EU Framework Program for Research and Innovation (SC5-18a-2014 - H2020)

Project Nr: 641538

Coordinating an Observation Network of Networks EnCompassing saTellite and IN-situ to fill the Gaps in European Observations

Deliverable D4.1
Observation inventory requirements, database schema and queryable fields

Version 1

Due date of deliverable: 31/07/2015
Actual submission date: 11/11/2015

Document control page				
Title	D4.1 Observation inventory requirements, database schema and queryable fields			
Creator	DN_52N			
Editors	DN_52N and SN_CNR			
Description	Report on the observation inventory requirements, database schema and queryable fields to be adopted during the project. The observation inventory will be based on GCI Information, DAB, and Copernicus services catalogues.			
Publisher	ConnectinGEO Consortium			
Contributors	ConnectinGEO Partners			
Type	Text			
Format	MS-Word			
Language	EN-GB			
Creation date	01/06/2015			
Version number	1			
Version date	11/11/2015			
Last modified by	DN_52N			
Rights	Copyright © 2015, ConnectinGEO Consortium			
Dissemination level		CO (confidential, only for members of the consortium)		
	X	PU (public)		
		PP (restricted to other programme participants)		
		RE (restricted to a group specified by the consortium)		
	When restricted, access granted to:			
Nature	X	R (report)		
		P (prototype)		
		D (demonstrator)		
		O (other)		
Review status		Draft	*Where applicable:*	
	X	WP leader accepted		Accepted by the PTB
		PMB quality controlled		Accepted by the PTB as public document
	X	Coordinator accepted		
Action requested		to be revised by all ConnectinGEO partners		
		for approval of the WP leader		
		for approval of the PMB		
		for approval of the Project Coordinator		
		for approval of the PTB		
Requested deadline				

Revision history			
Version	Date	Modified by	Comments
0.1	04-06-2015	DN_52N	Created the basic structure of the deliverable
0.2	18-06-2015	SN_CNR-IIA, DN_52N	Comments to the TOC, added more sections and comments for responsibilities
0.3		DN_52N	Add introduction and scope section text; incorporate changed by JM_CREAF and MvdB_ST
1	11-11-2015	DN_52N	Add missing sections; revise to conform WP4 architecture discussion at PTB Paris

Contributors	
Acronym	Full name
JM_CREAF	Joan Masó (CREAF)
DN_52N	Daniel Nust (52 North)
MvdB_ST	Maud van den Broek (S&T corp.)
SN_CNR	Stefano Nativi (CNR)

Table of Contents

1 Introduction 5

1.1 The ConnectinGEO observation inventory 5

1.2 Scope & purpose of the document 6

2 Requirements 7

2.1 Supported data sources 7
2.1.1 GEOSS DAB 7
2.1.2 Additional standardized data sources 8

2.2 Connection with external information 8
2.2.1 Introduction 8
2.2.2 Enrichers 9
2.2.3 URR (SEE IN Knowledge Base) 10
2.2.4 Forecasts & predictions 10
2.2.5 Copernicus services catalogues 10
2.2.6 User feedback 10
2.2.7 Scientific literature 11
2.2.8 Thesauri 12

2.3 Filtering and searching capabilities 13
2.3.1 Spatial filtering 13
2.3.2 Temporal filtering 13
2.3.3 Data availability 13
2.3.4 Data quality 13
2.3.5 Essential Variablesfeedback server and request for relevant feedback information via 13

2.4 Infrastructure and Deployment 14

2.5 Analysis requirements 14

3 Database schema 15

3.1 Fields 15

3.2 Queryables 16

1 Introduction

The exercise of creating a GEOSS database from the GEOSS catalogues has been done in the past to assess the quality of the metadata of the data related to GEOSS. This was done twice in the GeoViQua project and the last one published in Zabala et al.[1]. At that time GEOSS catalogue was aggregated in the GEOSS clearinghouse and the number of records was below 100.000.

Creating an observation inventory form the new GEOSS DAB will be more challenging due to a record count of almost 40 million records reported in the last DAB Report by CNR.

1.1 The ConnectinGEO observation inventory

The ConnectinGEO observation inventory (OI) is part of the ConnectinGEO methodology to identify gaps in the Earth observations in Europe and to assess the priority of these gaps. The methodology itself consists of several different threads to derive observation requirements and analyse the current state of observations. One of the threads is a bottom-up systematic analysis of observations and measurements that are currently undertaken and accessible through the GEOSS Common Infrastructure (GCI). The observations include space-based, airborne and in-situ platforms.
This document collects the requirements towards the OI. It contains questions that analysts, scientists or decision makers might have about the state of observations in Europe. These questions could potentially be posed to the OI respectively be answered by the metadata captured in it. Therefore this document translates these questions (the analysis requirements) together with ulterior requirements into an abstract database model and a list of fields that each database entry can have, so called queryables.

