7th Framework Programme
ENV.2010.4.1.2-2
Integrating new data visualisation approaches of earth Systems into GEOSS development

Project Nr: 265178

QUAlity aware VIsualisation for the Global Earth Observation system of systems

Deliverable D7.2
Report on current GCI and identification of issues in integration with GEOSS

Version 1.0

Due date of deliverable: 31/01/2012
Actual submission date: 31/01/2012

<table>
<tr><th colspan="5">Document control page</th></tr>
<tr><td>Title</td><td colspan="4">D7.2 Report on current GCI and identification of issues in integration with GEOSS</td></tr>
<tr><td>Creator</td><td colspan="4">ES_UAB</td></tr>
<tr><td>Editor</td><td colspan="4">ES_UAB and MN_UAB</td></tr>
<tr><td>Description</td><td colspan="4">Report on current GCI and identification of issues in integration with GEOSS</td></tr>
<tr><td>Publisher</td><td colspan="4">GeoViQua Consortium</td></tr>
<tr><td>Contributors</td><td colspan="4">GeoViQua Partners</td></tr>
<tr><td>Type</td><td colspan="4">Text</td></tr>
<tr><td>Format</td><td colspan="4">MS-Word</td></tr>
<tr><td>Language</td><td colspan="4">EN-GB</td></tr>
<tr><td>Creation date</td><td colspan="4">09/01/2012</td></tr>
<tr><td>Version number</td><td colspan="4">1.0</td></tr>
<tr><td>Version date</td><td colspan="4">31/01/2012</td></tr>
<tr><td>Last modified by</td><td colspan="4">MN_UAB</td></tr>
<tr><td>Rights</td><td colspan="4">Copyright © 2012, GeoViQua Consortium</td></tr>
<tr><td rowspan="5">Dissemination level</td><td></td><td colspan="3">CO (confidential, only for members of the consortium)</td></tr>
<tr><td>X</td><td colspan="3">PU (public)</td></tr>
<tr><td></td><td colspan="3">PP (restricted to other programme participants)</td></tr>
<tr><td></td><td colspan="3">RE (restricted to a group specified by the consortium)</td></tr>
<tr><td colspan="4">When restricted, access granted to:</td></tr>
<tr><td rowspan="4">Nature</td><td>X</td><td colspan="3">R (report)</td></tr>
<tr><td></td><td colspan="3">P (prototype)</td></tr>
<tr><td></td><td colspan="3">D (demonstrator)</td></tr>
<tr><td></td><td colspan="3">O (other)</td></tr>
<tr><td rowspan="4">Review status</td><td></td><td>Draft</td><td colspan="2">Where applicable:</td></tr>
<tr><td>X</td><td>WP leader accepted</td><td></td><td>Accepted by the PTB</td></tr>
<tr><td></td><td>PMB quality controlled</td><td></td><td>Accepted by the PTB as public document</td></tr>
<tr><td>X</td><td>Coordinator accepted</td><td colspan="2"></td></tr>
<tr><td rowspan="5">Action requested</td><td>X</td><td colspan="3">to be revised by all GeoViQua partners</td></tr>
<tr><td>X</td><td colspan="3">for approval of the WP leader</td></tr>
<tr><td></td><td colspan="3">for approval of the PMB</td></tr>
<tr><td>X</td><td colspan="3">for approval of the Project Coordinator</td></tr>
<tr><td></td><td colspan="3">for approval of the PTB</td></tr>
<tr><td>Requested deadline</td><td colspan="4"></td></tr>
</table>

