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

Authors and affiliation:

BGS	**E. Lewis**
	J. Passmore
ISPRA	**C. Cipolloni**
	E.Giulianelli
BRGM	**M. Beaufils**
	S. Grellet
CGS	**L. Kondrová**
	O. Moravcová

E-mail of lead authors:
edlew@bgs.ac.uk

Deliverable 3.1

Version: 02/05/2019

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

Deliverable Data		
Deliverable number	D3.1	
Dissemination level	Public	
Deliverable name	Data models, Standard Guidelines and Toolkits	
Work package	WP3, Standard and Interoperability Issues	
Lead WP/Deliverable beneficiary	Carlo Cipolloni (ISPRA)	
Deliverable status		
Submitted (Author(s))	02/05/2019	Edd Lewis, James Passmore, Carlo Cipolloni, Elio Giulianelli, Mikael Beaufils, Sylvain Grellet, Lucie Kondrová, Olga Moravcová.
Verified (WP leader)	21/05/2019	Carlo Cipolloni
Approved (Coordinator)	22/05/2019	Jorgen Tulstrup

GENERAL INTRODUCTION

Deliverable 3.1 – Data models standards guidelines and toolkits

An important principle for the GeoERA Information Platform project is to build on interoperable standards. To achieve this, this document contains an analysis on the existing reference standards that can be used in the different science projects and that can be implemented in EGDI.
It also provides toolkits, guidelines and examples for datasets and metadata harmonisation.

Executive Summary

There is a range of international standards applicable to most of the data produced by the GeoERA science projects, and where no standards exist there are examples of best practice from other domains which could be applied.

Key matrices of data types, available standards and existing enabling technologies can be found in tables ES1 & ES2.

ES1. Data type and applicable standard matrix

Standards	Data Type					
	Point/Line/Polygon	Grids	Volumes	Time Series	Downhole Logs	Unstructured Data
Catalogue Service	Y	Y	Y	Y	Y	Y
SOS				Y	Y	
SensorThings				Y	Y	
WCPS		Y	Y	Y	Y	
WCS		Y	Y	Y	Y	
WFS	Y			Y	Y	
WFS-T	Y				Y	
WMS	Y	Y			Y	
WMTS	Y	Y				
3D on the Web common Practice			Y			

ES2. Standards and supporting existing technologies/tools

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	

One of the main activities is the identification of the formats and structures of the data (and related metadata) needed to be distributed by the EGDI. Key requirements are:

- To identify the datasets to be supplied by the application;
- For each of these datasets, to identify shared data models (ideally INSPIRE or OGC data models), and if it is not using a data model, to schedule an extension of one of the data models involved that may better meet the EGDI data requirements.

These actions must be performed both for the input data and for the output data.

Summary of recommendations:

1. For each identified standard, the repository website of the standard should be examined in order to ensure the alignment with the current version.
2. Thorough review of the technical specifications of the selected data models must be performed to better understand the data structures and the correct understanding of the implementation described to correctly map their data in the selected models.
3. The extensions to the models should be limited as much as possible, especially those extensions which impact on the data structure. The mapping of the data into the data models available for the needed data should be carefully checked.
4. In case of data model extension, the reference thematic community (e.g. INSPIRE Community Forum, OGC Standards and Domain Working Groups) should be consulted to hear if similar extensions have been already proposed/discussed/solved.
5. Mapping rules should be defined before data transformation is performed (a common error in data harmonization is to start from scratch to use programs that perform transformation without first defining such rules).
6. When mapping data between models use an intermediate matching table and fill it in as detailed as possible to provide a better understanding of the overall harmonization, and save time in the last phase of the mapping implementation using the selected software tool.
7. When transforming a source schema to the INSPIRE target scheme, information from the source schema should be mapped to the destination schema.
8. It is recommended to complete as many metadata fields as possible, not only the mandatory ones.
9. It is recommended to use existing shared codes/vocabularies and a thesauri, and to avoid the use of free text keywords as much as possible.

Deliverable 3.2 will aim to identify gaps between the anticipated science project outputs and existing available data standards/best practices as described in this report.

TABLE OF CONTENTS

TABLE OF FIGURES

1 DATA STANDARDS

1.1 Existing Standards

1.1.1 Metadata

A description of the proposed architecture can be found in D5.2, section 1.1 which is based on the EGDI Metadata template.

Applicable standards are:

- ISO 19115 Geographic Information: Datasets,
- ISO 19119 Geographic Information: Services,
- ISO/TS 19139:2007 Geographic Metadata XML (gmd & srv) encoding, an XML Schema implementation derived from ISO 19115 & 19119,
- ISO 19110 Geographic Information: Methodology for feature cataloguing,
- ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set,

The metadata must by compliant with the latest INSPIRE Metadata implementation rules.

A common vocabulary of geological terms should be used, it is recommended to utilise the existing CGI Vocabularies (http://resource.geosciml.org/def/voc/) and INSPIRE codelists where possible. Furthermore, WP4 supports the GeoERA projects in delivering their own project vocabularies. If needed, EGDI (on behalf of EGS as a registered INSPIRE roof organization) may also propose official extensions to the INSPIRE codelists and potentially also to the CGI vocabularies where it is required by the science projects.

1.1.2 Data Models

1.1.2.1 Data modeling approach for the GIP

There are a number of existing data models accepted as international standards which could be applied to the GeoERA IP.

To be in line with European regulation on INfrastructure for SPatial Information in Europe (**INSPIRE - 2007/2/CE)** and not to reinvent the existing internationally approved standards, we want to push the GeoERA scientific projects to use the INSPIRE and OGC models. Complementing this, one has to add that EPOS Thematic Core Services on Geological Information and Modeling –TCS GIM) already shares that philosophy and enhances the data models available for GeoSciences.

For now, most of this knowledge representation is formalized in UML models (conceptual, logical models) building on the ISO 191xx series of standards.

Given the target of the GIP, and the mechanics implied, it is deemed reasonable to continue that approach for the GIP data models activities. This reasoning is already detailed in D5.1

section 4: “REACHING THE TARGET - USE OF DATA STANDARDS “ and D5.2 section 1.6: “Shared data specifications environment”.