Ulterior requirements come from different areas. First, to provide a coherent picture, the OI will integrate information from different sources, which can be broadly divided into core data sources and external information, which are both specified in this document because they also have implications on the database model. One major external information item concerns the EO requirements collected in other threads of the ConnectinGEO methodology, which must be connected to the OI.
The OI will become a database internal to the project that will provide data for different analysis tools which create plots, reports, or summary statistics. These infrastructure and deployment requirements are also presented.

[1] ZABALA A., RIVEROLA A., SERRAL I., DÍAZ P., LUSH V., MASÓ J., PONS X. y HABERMANN T. (2013) Rubric-Q: Adding Quality-related Elements to the GEOSS Clearinghouse Datasets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing" (IEEE J-STARS) 6 (3) 1676-1687

1.2 Scope & purpose of the document

This document collects requirements from different perspectives towards a technical platform to support the analysis of the state of observations in Europe. These requirements are translated into a database model for the ConnectinGEO observation inventory, effectively listing all the pieces of information that must be stored. This document provides the basis for a priority-aware implementation of the OI and captures the currently existing ideas for products (reports, statistics) that can be derived from the OI.

2 Requirements

2.1 Supported data sources

2.1.1 GEOSS DAB

The observations inventory will not harvest any data but just metadata about data that can be accessed through GEOSS. The primary source of information for this metadata is the Discovery and Access Broker (DAB). The DAB automatically harvest all the records available from a list of data catalogues. The DAB release periodic reports on the different sources of records that they regularly harvest. The last one available report is available online[2]. The following list presents selected large catalogues, their its number of records, and the source of the number from the report:

6	ArcGIS Online ESRI	185.000	P
20	CEOS WGISS Integrated Catalog (CWIC)	1.844	P
24	CSW/ISO (at http://eopower.grid.unep.ch:8080/gi-cat/services/cswiso?service=CSW&request=GetCapabilities&acceptversions=2.0.2)	1.463	G
25	CSW/ISO (at http://ims.geoportal.de/gdi-de/srv/en/csw?service=CSW&request=GetCapabilities&acceptversions=2.0.2)	133.154	G
27	Chile Geoportal	11.628	G
42	Digital Globe	586	G
43	EEA SDI Catalog	415	G
45	ESRI-GEOPORTAL (at http://geoss.esri.com/geoportal/csw/discovery?service=CSW&request=GetCapabilities&acceptversions=2.0.2)	27.241	G
59	FedEO	2.210[3]	See footnote
62	GEONETWORK (at http://ims.geoportal.de/dgeocat/srv/eng/csw?service=CSW&request=GetCapabilities&acceptversions=2.0.2)	135.047	G
68	Geoscience Australia	21.743	G
69	Geoscience Data - Données Géoscientifiques (WMS)	10.266	G
70	Global River Discharge Datasets (GRDC/GEOWOW) - Kisters AG	1.928	G

[2] http://reporting.geodab.eu/DAB-Report-2015-10-02.html

[3] The number of granules is obtained by issuing a request with no constraint to each collection, and summing the total number of records

71	HIS Central US	13.229.445[4]	See footnote
73	INPE CDSR	899.492	G
74	IRIS Event	4.179.043	G
75	IRIS Station	544.991	G
76	ISPRA Monitoring network	15.948	G
83	NASA Global Change Master Directory	28.108	G
86	NOAA Unified Access Framework Catalog	3.988	G
88	New Zealand government geodata catalog	7.118	G
89	ORNL DAAC WCS Server (WCS)	8.512	G
90	ORNL DAAC WMS Server (WMS)	8.438	G
93	PANGAEA	346.304	G
101	South African Environmental Observation Network	12.499	G
103	US Data Gov	9.895	G
104	US NODC Collections	28.690	G
105	United States Geoscience Information Network Metadata Catalog	44.750	G
121	WIS GISC DWD	145.158	G
129	Webservice Energy Catalog	1.664	G

The total number of record is: 38.458.038

Source statements:

- P = Declared by Provider
- G = Number of records and granules harvested by GEODAB

2.1.2 Additional standardized data sources

If an organization wants to contribute standardized metadata that is currently not part of the DAB but in theory compatible (e.g. the data is published in an OGC Catalogue), CNR and GEO together are open to consider to add their metadata to the DAB. In particular, Task 5.4 of ConnectinGEO proposes to do this for the in-situ networks that are found not to be part of GEOSS yet. ConnectinGEO will collaborate with them in their integration. In general, this means that there is no need to support any other data source during the project.