Revision history			
Version	Date	Modified by	Comments
0.0.1	10-01-2011	MN_UAB	Initial document description (PD_CREAF, PS_UAB, MN_UAB, JM_CREAF) and GCI analysis summary (AZ_UAB, MN_UAB, XP_UAB)
0.1	20-01-2012	ES_UAB	Created the basic content of the deliverable
0.2.0	23-01-2012	MN_UAB	Structuring the content
0.2.1	25-01-2012	ES_UAB	Adding and deepening contents (AZ_UAB, MN_UAB, PS_UAB, IS_CREAF, PD_CREAF, JM_CREAF, KY_UREAD, JB_UREAD, DC_ASTON, LB_ASTON, VL_ASTON, DN_52N)
0.2.2	26-01-2012	MN_UAB	Review of methods and results
0.3.0	27-01-2012	MN_UAB	Revision by GeoViQua partners (PD_CREAF, JM_CREAF, OG_CREAF, VG_ESA)
0.3.1	27-01-2012	ES_UAB	
0.4.0	28-01-2012	ES_UAB	EuroGEOSS Conference comments (JM_CREAF, PD_CREAF, IS_CREAF, MN_UAB, ES_UAB)
0.5.0	29-01-2012	MN_UAB	Revision (XP_UAB, JM_CREAF) and minor edits (AZ_UAB, GM_UAB, PS_UAB, AR_UAB)
0.6.0	30-01-2012	ES_UAB	Revision by GeoViQua partners (KY_UREAD, VG_ESA, DC_AST, ST_FRAUN)
0.7.2	30-01-2012	LB_AST	Minor edits. Clarifications.
1.0	31-01-2012	JM_CREAF	Coordinator acceptance

Institution	Contributors
52N	DN “Daniel Nüst”
AST	LB “Lucy Bastin”, DC “Dan Cornford”, VL “Victoria Lush”
CREAF	IS “Ivette Serral, JM “Joan Masó”, PD “Paula Díaz”, OG “Óscar González”
ESA	VG “Veronica Guidetti”
FRAUN	ST “Simon Thum”
UAB	AR “Anna Riverola”, AZ “Alaitz Zabala”, ES “Eva Sevillano”, GM “Gerard Moré” MN “Miquel Ninyerola”, PS “Pere Serra”, XP “Xavier Pons”
UREAD	JB “Jon Blower”, KY “Kevin Yang”

Table of Contents

Abstract

This deliverable is related to task 7.1 “Document current GCI services enabling quality-aware visualisation and search” in Work Package 7 (WP7) “Pilot case studies” in the GeoViQua 7th Framework Programme ENV.2010.4.1.2-2 project framework.

In collaboration with CREAF, WP7 activities linked to task 7.1 and WP3 activities linked to task 3.1 “Metadata extraction quality components” are fulfilled in this document and continued in the corresponding Deliverable 3.1 “Metadata extraction quality components”.

The GEOSS Common Infrastructure (GCI) has assembled metadata catalogue services focusing on retrieving the highest number of components, with the purpose of facilitating data discovery and access. GEO Portal users try to find fit-for-purpose data comparing metadata records.

Specific statements contained in metadata standards, suchlikesuch as quality elements, lineage and data sources are useful variables, however neither used to sort the results nor clearly shown in dataset descriptions. Whether the reason is that quality description is a complex subject not having received due attention, or else, there is not enough information, the matter is vital to the EC FP7 GeoViQua project intended to develop tools to elicit quality information in GEOSS.

A methodology for globally assessing the available quality information in the GEOSS Clearinghouse is presented. First, metadata documents are extracted from the GCI catalogue. Next, metadata elements are stored in tables that enable evaluating the relevant content. The analysis provides an overview of the current state of the metadata records, identifying the most widely used descriptors and protocols, and assessing the level of quality comprehensiveness.

Results show that quality indicators and lineage statements (following ISO standard ISO 19115:2003 –ISO-TC211 2003-), although sometimes present, are astoundingly scarce (around 19%). Not unexpectedly, the most frequent quality parameters are positional accuracy and completeness. However, surprisingly the Clearinghouse catalogue comprises a diversity of quality indicators, including quality measurements in the form of coverage results, although in a very few documents. Quality measures include conformance and quantitative results, the latter often lacking measurement units and descriptions. Some statements referring to usage are also embedded in the sample.

Another point is that character strings are provided in local languages, at present affecting geosearch and semantic Brokers. Interoperability and the functionality to assist with quality information in the quality information provision and flow are essential in the evolving GEOSS architecture.

On the whole, results are better than anticipated, especially when comparing with other spatial data infrastructures. Nonetheless, the user requirements surveyed in GeoViQua differ significantly from the statements considered in the current ISO standards. In the end, the Quality Model should be changed to allow for a matching between the soft knowledge demanded by users, the current quality indicators enhancement and the detected gaps in the existing elements in the standards.