Indeed there are trends to better specify (and exchange) datasets and some GIP-P partners are actively involved in this (see D5.1 and D5.2 for more details).
However, web semantics standards are considered mature enough for 2 areas of interest for the GIP-P:

- Datasets and services metadata according to DCAT and EPOS ICS-C requirements (EPOS_DCAT_AP)
- Vocabulary exchange: in the sense of CodeList/Thesaurus and not in the sense of setting up new ontologies to model a complete domain. SKOS, RDF, RDFS, PROV, VANN and many other models can be involved in this.
  This CodeList exchange is important to be driven properly as it fits perfectly within the INSPIRE register federation dynamic. In case the GIP-P produces project specific code lists or vocabularies these shall be registered under a europe-geology.eu subdomain as it must be under the control of EuroGeoSurveys after the end of GeoERA.
  CodeList/Thesaurus population should be done in coordination with WP4 as this is the goal of this WP. It should also be populated keeping in mind the existing CGI Vocabularies (http://resource.geosciml.org/def/voc/) where possible. CGI vocabularies can be extended where required by the science projects.

As a result, all data modelling activity within the GIP-P shall use UML models building on the ISO 191xx series of standards.
As summarized in D5.1 section 4, complementing this approach and when deemed useful for reuse at the GeoERA Information Platform Level (or by any external system), an ontology will be generated building on the discussion held at the level of OGC.

1.1.2.2 Simple Knowledge Organization System (SKOS)

“The Simple Knowledge Organization System” is a common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies. Using SKOS, a knowledge organization system can be expressed as machine-readable data. It can then be exchanged between computer applications and published in a machine-readable format on the Web.” (https://www.w3.org/TR/skos-reference/#L895)

1.1.2.3 RDF

RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. (https://www.w3.org/RDF/)

1.1.2.4 DCAT

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. This document defines the schema and provides examples for its use.

By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation. (https://www.w3.org/TR/vocab-dcat/)

Use of DCAT is not recommended for the GIP where ISO19XXX is appropriate.

1.1.2.5 INSPIRE themes model

The INSPIRE Directive has defined with Commission Regulation (EU) No 1312/2014 of 10 December 2014 amending Regulation (EU) No 1089/2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data services, the 34 Data themes that are in the scope of INSPIRE. For each of those, a Data Specification Technical guideline has been defined. The aim for INSPIRE is to ensure interoperability to provide the possibility to combine spatial data and services from different sources across the European Community.
The Data themes are subdivided into three annexes, taking in consideration the main scope of these datasets, and the list is available in the INSPIRE Data Specification page (https://inspire.ec.europa.eu/data-specifications/2892), in which it is possible to navigate through each data model in order to identify the right scope.
The data model can be navigated in a specific section (https://inspire.ec.europa.eu/data-model/approved/r4618/html/), while the schema repository for all the INSPIRE data models is available at https://inspire.ec.europa.eu/schemas/.

1.1.2.6 GeoSciML v 4.1

GeoSciML v 4.1 was published in Jan 2017 and is the result of a collaborative project to develop a GML based exchange language and data model by the IUGS Commission for the Management and Application of Geoscience Information (CGI).

“The core purpose of GeoSciML remains largely unchanged, covering the representation of geologic units, earth materials and geologic structures. Geologic structures include shear displacement structures (brittle faults and ductile shears), contacts, folds, foliations, lineation’s and structures with no preferred orientation (e.g. ‘miarolitic cavities’). The Earth Material package allows for the description of compound materials, such as rocks or unconsolidated materials, as well as their individual components, such as minerals, and includes the relationships between the components. Provision is made for description of alteration, weathering, metamorphism, particle geometry, fabric, and petrophysical data. Mapped features describe the shape of the geological features using standard GML geometries, such as polygons, lines, points or 3D volumes. Geological events provide the age, process and environment of formation of geological features. Geological sampling, logs, and observations from boreholes and outcrops can also be delivered using the GeoSciML

extension of the OGC standard for Observations and Measurements (O&M).” (O Raymond, B Simmons, E Boisvert, 2010. Information Models for the Australian Geoscience Community: GeoSciML, EarthResourceML and GroundWaterML)

An example of use can be found at http://onegeology.brgm-rec.fr/mapClient/

Full documentation is available at https://www.opengeospatial.org/standards/geosciml

1.1.2.7 Groundwater ML2 (GWML2)

GroundwaterML2 (GWML2) was developed under the auspices of the OGC Hydro Domain Working Group to facilitate exchange of groundwater data and is and extension of two existing GML standards, O&M and GeoSciML. GWML is an OGC standard and can be used with OGC web services such as WFS, SOS or WCS to share groundwater data.

A compilation of available ground water endpoints (most of them based on GWML2) was set up in order for World Meteorological Organization Commission for Hydrology (WMO CHy) to begin GWML2 and related standards testing phase. The compilation is available here: https://external.opengeospatial.org/twiki_public/HydrologyDWG/GINsForWMOCHy. Some examples of data consumption by desktop and web client are also provided
Now the standard is in the right tracks for adoption at WMO level by December 2019 to be used within WMO's Hydrological Observing System (WHOS).

Full documentation is available at https://www.opengeospatial.org/standards/gwml2

1.1.2.8 EarthResourceML

“EarthResourceML is an XML-based data transfer standard for the exchange of digital information for mineral occurrences, mines and mining activity. The model describes the geological features of mineral occurrences, their commodities, mineral resources and reserves. It is also able to describe mines and mining activities, and the production of concentrates, refined products, and waste materials.” (http://www.earthresourceml.org/) ERML uses ISO and OGC data standards, including GML v3.2, SWE Common v2, and GeoSciML v3.2

Full documentation can be found at http://www.earthresourceml.org/earthresourceml/2.0/doc/ERML_HTML_Documentation/

EarthResourceML 2.0 is the preferred standard for mineral resource data sharing initiatives and projects, such as

- European Union’s INSPIRE directive, EURare, Minerals4EU, and ProSUM projects,
- The Australian AuScope, and Geoscience Portal projects.

After 2015 small modifications, the full INSPIRE Mineral Resource model and CGI EarthResourceML models are identical.

There is also EarthResourceML Lite v2.0.1 released in October 2018.

“EarthResourceML-Lite is a model and schema for simple map services (e.g., WMS and WFS Simple Features). It is an abridged version of the full EarthResourceML model and can be

used to deliver simplified views on mineral occurrences and their commodities, mines, mining activities and mine waste products.