2.2 Connection with external information

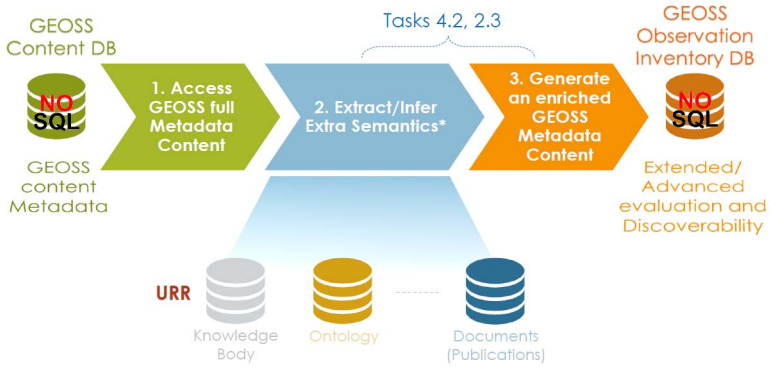
2.2.1 Introduction

The monitoring and analysis of the current observations contributing to GEOSS through the GCI via the DAB uses the created observation inventory

[4] See "Number of Results" section at http://essi-lab.eu/do/view/GIcat/HYDRODetails

(OI). This is a systematic approach to recognize the different observed properties and their types. The process can be divided into three steps:

1. Accessing GEOSS full metadata content
2. Extract relevant metadata and infer extra semantics
3. Generate enriched GEOSS metadata content

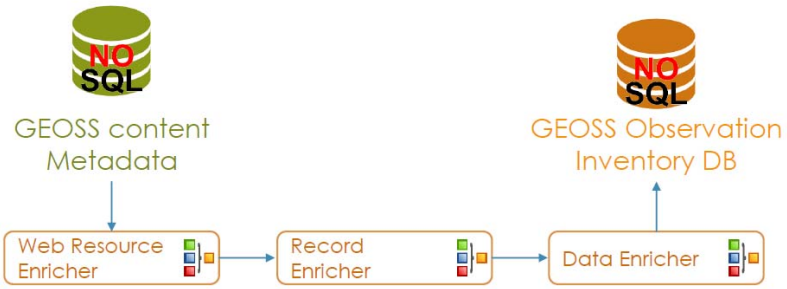


In this chapter of the document, we describe the sources for the extra semantics and additional metadata to be used in the second step. Within this step, as set of rules is applied to enrich the full metadata content.

2.2.2 Enrichers

The following enrichers, following the enricher pattern[5], can provide the critical missing pieces of information that make the analysis of the OI more powerful than previous approaches:

- Web resource enricher
- Record enricher
- Data enricher
- Document enricher



The enrichers will add tags[6] to the full metadata records and applied in an enrichment sequence, which eventually inserts the enriched record in the OI.

[5] http://www.enterpriseintegrationpatterns.com/patterns/messaging/DataEnricher.html

[6] https://en.wikipedia.org/wiki/Tag_%28metadata%29

2.2.3 URR (SEE IN Knowledge Base)

The Socio-Economic and Environmental Information Needs (SEE IN) Knowledge Base is the evolution of the User Requirements Registry (URR). This database will be used as a “user needs” database and will be useful for the gap analysis study. The information in this database will not be part of the OI but will be used later on in the project to compare, offerings (data) with needs and determine the gaps.

2.2.4 Forecasts & predictions

Models, forecasts and projections are also part of this "data need" but they also are capable of creating new data offerings on demand. Unfortunately we do not have any models database in GEOSS so far and will not be considered in this project threath. Requirements from models will be introduced in the gaps database using other project threads such us expert interviews and assessments.

2.2.5 Copernicus services catalogues

Copernicus service catalogues are or will be integrated into the GCI and are therfore available through the GEODAB and included in this thread.