This work proves the incompleteness of metadata and harvesting processes, seriously threatening a proper interpretation and GCI efficacy, providing a consistent analysis methodology and corrective proposals. Data quality concepts and existing standards are reviewed under the scope of GeoViQua activities, resulting in a global conception of components and tools to be developed according to the state of the art as regards quality awareness, availability and implementation in GEOSS infrastructure.

1. INTRODUCTION

1.1 *GEOSS components under the GeoViQua framework*

This deliverable is focused on analysing quality information in the Global Common Infrastructure (GCI) in GEOSS (Global Earth Observation System of Systems).

The Global Earth Observation System of Systems (GEOSS), coordinated by the Group on Earth Observations (GEO) (Global Earth Observation System of Systems 10-Year Implementation Plan, 2005), is a public infrastructure that interconnects a diverse and growing array of data, instruments and systems to allow monitoring and forecasting changes in the global environment. GEOSS gives support to policy-makers, resource managers, scientific researchers and many other experts and decision makers in their daily work with Earth Observation (EO) data.

The GCI consists of Web-based Portals, a Clearinghouse for searching data, information and services, and registries containing information about GEOSS components, standards, and best practices. That is to say, the GCI Clearinghouse is the GEOSS registry and metadata catalogue.

The goals of GCI in 2011 are to facilitate access to datasets and other resources supporting the Earth Observation Priorities identified in User Interface Committee (UIC) report, by enabling discovery and facilitating access to data-core contributions. Other proposed actions are the adaptation of interfaces in order to permit effective data finding and downloading ('fewer clicks to the data'), and the improvement of usability/user experience.

The GCI is subjected to continuous development and improvement of its functionality, from infrastructure and back-end services to context-driven applications and human-computer interfaces. A major target for this development is the wide variety of Societal Benefit Areas (SBA) that GEOSS addresses: Health, Disasters, Weather, Energy, Water, Climate, Agriculture, Ecology and Biodiversity.

In turn, the GEO Portal is aimed at the discovery and visualization of EO data in an integrated, standardized and interactive way, enabling a wide use by the scientific community when dealing with representation and modeling of Earth Systems.

Under the GEO umbrella several tasks address the dynamic architectural implementations in GEOSS, including components and tools development, development which must guarantee multidisciplinary interoperability in addition to enhancing quality elicitation and awareness.

QUAlity aware VIsualisation for the Global Earth Observation system of systems (GeoViQua) is a European project framed within GEOSS acting as a testing platform, evaluating GEOSS progress and providing reports on usability, visualisation, quality parameterisation and search with a special remark focus on quality issues. Thus, continuous inputs and feedbacks are implied between GEO tasks and GeoViQua tasks and even inside GeoViQua's tasks and work packages (WP), assuring coherence between projects, targets and tasks.

Partners involved in WP7 coordinating the deliverables are CREAF, UAB, 52N, AST, UREAD, CEA, ESA and S&T. This work package has been assigned 15.5% of the total workload in GeoViQua, as defined in the Description of Work (DoW). In particular, this document is related mainly to task 7.1 and task 7.2 in WP7, coordinated by UAB in GeoViQua. Extracted from the DoW:

"Task 7.1. Explore GCI services and components and GEO Portal search and visualisation tools. Evaluation of usability and detection of gaps. Selection of pilot cases".

This task therefore revises reviews GEOSS datasets (available through the GEO Portal) and, when available, associated quality data. The selection of datasets and pilot cases to be used in the practical implementation of GeoViQua scenarios was already reported in Deliverable 7.1. "Document pilot case studies in a standard format".

The GEO Portal (http://www.geoportal.org/web/guest/geo_home) provides convenient access to the full range of GEOSS data and information, linking to GCI where components and services are being deployed and catalogued. The current GEOSS Web Portal is provided on the basis of ESA's activity. GeoViQua solutions are expected to be integrated into the current GEOSS Web Portal logic.

As regards task 7.2, extracted from the DoW:

"Task 7.2. Quality product parameterisation. Classification of main factors of geospatial datasets, with special emphasis in quality components".