The v2.0.1 release corrects minor omissions in the CommodityResourceView and MiningWasteView schemas but is otherwise compatible with v2.0.0." (http://www.earthresourceml.org/)
Full documentation can be found at http://www.earthresourceml.org/earthresourceml-lite/2.0.1/documentation

1.1.2.9 Observations and Measurements

Observations & Measurements (ISO 19156 / OGC 10-004r3)

"This standard specifies an XML implementation for the OGC and ISO Observations and Measurements (O&M) conceptual model (OGC Observations and Measurements v2.0 also published as ISO/DIS 19156), including a schema for Sampling Features. This encoding is an essential dependency for the OGC Sensor Observation Service (SOS) Interface Standard. More specifically, this standard defines XML schemas for observations, and for features involved in sampling when making observations. These provide document models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities."

Numerous domain standards and services standards are built on O&M. It also triggers sibling activities within W3C (ex SSN/SOSA).

At INSPIRE level a specific guidance document as produced on "the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE" (INSPIRE D2.9: https://inspire.ec.europa.eu/id/document/tg/d2.9-o%26m-swe).

Full documentation available at: https://www.opengeospatial.org/standards/om

1.1.3 Data Services

Open Geospatial Consortium (OGC) web services offer a cost efficient and open source technology that permits transfer of standardized data from distributed sources, removing the need for data to be regularly uploaded to a centralized database. When combined with community defined exchange standards or schemas, OGC services offer the ability to access the latest data from source agencies in a consistent format." (O Raymond, B Simmons, E Boisvert, 2010. Information Models for the Australian Geoscience Community: GeoSciML, EarthResourceML and GroundWaterML)

1.1.3.1 SPARQL Endpoint

"SPARQL is an RDF query language—that is, a semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is recognized as one of the key technologies of the semantic web. On

15 January 2008, SPARQL 1.0 became an official W3C Recommendation, and SPARQL 1.1 in March, 2013." (https://en.wikipedia.org/wiki/SPARQL)

"A SPARQL Endpoint is a Point of Presence on an HTTP network that's capable of receiving and processing SPARQL Protocol requests. It is identified by a URL commonly referred to as a SPARQL Endpoint URL." (https://medium.com/virtuoso-blog/what-is-a-sparql-endpoint-and-why-is-it-important-b3c9e6a20a8b)

A SPARQL endpoint could be implemented for all GeoERA projects/applications.

1.1.3.2 X3D

X3D is an ISO-ratified, royalty-free open standards file format and run-time architecture to represent and communicate 3D scenes and objects. X3D has evolved from its beginnings as the Virtual Reality Modeling Language (VRML) to the considerably more mature and refined ISO X3D standard. X3D provides a system for the storage, retrieval and playback of real time 3D scenes in multiple applications, all within an open architecture to support a wide array of domains and user scenarios.

As an open standard X3D can run on many platforms, but importantly can render 3D models in most web browsers without the requirement for additional or proprietary applications. Further, once models are developed utilizing X3D, these easily port to alternative platforms like holographic, head-mounted or other display devices.
(http://www.web3d.org/x3d/what-x3d)

1.1.3.3 Catalogue Service (Catalogue Service for the Web – CSW)

"Catalogue services support the ability to publish and search collections of descriptive information (metadata) for data, services, and related information objects. Metadata in catalogues represent resource characteristics that can be queried and presented for evaluation and further processing by both humans and software. Catalogue services are required to support the discovery and binding to registered information resources within an information community.

OGC Catalogue interface standards specify the interfaces, bindings, and a framework for defining application profiles required to publish and access digital catalogues of metadata for geospatial data, services, and related resource information. Metadata act as generalized properties that can be queried and returned through catalogue services for resource evaluation and, in many cases, invocation or retrieval of the referenced resource. Catalogue services support the use of one of several identified query languages to find and return results using well-known content models (metadata schemas) and encodings."

Full documentation available at: https://www.opengeospatial.org/standards/cat

1.1.3.4 Sensor Observation Service (SOS)

"The SOS standard is applicable to use cases in which sensor data needs to be managed in an interoperable way. This standard defines a Web service interface which allows querying observations, sensor metadata, as well as representations of observed features. Further, this

standard defines means to register new sensors and to remove existing ones. Also, it defines operations to insert new sensor observations. This standard defines this functionality in a binding independent way; two bindings are specified in this document: a KVP binding and a SOAP binding."

Full documentation available at: https://www.opengeospatial.org/standards/sos

1.1.3.5 SensorThings API

"The OGC SensorThings API provides an open, geospatial-enabled and unified way to interconnect the Internet of Things (IoT) devices, data, and applications over the Web. At a high level the OGC SensorThings API provides two main functionalities and each function is handled by a part. The two parts are the Sensing part and the Tasking part. The Sensing part provides a standard way to manage and retrieve observations and metadata from heterogeneous IoT sensor systems. The Tasking part is planned as a future work activity and will be defined in a separate document as the Part II of the SensorThings API."

It is built on O&M and, as summarized in D5.2 section "1.3 Observation features OGC services strategy", is often wrongly reduced to the REST binding of SOS 2.0.

Full documentation available at: https://www.opengeospatial.org/standards/sensorthings

1.1.3.6 Web Feature Service (WFS)

The Web Feature Service (WFS) represents a change in the way geographic information is created, modified and exchanged on the Internet. Rather than sharing geographic information at the file level using File Transfer Protocol (FTP), for example, the WFS offers direct fine-grained access to geographic information at the feature and feature property level.

This International Standard specifies discovery operations, query operations, locking operations, transaction operations and operations to manage stored, parameterized query expressions.

Discovery operations allow the service to be interrogated to determine its capabilities and to retrieve the application schema that defines the feature types that the service offers.

Query operations allow features or values of feature properties to be retrieved from the underlying data store based upon constraints, defined by the client, on feature properties.

Locking operations allow exclusive access to features for the purpose of modifying or deleting features.

Transaction operations allow features to be created, changed, replaced and deleted from the underlying data store.

Stored query operations allow clients to create, drop, list and described parameterized query expressions that are stored by the server and can be repeatedly invoked using different parameter values.

This International Standard defines eleven operations:

- GetCapabilities (discovery operation)
- DescribeFeatureType (discovery operation)
- GetPropertyValue (query operation)
- GetFeature (query operation)
- GetFeatureWithLock (query & locking operation)
- LockFeature (locking operation)
- Transaction (transaction operation)
- CreateStoredQuery (stored query operation)
- DropStoredQuery (stored query operation)
- ListStoredQueries (stored query operation)
- DescribeStoredQueries (stored query operation)

In the taxonomy of services defined in ISO 19119, the WFS is primarily a feature access service but also includes elements of a feature type service, a coordinate conversion/transformation service and geographic format conversion service.

D5.2 in its section 1.2 and 1.13 clarifies the position to adopt regarding WFS 3.0.

