName":"ShearDisplacementStr
uctureView"}'
```

- run pentaho
- grab wfs data

ANNEX C. ANNEX: FROST Service

3. Basic information

The HOVER FROST Service is an experimental SensorThings API implementation serving ground water observation data from the EGDI database. The service is not part of the EGDI central system yet. The aim of the pilot was to prove the concept of using the FROST© server for publishing geoscientific information on a high level of FAIRness, to test feasibility and assess the required efforts and resources. FROST provides a REST API that complies with the OGC standard. It is anticipated that such services will play more significant role in the next phase of EGDI development, for example in the CSA project.

More information is available on the GeoERA GIP e-Learning Platform at "Lesson 2 - HOVER Best Practice example - HOVER Data harmonisation example"

The experimental FROST service was developed by MBFSZ based on initial HOVER project data restricted to the area of Hungary and Slovenia.

4. Architecture overview

The FROST server's REST API is built on top of PostgreSQL. Data upload, search and download are possible through http using json structures and the SensorThing API query language.

Figure 8: Overview HOVER Frost Service component and data flow

The server itself is available in a Docker container installed at BRGM. In the current setup it is a standalone service, but in future implementations it can be directly connected to the EGDI database.

5. Where the source is stored

FROST is an Open Source project. Source is available on Github at:

https://github.com/FraunhoferIOSB/FROST-Server

To populate the HOVER FROST pilot dataset a python program was developed. It is located on the Gitlab server at GEUS to the project partners only (for confidentiality and security concerns):

https://geusgitlab.geus.dk/egdi/frost

6. How to build the source

There is no need to build, the code can be run on any standard python interpreter.

7. The services it depends on

Currently there is no service it depends on. In future implementations integrated FROST services must read from EGDI PostgreSQL databases in order to work with the other components.

8. The services it provides

The HOVER FROST Pilot site is available at:

https://geoera.brgm-dev.fr/FROST-Server/v1.1/

Example query to find geothermal wells within 2 Km:

https://geoera.brgm-dev.fr/FROST-Server/v1.1/Locations?$count=true&$filter=geo.distance(location,geography'POINT(4924000 2581000)') lt 2000&$select=name&$expand=Things

Figure 9: FROST server response with an example result set

9. The log files (where they are)

Log files are stored at their standard location in the FROST Docker container.

BIBLIOGRAPHY

Interoperability design and implementation choices reference: https://www.ai4eu.eu/sites/default/files/inline-files/Deliverable_AI4EU_D2.7_M8_vfinal.pdf

GIP-P Standard validation procedures: https://geoera.eu/wp-content/uploads/2020/01/D3.3-Standards-validation-procedures.pdf

GIP-P D5.1 blue print on data and services: https://geoera.eu/wp-content/uploads/2019/10/D5.1.v1-GIP_blueprint-Data_and_services_architecture.pdfGIP-P D5.2 Central-System-Specification https://geoera.eu/wp-content/uploads/2019/12/D5.2.v1-GeoERA-Central-System-Specification.pdf

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

Jupyter notebook as VRE examples : https://www.eosc-hub.eu/training-material/egi-jupyter-notebooks-examples

Autoscaling Kubernetes: https://medium.com/nerd-for-tech/autoscaling-in-kubernetes-hpa-vpa-ab61a2177950

SENSORTHING API: https://en.wikipedia.org/wiki/SensorThings_API

Horizon Europe calls: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-infra-2021-tech-01-01