Results from WP2 (user requirements detection), WP 3 (quality elicitation; in particular, tasks 3.1, 3.2, 3.4 and 3.5) and WP6 (encoding) are inputs for this task. Outputs will provide a valuable feedback for WP2 and WP3 and, in turn, an input to task 7.4, related to the GEO Label. Currently, Deliverable 3.1 is running in parallel to Deliverable 7.1. In collaboration with CREAF, WP7 activities linked to task 7.1 and WP3 activities linked to task 3.1 "Metadata extraction quality components" are fulfilled in this document and continued in the corresponding Deliverable 3.1 "Metadata extraction quality components". As stated by **Parr and Lu (2010)**, recent geographic information research priorities include visualization, uncertainty and spatial data infrastructure issues, all of which are tackled within GeoViQua.

In summary, GeoViQua will provide a set of scientifically developed software components and services that facilitate the creation, search and visualization of quality information on Earth Observation data integrated and validated in the GCI. In order to provide innovative tools that will enhance the current infrastructure capabilities, an analysis of the quality components that can be discovered and accessed through the GCI by the community of users is the mandatory first stage.

The present deliverable (D7.2 "Report on current GCI and identification of issues in integration with GEOSS") reports the progress on WP7, focusing especially on Tasks 7.1 and 7.2, and making use of the results of the analysis of the GCI content.

The following diagram presents an overview of GeoViQua components. More specifically, this document is focused on the path from the Catalog Services for the Web CSW Clearinghouse. This path goes through the EuroGEOSS discovery Broker and its extended GeoViQua quality-aware component to the enhanced geosearch, visualization and intercomparison quality-aware GeoViQua tools which are accessible to the user. To make this possible, the first step is an exploration of the contents of the quality components Clearinghouse, which is the main subject in this deliverable. It is necessary to track the process and survey whether information is missed or corrupted on the way to the user or else is smoothly transmitted and arrives straightforwardly. However, before this step we need to find out the initial state of the art regarding quality information in the documentation of data products and services. Another relevant point tackled in GeoViQua is user feedback, implying documentation updating and enriching through, for instance, validation processes, allowing for a quantitative and qualitative evaluation of quality and usage that should be implemented through a bidirectional path within the system architecture.