2.2.6 User feedback

Within the GeoViQua project a tool has been developed for the provision of qualitative information to GEOSS dataset. This has resulted in the creation of a "User feedback" server. This server holds the feedback information that has been provided by expert or non-expert users such as comments, tags, or ratings,
but also specific information on certain areas within a dataset (discovered issues) or references to papers that have made use of the data.
The user feedback server is a source that can be used to complement the information retrieved via the GEO-DAB. Via the search API it is possible to scan the feedback information on notes of missing, incomplete data, discovered issues. The amount of usage reports and publications attached
can also provide information on the quality of the data.
The observation inventory will be coupled to the user feedback server allowing to add new feedback and retrieve existing feedback. To support gap analyses the feedback server will be extended with the possibility to create feedback on missing datasets. The design and implementation of this will be covered task WP4.4. The following requirements in relation to the OI, can be foreseen at this stage:

1. The user feedback server shall be coupled to the Observation Inventory by providing links to related user feedback items from the OI.
2. The user feedback server shall support the provision of feedback on missing data, which has no data identifier.
3. The user feedback server shall support retrieval of existing feedback and the creation of new feedback items.

2.2.7 Scientific literature

Scientific publications can provide extremely detailed secondary information for datasets. Apart from the content, the mere existance of scientfic publications mentioning a (GEOSS) data set is a piece of information. Therefore, the OI must support a semi-automatic process to connect datasets with scientific literature. This should include the following two approaches:

1. *Recognize references* to scientific literature in (standardized metadata)
    a. Parse relevant ISO metadata fields
    b. Recognize DOI links in full text, i.e. anywhere in the document
2. *Search online* scientifc databases and APIs with selected search terms derived from the dataset's metadata in a semi-automated manner.
    a. Possible search terms are dataset title, short names, or keywords.
    b. Online scientific APIs shall be evaluated for their usability in a semi-automated process.
    c. A tool shall be provided to attach references to entries in the inventory to potential matches in scientific databases. These references shall include the full URL to access the search result, a timestamp, and a status field (new, confirmed, and potentially more).
    d. A human analysist shall be ableprovided with a tool or instructions to query the OI for all unreviewed potential links and how to confirm or remove them.

A non-comprehensive list of (lists of) potential databases to acces search is as follows:

- http://www.programmableweb.com/news/195-science-apis-springer-epa-and-ncbi/2012/03/28
- Springer API: https://dev.springer.com/
- Elsevier API: http://dev.elsevier.com/
- Berkeley library – APIs for scholarly resources: http://guides.lib.berkeley.edu/information-studies/apis
- rOpenSci: http://ropensci.org/
- PLOS Search API: http://api.plos.org/solr/faq/
- Microsoft Academic Search API: http://api.plos.org/solr/faq/
- Web of Science Web Services: http://wokinfo.com/products_tools/products/related/webservices/
- EarthDoc (EAGE): https://www.eage.org/?evp=1971
- GeoRef AGI: http://www.americangeosciences.org/georef/how-access-georef
- University of Oregon GEOBASE: https://onesearch.uoregon.edu/databases/database/ORG04090
- Elsevier GEOBASE: https://en.wikipedia.org/wiki/GEOBASE
- Monash university – databases by subject: Geosciences: http://guides.lib.monash.edu/subject-databases/geosciences

2.2.8 Thesauri

Some thesauri will be adopted as described coming from the unified project methodology described in D6.1 For example, a common thesauri for EV should be used derived form the work in WP2 and included in the deliverable D2.3.

2.3 Filtering and searching capabilities

The previous data sources exposed before will be harvested into the IO by CNR. The harvest will consist in a collection of records allowing manually created filtering criteria.

2.3.1 Spatial filtering

Only metadata about data over Europe will be collected. Spatial fragmentation can be a criterion for discarding datasets for the study.

2.3.2 Temporal filtering

We need data with a certain degree of data continuity. The lack of temporal continuity could be also a criteria for excluding a record from the study.

2.3.3 Data availability

Another criterion for filtering could be the capacity to access the data. Records that have no connection to data could be considered not appropriate for this study. This can be derived from data by scanning metadata records for links to data services and even checking if the linked data services exist.

2.3.4 Data quality

Data quality is one of the obvious criteria for discarding records in this study. Nevertheless, we know from the GeoViQua project that the number of record without data quality information is high and it cannot be a criterion for filtering because too many datasets would be lost.

2.3.5 Essential Variablesfeedback server and request for relevant feedback information via

One of the User Feedback API more clear cases to connect with the IO is the thesauri of Essential Variables (EV). One of the conclusions of the Bari Workshop[7] was that different communities have different levels of maturity for their EV. We can classify EV communities into the following broad groups:

- Communities that have defined their EVs
- Communities that have EV candidates but consensus process is under way
- Communities that does not have EV but there are some obvious candidates
- Communities that have not considered the EV or EV cannot be applied

It is clear that we need a list of the EV and to associate each record to them. Records that cannot be associated to them will be out of the scope of this study.