Full documentation available at: https://www.opengeospatial.org/standards/wfs

1.1.3.7 Web Map Service (WMS)

The OpenGIS® Web Map Service Interface Standard (WMS) provides a simple HTTP interface for requesting geo-registered map images from one or more distributed geospatial databases. A WMS request defines the geographic layer(s) and area of interest to be processed. The response to the request is one or more geo-registered map images (returned as JPEG, PNG, Vector tiles, etc) that can be displayed in a browser application. The interface also supports the ability to specify whether the returned images should be transparent so that layers from multiple servers can be combined or not.

Full documentation available at: https://www.opengeospatial.org/standards/wms

1.1.3.8 Web Map Tile Service (WMTS)

A WMTS enabled server application can serve map tiles of spatially referenced data using tile images with predefined content, extent, and resolution.

Full documentation available at: https://www.opengeospatial.org/standards/wmts

1.1.3.9 Web Coverage Processing Service (WCPS)

"The OGC® Web Coverage Processing Service (WCPS) defines a protocol-independent language for the extraction, processing, and analysis of multi-dimensional coverages representing sensor, image, or statistics data."

Full documentation available at: https://www.opengeospatial.org/standards/wcps

1.1.3.10 Web Coverage Service (WCS)

"A Web Coverage Service (WCS) offers multi-dimensional coverage data for access over the Internet. WCS Core specifies a core set of requirements that a WCS implementation must fulfill.

More information on spatio temporal coverage / datacube standards can be found at http://myogc.org/go/coveragesDWG, including tutorials and webinars, conformance testing, background information, and updates on standardization progress."

Full documentation available at: https://www.opengeospatial.org/standards/wcs

2 DELIVERED DATA FROM THE GEOERA SCIENCE PROJECTS

Please see deliverable D2.2.1 for full description of anticipated data produced by the science projects which require delivery via the GIP-P.

The data to be delivered via the GIP-P can be categorised into two types:

1. Structured Data
    a. Mostly possible to use open standards, see Table 1
    b. Tools to deliver this data are outlined in Table 2
    c. 3D models no directly applicable international standards specifically for the geoscience domain however X3D is the cross domain ISO-ratified standard.
2. Unstructured data
    a. These are documents, PDF's, photos, videos, scanned downhole logs.
    b. Would commonly be held in a central repository. WP5 will identify the best architecture to ensure that these outputs are findable and accessible.
    c. Although no standards can be applied to the data itself, the metadata of these records should conform to international ISO standards.

Table 1. Data type and applicable standard matrix

Standards	Data Type					
	Point/Line/Polygon	Grids	Volumes	Time Series	Downhole Logs	Unstructured Data
Catalogue Service	Y	Y	Y	Y	Y	Y
SOS				Y	Y	
SensorThings				Y	Y	
WCPS		Y	Y	Y	Y	
WCS		Y	Y	Y	Y	
WFS	Y			Y	Y	
WFS-T	Y				Y	
WMS	Y	Y			Y	
WMTS	Y	Y				

Table 2. Standards and supporting existing technologies/tools

Standards	Existing Technologies					
	MICKA / GeoNetwork	52°North Sensor Observation Service	Fraunhofer IOSB Frost-Server	Geoserver / ArcServer / MapServer / TinyOWS	Rasdaman / Petascope	Epimorphics UKGovLD - Registry Core
Catalogue Service	Y					
Semantic web for codeList						Y
SOS		Y		Y		
SensorThing			Y			
WCPS					Y	
WCS				Y	Y	
WFS				Y		
WFS-T				Y		
WMS				Y	Y	
WMTS				Y		

WPS				Y		

2.1 Data requirements

Based on the deliverable D.2.1.1 "Requirements to the GIP-project by the three other themes", we have mapped the GeoERA projects user requirements analyzing the input and output data handled. The granularity of this information is related to the level of detail provided by the projects themselves. In some case we have identified an applicabledata structure, and the appropriate data model and or application schema was easily manageable. In other cases the information provided by a project is quite generic and we have only been able to identify the main data type.
Table 3 presents an overview of data types identified as project input or output is presented together with the data model that has been identified to enable a mapping of the attribute in harmonized way. A more accurate analysis on data mapping will be done by the task 3.2 "Data model gap analysis and technical requirements".

Table 3 – project data type mapping with standard data models.

Project Name	DATA TYPE	DATA STANDARD
(GW) - RESOURCE project	Groundwater composition and age	*GroundWaterML2 O&M*
	Cross-border patterns of groundwater depletion and recharge	*GroundWaterML2 + O&M*
	Geological and hydrological data specifying delimitation of aquifer (thickness, depth, groundwater flow directions and flux)	*GroundWaterML2 + O&M*
	Hydrological dataset for numerical modelling	*OGC GeoScience DWG 3D Model + EPOS TCS GIM ModelView*
	Multiple layers of information specifying lithology, depth and extent of Aquifers and Aquitards	*GroundWaterML2 + GeoSciML*
(GW) - VOGERA project	Fault zones	*GeoSciML (GeologicalStructure)*
	Boreholes	*GeoSciML (Borehole) + EPOS Borehole model*
	Physico-chemical data: stable isotopes, time indicators temperature	*GeoSciML + O&M*
	Geophysical methods	*INSPIRE GE*
	3D models	*OGC GeoScience DWG 3D Model (future) + EPOS TCS GIM ModelView. X3D*
	Model of vulnerability maps	*GroundWaterML2*
(GW) - HOVER project	Database for concentrations of dissolved elements and associated parameters to define thermal and mineral water	*GroundWaterML2 + O&M*
	Boreholes	*GeoSciML (Borehole)*