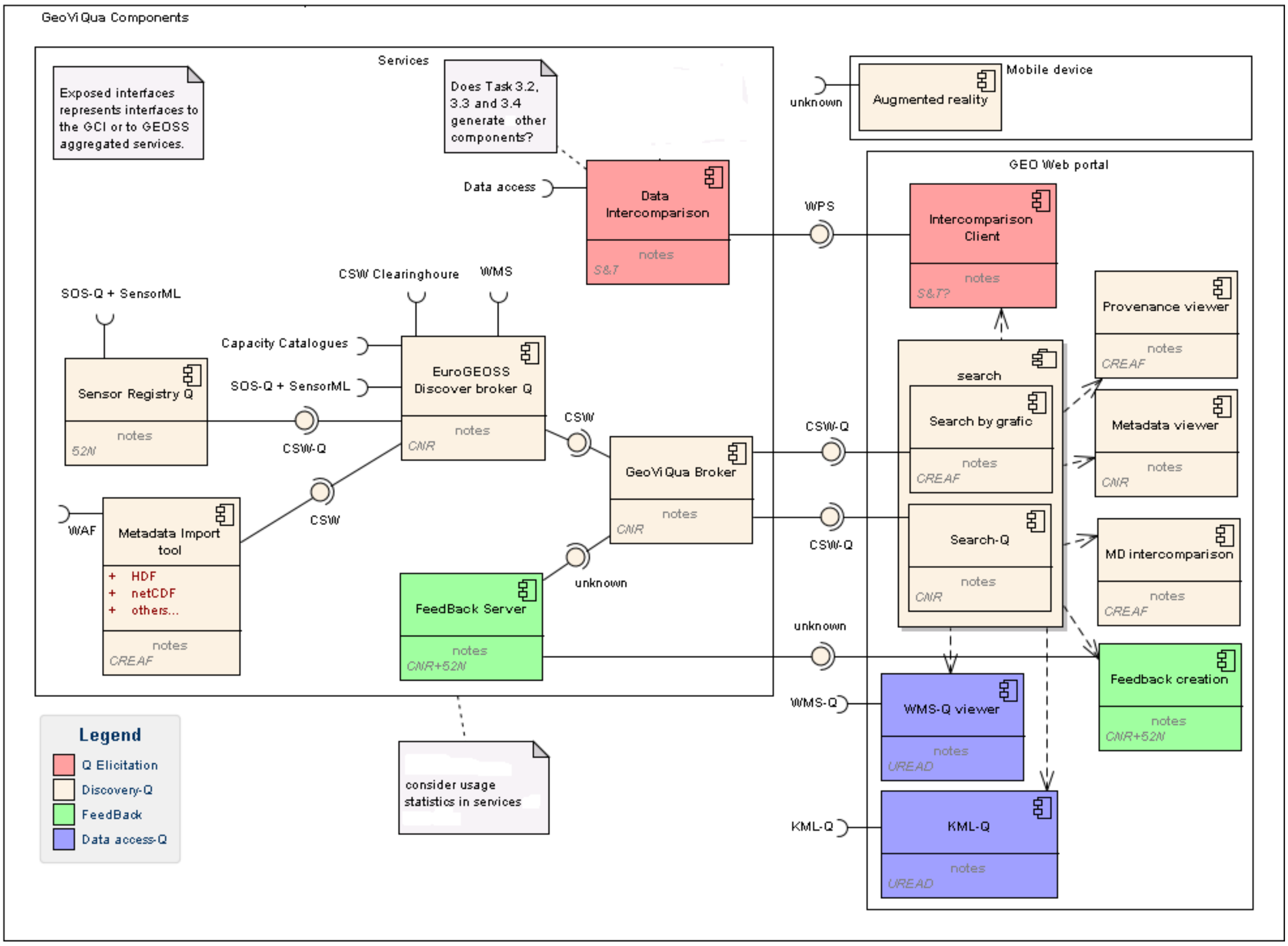


Figure 1. Overview of GeoViQua quality-aware components (note the –Q extension for Quality awareness components and tools)

A few notes should be stated on the diagram for clarification purposes. First, it must be noted that GeoViQua components are dynamic and respond to the processes in GEOSS. At present, the functionality for the "Inter-comparison" components is not yet clear. Requirements related to this component shall be further clarified in order to assess the feasibility of this integration into the GeoViQua component. Therefore, although the component would enhance expert user assessment of quality it may be considered an add-on to GeoViQua rather than one of its components. This component is more a data processing functionality than a quality access and visualisation tool.

As regards the WMS-Q viewer, the challenge is to find a way to translate quality information available from ISO 19115:2003 (**ISO-TC211 2003**) and the future GeoViqua Quality Model (see section 2.3 Quality elements and measures) into graphical information displayed via WMS services. As yet, the WMS service does not support quality information. Likewise, the inclusion of quality fields in the evolution of KML standard within GeoViQua (KML-Q viewer component) poses similar challenges: some limitations may be experienced in the visualization of KML files on the 3D Globe.

Finally, on the Feedback Creation component, the interface and protocol are to be defined. An advantage is that in this respect there shall be no technical constraint on the Portal side about which protocol to use and how reviews/comments/ratings could be sent to users. For instance, on the client side (Feedback Creation) there are several available tools and javascript libraries which could be used to facilitate client-server communication and display rating information, using a 5-stars approach, on the user interface.

1.2 *GEOSS Common Infrastructure recent improvements*

In the GEO Web (http://www.earthobservations.org/gci_gci.shtml), there is a brief description of the main four elements in the GCI:

-The **GEO Portal** provides the direct Web interface through which the user accesses GEOSS and searches for information and services

-The **GEOSS Clearinghouse** is the engine that drives the entire system. It connects directly to the various GEOSS components and services, collects and searches their information and distributes data and services via the PortalPortal to the user

-The **GEOSS Components and Services Registry** is similar to a library catalogue. All of the governments and organizations that contribute components and services to GEOSS provide essential details about the name, contents, and management of their contribution. This assists the Clearinghouse, and ultimately the user, to identify the GEOSS resources that may be of interest.