[7] http://www.gstss.org/2015_Bari/

Note that this means that level 0 and level1 data will be automatically discarded from the study, such as satellite data imagery. It can be used to derive higher level that can describe an essential variable but cannot be associated to an EV per se.

2.4 Infrastructure and Deployment

To minimize the devlopment overhead and improve performance, the infrastructure for the OI should be coupled closely with the existing GEOSS DAB systems. This allows to make use of the native no-SQL APIs to read the full metadata content of each GEOSS harvested resources and populate the OI.

In combination with the previously mentioned enrichment chains, this provides a flexible and scalable solutions, as enrichers can be added and removed as needed and also operate independently on the data.

2.5 Analysis requirements

Several data sources will be harvested for metadata as described in the previous chapters. This data is the basis for and analysis of the state of observations in Europe from a thematic perspective, which could include different application domains. Generally, requirements from these domains can be divided into two sections:

(a) Information we want **based on the present situation**, e.g. EVs must be part of the current GEOSS content shared by the providers.
   a. EVs
   b. Property and features (e.g. resolutions, extent, accuracy, ….)
   c. Accessibility level (reality vs. announcement)

(b) **A priori expectations** useful to test the new EO inventory to be developed, e.g. the expected Copernicus services.
   a. URR content
   b. Granularity of data
   c. Accessibility level (expectations)
   d. Quality level

The following **expert requirements** were submitted by members of the ConnectinGEO consortium:

- Ad-hoc aggregations: we need means to aggregate data within the inventory (temporally, spatially, by country) to create views on the data
- We must distinguish between remote sensing (EO) and in-situ data

- We need "level" information[8] for EO data (datasets rely on each other, so records must be separable between the different levels), for example "low level" observations vs. "high level" observations
- We need information about the recentness of the data, i.e. when was the last metadata and/or data update
- We need policy information
- We need to know if the services linked to from the metadata still online

For the analysis it is useful to use existing vocabularies to describe observed properties, EVs etc. The following list contains the vocabularies currently supported by the DAB. If available the unique identifiers from the vocabularies shall be used to identify queryable values, keywords, field values etc.

The list of currently used vocabularies in the DAB is as follows:

- GCMD - Earth Science Keywords
- GEMET - INSPIRE themes
- GEMET
- GEOSS - Earth Observation Vocabulary
- GEOSS - Societal Benefit Areas
- INSPIRE - Feature Concept Dictionary
- INSPIRE - Glossary
- ISO - 19119 geographic services taxonomy EuroVoc (domini 36, 48, 52, 56, 60 e 66)
- GEOSS AIP-3 Hydrosphere Vocabulary Cadastre and Land
- Administration thesaurus (CaLAThe) EuroGEOSS Drought Vocabulary

The domain requirements and expert requirements are taken into consideration for the following database schema.

3 Database schema

The requirements presented in this document must be translated into a concrete list of database fields and queryables that must be stored in the inventory. In this section, an abstract data model is specified.

3.1 Fields

Due to the amount of data that is present in the GEOSS DAB we suggest to minimise the observation inventory fields to the minimum possible number of actually required fields. This will decrease the load of the OI and potentially speed up the analyses build on top of the OI. Connection to the original complete record will be always possible through the metadata resource identifier, which can be used by applications to build a URL to access the full metadata record.

[8] https://earth.esa.int/web/guest/-/data-types-levels-formats-7631 and http://www.cesbio.ups-tlse.fr/multitemp/?p=3202

The following fields and cardinalities are part of an **OI entry**, the core element in the inventory. They can all be extracted from the ISO 19115 core metadata:

- Inventory ID (internal)
- Metadata resource identifier (1), public, e.g. GEOSS DAB ID
- Title (1)
- Abstract/description text (0..*)
- Producer name (0..*)
- Topic category (1..*)
- Theme/topic (0..*)
- Measurement (0..*)
- Keywords (0..*)
- EV (0..*)
- Spatial extent (0..1)
- Temporal extent (0..1)
- Resolution/scale (0..1)
- Series/parent resource (0..1)
- Link to data (0..*)
- Published or planned (0..1)
- Availability/license/policy/restrictions (0..n)
- Other links (0..n), e.g. additional resources

3.2 Queryables

All fields are queryable in the OI, each having an appropriate index structure to allow good performance for expected as well as currently not foreseen questions and queries.