	Database for concentration of elements of natural origin per typologies	*GroundWaterML2 + O&M*
	European exposure maps of selected elements (and indicators)	*INSPIRE AM*
	Atlas of geological/hydrogeological settings vulnerability maps	*ISO19115*
	Maps of groundwater-N travel time	*GroundWaterML2 + O&M*
	Database for concentration of groundwater age indicators and vulnerability classes	*GroundWaterML2 + O&M*
	Maps and cross sections showing distribution of groundwater age and vulnerability classes in selected European aquifers.	*GroundWaterML2 + GeoSciML*
	Maps and cross sections showing vulnerability of the upper aquifer to pollution.	*GroundWaterML2*
(GW) - TACTIC project	Hydrogeological parameters: (e.g. porosity, hydraulic conductivity, cation exchange capacity)	*GroundWaterML2 + O&M*
	Hydrogeological time series: Water tables; Head/River/concentrations; Rainfall, temperature, potential evaporation; real time data.	*GroundWaterML2 + WaterML2 + INSPIRE AC*
	Borehole, hydrochemical and geophysical logs (data and time series)	*GeoSciML (Borehole) + INSPIRE GE + OGC O&M + GWML2 (logs) + EPOS TCS GIM Boreholes*
	Soil maps/soil properties: Land use; Specific model outputs (e.g. min, max, mean heads, or changes); Climate grid; Satellite	*INSPIRE SO + INSPIRE LU + INSPIRE AC + EF + O&M*
	3D data: Hydrogeological model – structures; Hydrogeological parameters; Model outputs	*OGC GeoScience DWG 3D Model + EPOS TCS GIM ModelView*
(RM) - EuroLithos project	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Location of current and relevant old/historic ornamental stones mining districts, mining sites or isolated quarries	*INSPIRE MR/ EarthResourceML*
	Land use planning constraints and threats	*INSPIRE LU + INSPIRE AM*
(RM) - FRAME project	Mineralisations and deposits on land and the marine environment (Data from Minerals4EU, ProMine and OneGeology Europe exists as WMS/WFS)	EarthResourceML

	Secondary resources (mining waste) - *Data from ProSUM Mining Waste (to be delivered) will be provided as WMS/WFS*	EarthResourceML
(RM) - MINDeSEA project	Marine geology (will reuse part of EMODnet, ISA, Interridge programs and Geo-Seas)	*GeoSciML*
	All other Marine information about SMS, Placers, Nodules (will reuse part of EMODnet, ISA, Interridge programs and Geo-Seas)	*INSPIRE EF + INSPIRE OF + OGC O&M + INSPIRE MR /EarthResourceML*
(RM) - Mintell4EU project	Data based on Mineral4EU, i.e. mineral occurrences, mines and stastistical data on country level.	EarthResourceML
(RM) - GARAH project	Boreholes, wells, outlines of formations, basin outlines, horizon interpretations,	*GeoSciML (Borehole) + EPOS TCS GIM Boreholes*
	Faults	*GeoSciML (GeologicalStructure)*
	Temperature maps	*GeoThermal + INSPIRE ER*
	Bathymetry	*INSPIRE EL*
	Geothermal gradients, seafloor temperature, seafloor T heat flow	*GeoThermal + INSPIRE ER*
	Sedimentation rates in 4D	*3D model/X3D*
	Fishing activities	*INSPIRE PF*
	Gas hydrates below seafloor, gas stability map	*INSPIRE ER*
(GE) - Geoconnect³d project	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Faults, Fault systems	*GeoSciML (GeologicalStructure)*
	Chemical analyses of springs water	*WaterML + INSPIRE EF + OGC O&M*
	Wells measurements	*GroundWaterML2*
(GE) - HIKE project	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Faults	*GeoSciML (GeologicalStructure)*
(GE) - 3D geomodeling	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Wells observations	*GroundWaterML2 + OGC O&M + INSPIRE EF + OGC O&M*
	Reservoir proprieties	*GroundWaterML2*
	2.5D Time model (xyz): 2.5D Time model (xyz) 2.5D Velocity maps (xyz) 3D Structural model 3D Harmonized model of lithostratigraphic layers	*3D model (X3D) + OGC O&M + Metadata + EPOS TCS GIM ModelView*

	Geothermal properties related to wells (porosity & permeability) + 2D Geothermal property maps Example datasets and models containing uncertainty information 2D Maps of Cenozoic reservoirs (extent + depth) 2D Map of extent & depth of salt/fresh groundwater barrier Uncertainty in geomodels Metadata	
(GE) - Muse project	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Conflicting layers	
	Wells, borehole observations, sample measurements	*GeoSciML (Borehole)(EPOS-GIM/OGC Geoscience model + GroundWaterML2 + OGC O&M*
(GE) - Hotlime project	Geology (geologic unit)	*GeoSciML (GeologicalUnit)*
	Conflicting layers	
	Wells, boreholes observations, samples measurements	*GeoSciML (Borehole) + GroundWaterML2 + OGC O&M + EPOS TCS GIM Boreholes*
	Faults	*GeoSciML (GeologicalStructure)*

3 DATA PROCESSING

3.1 Data and Metadata Harmonization Process

In this section, the general approach for data and metadata harmonisation is described.
In Figure 1 is the schema of the overall data harmonization process, consisting of an initial phase in which the source dataset and its associated data model are analysed. After we have identified and selected the appropriate target schema that best fits the purpose with regard to the source dataset and the objective of the transformation, the corresponding data specification is thoroughly analysed:

- For the target data models corresponding to the INSPIRE or OGC data Models (or the ISO 191xx series), the descriptive version of the data specification and its UML representation are available on the INSPIRE, OGC or other reference website;

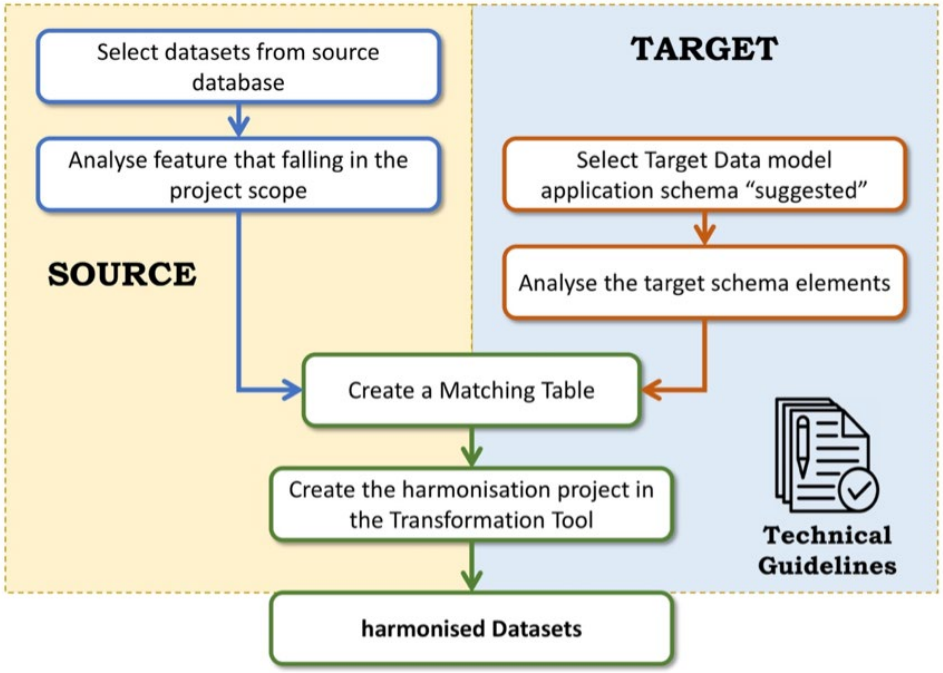

Figure 1 Data harmonization overall process

After this phase and before using any transformation tool, the most crucial harmonization step consists of filling-in the matching table (which is presented in the Github repository: https://github.com/GeoEra-GIP) to map the data in right way. This should be carried out alongside with the generation of an example XML instance file for the considered features. This can be done hand-made using XML aware tools such as XMLSpy or Oxygen and helps identify the gaps by actually manipulating real data examples.