-The **GEOSS Standards and Interoperability Registry** enables contributors to GEOSS to configure their systems so that they can share information with other systems. This Registry is vital to the ability of GEOSS to function as a true system of systems and to provide integrated and cross-cutting information and services. Contributors can also share ideas and proposals informally via the associated Standards and Interoperability Forum.
Finally, the ancillary Best Practices Wiki provides the GEOSS community with a means to propose, discuss and converge upon best practices in all fields of earth observation.

Each element of the GEOSS Common Infrastructure has been contributed by GEO members and participating organizations. To better understand how the Common Infrastructure functions, see the diagram below:

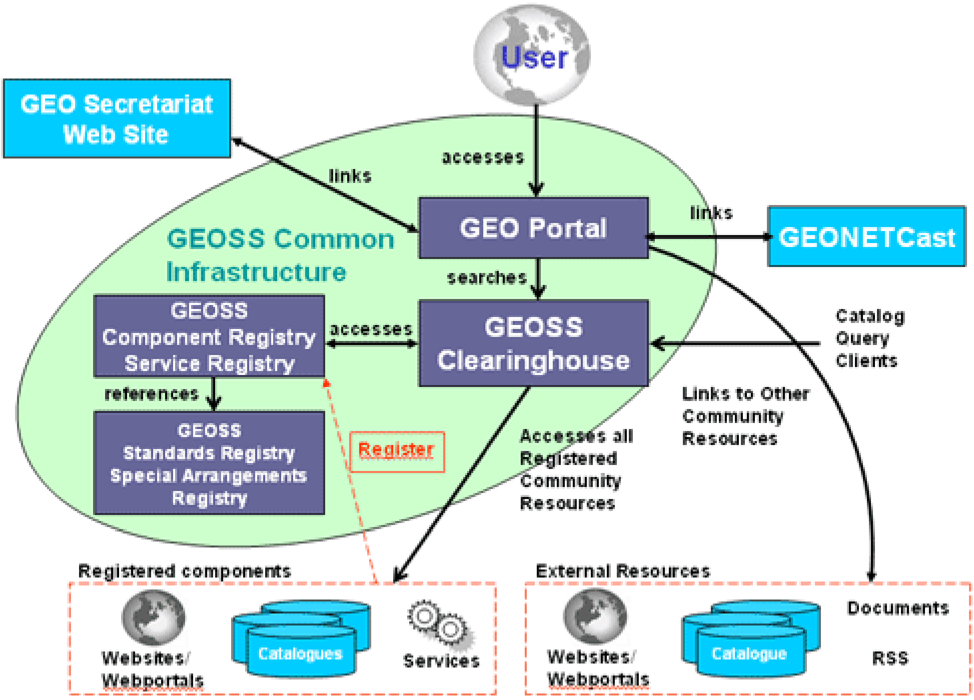


Figure 2. Overview of the GCI architecture in the GEO Web.

However, GEOSS is constantly evolving and since the beginning of GeoViQua, this general schema has been changed.

GeoViQua is actively participating in the development of the GCI new infrastructure. Improvements in the GEO Portal are directed to dataset visibility, promotion of GEOSS data-core, assistance in data discovery, and provision of visual guidance for resources quick recognition. Some tools that are being embedded in the new version of the GEO Portal are a 3D globe based on Google Earth plug-in, a new map viewer based on OpenLayers API allowing direct load of Web Map Service (WMS), advanced data access allowing a direct browsing of remote services through an open search description document, easy query and retrieval. Currently, the GEO Portal is being tested (http://scgeoviqua.sapienzaconsulting.com) in a controlled environment leveraging the portal functionalities and portlet technology.

In this sense, user requirements identified in WP2 within GeoViQua that are relevant to Portal testing environment are related to quality search, quality visualization, peer-reviewing and sharing concerns.

A relevant component introduced into the GCI architecture is the EuroGEOSS Broker, which now links the GEO Portal and the Clearinghouse. The GEO Portal interaction with the broker is made via Catalogue Service for the Web CSW 2.0.2. The portal data display will depend on results provided by the catalogue through the broker, whereas the portal functionality will depend on the metadata organization in the database, the catalogue service interfaces allowing specific queries, and results provided by catalogue and Broker.