Performing this exercise and analysing and solving the eventual matching problems, any GeoERA project participant will be facilitated to carry out the transformation step. The transformation can be done using a dedicated software tool such as Hale but could also be done manually.
After the creation of the harmonised dataset, it has to be validated using tools and procedure described by deliverable 3.3 and then published as a network service, following the standard presented in this document and procedure that will be described in the GIP WP8.

With regard to the metadata harmonization process, the basic principles for providing metadata for the EGDI platform are stated in the INSPIRE Metadata Regulation on https://inspire.ec.europa.eu/Legislation/Metadata/6541 and described in detail in Technical Guidelines for implementing the Metadata regulation on https://inspire.ec.europa.eu/Technical-Guidelines2/Metadata/6541.
As an addition to the original INSPIRE regulation, it is strongly recommended to create bilingual metadata records (English as the primary language, national language as the secondary) to provide a proper English translation of the geoscientific language (as opposed to using an automatic translation service), and to raise the discoverability of the scientific outputs of the projects. As the minimum, the title, abstract and keywords should be filled in in both languages.
To enable effective filtering, the use of predefined keywords for results of projects is required (and specific GIP-P multilingual keywords developed in WP4 will be used to identify results of the GeoERA projects). The names of data providers, projects and countries have to be harmonized as well.
To describe the logical relationships between data resources and to facilitate the user experience throughout the EGDI platform (portal and catalogue), it is required to relate metadata of datasets and the web services that are using them.
Requirements for additional metadata elements might come from the proposed functionality of the whole EGDI adopted by the GIP-P (i.e. results of the work of WP5 and WP7) and from the envisaged outputs of GeoERA projects (i.e. 3D models) and will be analyzed subsequently.

Full instructions and support will be provided by WP8. This will include cookbooks to follow, training facilities and individual support.

3.2 Tools

3.2.1 Geospatial Metadata Creation/Management/Distribution

The EGDI metadata catalogue (https://egdi.geology.cz/), which is used as a base for future developments through the GIP-P project, is based on MIcKA software (currently version 5, version 6 is tested within the frame of WP7).
The EGDI metadata profile is compliant with the requirements of the INSPIRE Directive for metadata and the EN ISO 19115/19119 standards. Only digital and structured information (spatial datasets or dataset series, spatial data services - WMS, WFS and web applications) is described by metadata in this catalogue. MIcKA provides tools for compilation of those metadata in a standardized format. Functions of transactions and harvesting are also supported. In addition to basic CSW functionality, the GeoDCAT-AP, KML, ATOM, OAI-PMH and other outputs are currently available. Although it is possible to create metadata directly in the EGDI catalogue, the preferred option is to maintain the metadata in the data provider's catalogue (national or project-specific) and use the harvesting mechanism (via CSW) to transfer the metadata records in the EGDI catalogue, which should serve as the central access point to metadata concerning structured geoscientific data sources.
Metadata are freely accessible to the public for viewing and searching but inserting and editing is for authorized users only. Each GeoERA project shall have an active metadata contact responsible for creation and maintenance of metadata. On https://egdi.geology.cz/?ak=cookbook the current version of the cookbook for creation of metadata for use in the GIP-P is available. Within the frame of the WP8 of the GIP-P project,

an updated version will be created, that would take into account any possible additions to the metadata profile to cover other OGC services (SOS, SensorThings API, WCS metadata).
A built-in metadata validator is part of the metadata catalogue for metadata records to ensure compatibility with INSPIRE. It may be modified to meet any specific need (e.g. use of mandatory project thesaurus keywords).

This is the recommended metadata tool for use by the GeoERA GIP.

3.2.2 Schema Mapping

HALE

A dedicated section in the GitHub project repository (https://github.com/GeoEra-GIP) has been created by WP3, where the main HALE resources are made easily accessible for the project partners:

- The main HALE page
- Download HALE
- Hale Documentation
- HALE Blog
- HALE video tutorial

In this sub-section, the use the HALE is presented by means of an example made to transform the geologic unit data using GeoSciML 4.1 and INSPIRE GE. In addition, the use of the groovy scripts, allowing an even more flexible and customised use of the HALE tool is described at the end of this subsection.
The following steps are presented in detail:

- Overview of the Hale transformation project
- What an alignment is
- Steps to Data Harmonization
- The Hale Workbench
- The default perspective
- The data perspective
- The map perspective
- Overview of the Hale transformation project

What an alignment is

The alignment is the mapping between source and target schemas. It defines relations between source and target entities (types or properties). Based on the defined relations, a transformation is derived. Each relation is represented in the Alignment by a mapping cell. In the image below a mapping cell is represented as it would be displayed in the Alignment view. In the example, the type GeoData from the source schema is mapped to the type Geologic Unit in the target schema (downloaded from the GeoSciML and/or Inspire website), the relation is represented by the Retype function (figure 2).

Figure 2 Retype function to map source with the target FeatureType element.

Steps to Data Harmonization

When creating a Hale alignment project, you follow these main steps:

1. Import the schema of the source data. (e.g. the shapefiles OSTGIS feature Geo100kData + Lithology100k.Table)
2. Import the target schema (e.g. http://schemas.opengis.net/gsml/4.1/geoSciMLBasic.xsd)
3. Import the source data set (e.g. POSTGIS feature Geo100kData + Lithology100k.Table)
4. Identify the relevant target types
5. Identify the information present in the source classes and information needed in the target ones
6. Identify relations between the source and target properties, then add them to the mapping

The Hale Workbench

The geom property of the shapefile is mapped to the geometry "AbstractCurve.CompositeCurve" element in the target schema. For each value in the geom, the rename function adds the same value to the CompositeCurve property in the transformed data.

In the HALE data perspective, you can examine the source and transformed data, e.g. comparing a source instance with the corresponding transformed instance. Through a filter query you can select certain instances for analysis.

In the following a short description of the perspective's views in the previous shot is provided:

- The Source Data view displays samples of the loaded source data (i.e. our source shapefile). A filter query can be used to control which instances are displayed.
- The Transformed Data view displays samples of the transformed data. By default it is synchronized to the Source Data view and contains the transformation result of the instances represented there.
- The Alignment view displays the current alignment per type relation and allows editing or removing mapping cells.
- The Properties view displays information on the current selection, in the above image this is the explanation of the mapping cell selected in the active Alignment view.

In the HALE data perspective, the Report List tab provides an overview of the last completed processes on the data transformation and their status (success / failed).

The HALE Map view provides a cartographic representation of the data. Source and transformed data are displayed alongside each other, with different layouts to choose from. The map can be used to select instances for examination in the data views, or vice versa.

HALE: Groovy Transformation Scripts

As highlighted and reported in the on-line HALE Users guide, while the transformation functions delivered with HALE cover a lot of issues, you may need to provide your own functions or customize existing ones (for example to conditionally execute a transformation). It is possible to combine the regular HALE transformation functions with Groovy scripts. HALE provides easy-to-use APIs for accessing and creating complex instances. To author the scripts, a script editor, is included that supports syntax highlighting and script validation. Example code for groovy property transformation and groovy type transformation is available in Hale

3.2.3 Data Transformation - INSPIRE Network Services/ Web Services

Data transformation using HALE can be published using tools such as GeoServer/MapServer.

HALE can also pre-generate a GeoServer app-schema configuration to ease in the set up of WFS flows compliant with an application schema (ex: an OGC standard, INSPIRE data specification).
The Geospatial data stored as GeoServer source data can be associated with complex object-oriented information models thanks to the application schema extension of the GeoServer software. The app schema module takes one or more of these simple feature data archives and applies a mapping to convert simple feature types into one or more complex feature types that conform to a GML application schema.
Conversely, the data harmonized and transformed by Hale into GML can be archived and delivered in the structure transformed by other tools such as MapServer or GeoServer itself.

3.3 Example of data harmonization

3.3.1 Groundwater examples

An example of use can be found at http://gin.gw-info.net/service/api_ngwds:gin2/en/wmc/standard.html

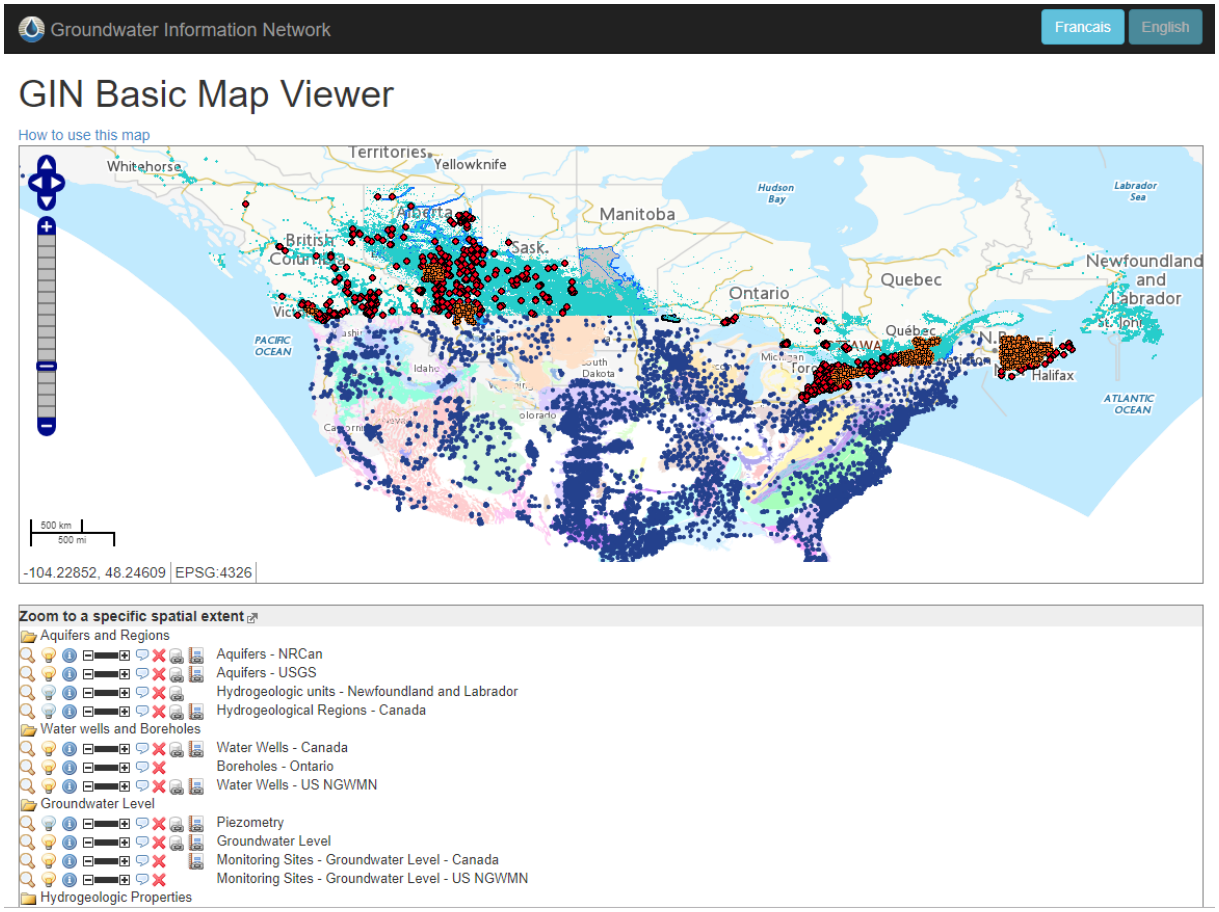

Figure 3. Groundwater Information Network portal showing harmonized groundwater data for North America.

3.3.2 *Mineral resources examples*

http://portal.geoscience.gov.au/

Figure 4.AUSGIN Portal showing harmonized mineral occurrence data

3.3.3 *Geology examples*

): http://www.europe-geology.eu/onshore-geology/geological-map/onegeologyeurope/

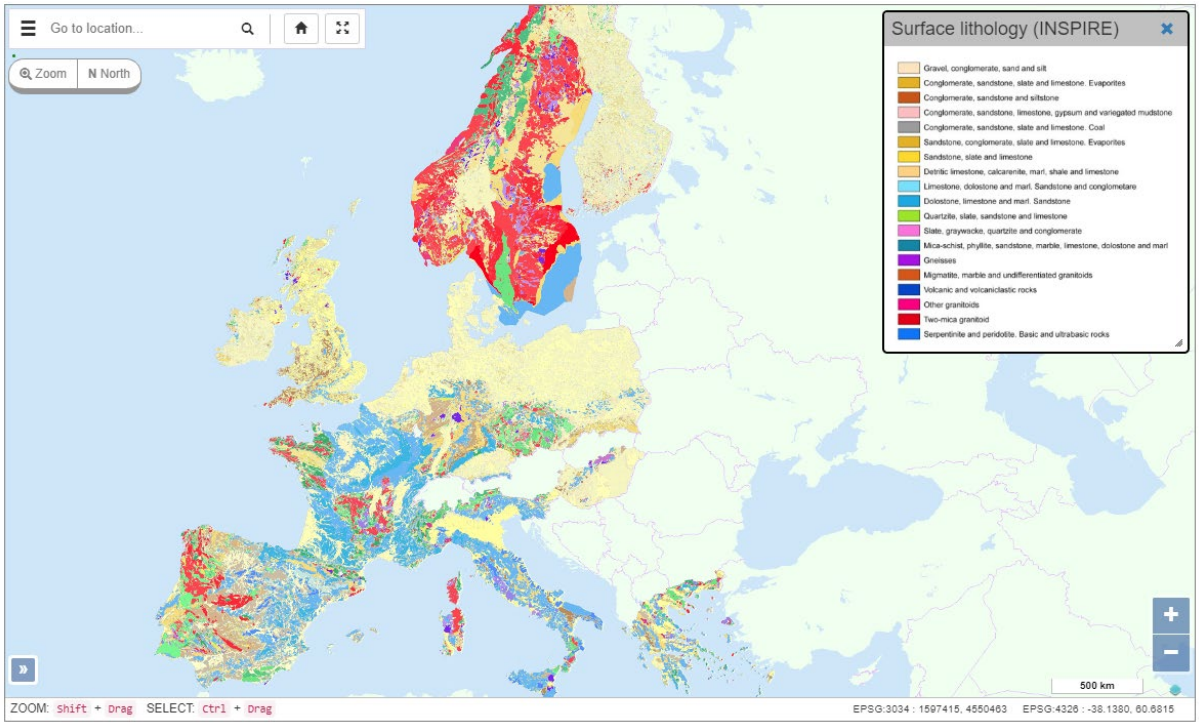

Figure 5. OneGeology Europe web portal with harmonized geology.

4 RECOMMENDATIONS AND CONCLUSIONS

We recognize that one of the main activities is the identification of the data (and related metadata) needed to the project application. For this reason, in order to respect this assumption, it is necessary:

- To identify the data necessary for the application;
- For each of these datasets, to identify shared data models (mainly referring to INSPIRE and OGC data models) in which each of these data is included;
- If it is not included in any data model, to schedule an extension of one of the data models involved that may better include information.

This action must be performed both for the input data and for the output data.

Recommendation 1
It is recommended to monitor the repository website of the international standard organisations in order to ensure the alignment with the most updated versions of the data models and of the INSPIRE implementing rules and Technical Guidelines as well as to check the on-going discussions and open issues published in the INSPIRE Community Forum related to the involved data themes.

After completing the whole interactive workflow, a user will be able to:

- Identify the properties of the data to be transformed according to the Data model definitions;
- Identify the missing information in the datasets as required by the OGC or INSPIRE object definitions. (Gap analysis);
- Identify a potential extension of the data model, to cover its entire data set.

Recommendation 2
A deep analysis of the technical specifications of the selected data models must be carefully performed to better understand the data structures and the correct understanding of the implementation described to correctly map their data in the selected models.

Recommendation 3
Limit as much as possible the extension of the models, and prioritise those extensions which impact on the data structure. Check carefully the mapping into the data models available for the needed data.

Recommendation 4
In case of data model extension, check the reference thematic community (e.g. INSPIRE Community Forum) if a similar extension has already been proposed/discussed/solved.

The referenced methodology proposed and applied by the exploitation of the harmonization toolkit delivered by the project has been designed and refined for the harmonization process and dedicate the time necessary for a complete understanding of the destination scheme, for the detailed compilation of the correspondence tables before implementing the effective harmonization process using the selected tools.
A common error in data harmonization is starting from scratch to use programs that perform data transformation without first defining the mapping rules.

Recommendation 5
It is recommended to Design data mapping using the matching table and filling them in as detailed as possible to provide a better understanding of the overall harmonization and to saves time in the last phase of mapping implementation using the selected software tool. Setting up instance XML file parallel to feeding a matching table is highly useful.

Recommendation 6
In the process of harmonizing a source schema with respect to an INSPIRE target scheme, it is important to map all information from the source schema to the destination schema, regardless of whether this information is classified as mandatory or not in the destination schema. In other words, do not limit the filling of mandatory fields but provide all available information.

There are some aspects that must be taken into account in order to provide interoperable metadata. These are:

- The completeness and correctness of the information that must be guaranteed by well-defined rules for filling in some free text metadata fields (such as abstract, lineage, ...)
- The selection of keywords to describe the data and to search for data that should be mainly:
  - or Independent language;
  - Or Shared by the different thematic communities interested in the described data.

This last aspect mainly refers to the semantic aspects of the content defined in the metadata.

Recommendation 7
The compilation of the metadata fields must be detailed as much as possible, taking into account in general the possible applications in which these data can contribute. Although many metadata fields are defined as non-mandatory by the reference regulation (INSIPRE/ National/Regional), it is suggested to fill in as many information as possible, interpreting all the metadata fields as mandatory.

Recommendation 8
Describe the data uniformly by selecting keywords from a list of existing shared codes and a thesauri, avoiding the use of free text keywords as much as possible.

The harmonization of data and metadata based on a reference target scheme that represents an international reference standard, such as the INSPIRE implementation rules, requires a series of procedures to ensure the compliance of the data and the metadata produced with these rules. The validation services provided by the GIP project meet this expectation through the exploitation and extension of existing official services. The procedures for validating the metadata and the data produced will be described in detail in D.3.3 and a specific validation service to support the procedure will be developed and delivered by the final platform.

The recommendations produced by this document represent a general guide to all the scientific projects of GeoERA.