As regards the EuroGEOSS Broker, GeoViQua is also directly involved: the Broker is a core component particularly of Work Package 4 entitled "Enhanced geosearch components". Among other benefits, the Broker will improve search capabilities within GEOSS. All this results in a new GCI architecture picture:

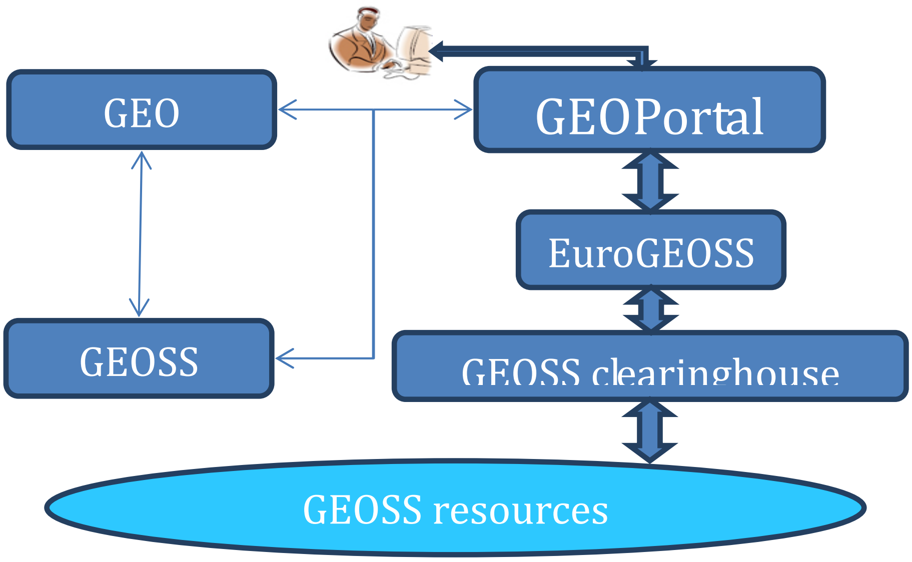


Figure 3. New GCI architecture as presented in GeoViQua 3rd Project Technical Board Meeting (Reading, United Kingdom, December 2011). (SAP-SD-ESA-EOP-PRE-0082, note)

Since November 2011, the Broker has been linked to several federated community catalogues, for instance GBIF, INSPIRE, and others, up to approximately ten catalogues. The current list of capacity catalogues accessible through the brokered GCI as reported in the EuroGEOSS "Advancing the vision for GEOSS" symposia includes GENESI-DEC, CWIC, NASA GCMD, Sea Data Net, NOAA UAF, NASA GSFC-DAAC, EEA-SDI, SAEOS, IODE, One Geology, Enviro GRIDS/WaterML and INPE. Only the INPE and Enviro Grids/Water ML catalogues architecture have required the adoption of new access protocols. In addition, the Broker is linked to some Web-accessible folders and Web 2.0 resources. These other catalogues will be analysed in other deliverable in GeoViQua (Deliverable 3.1 "Metadata extraction quality component" that would be the following chapter to Deliverable 7.2 "Report on current GCI and identification of issues in integration with GEOSS").

An important difference between the Clearinghouse and these other catalogues is that the harvesting process is expected to extract Metadata from the Catalog Services for the Web (CSW) Catalogue. Through the Broker, from the GEO Portal the user has direct access to these metadata. In contrast, up to date, the EuroGEOSS Broker does not provide the functionality for extraction of the Metadata in the source XML files from the rest of catalogues that are accessed through it.

In order to enable quality metadata retrieval, standardization of quality parameters has been identified as an essential step. Accordingly, the Clearinghouse and the Broker shall be able to implement the standards. It must be remarked that, as is shown in the figures above, the Portal is at the end of the chain. Thus, the search and visualization of quality information will depend on the availability and search capabilities of information from the other GCI components (*i.e.,* EuroGEOSS Broker, Clearinghouse, Community Catalogues).

As regards visualization of data quality in the Portal, possible strategies include statements about quality information availability (*e.g.,* via icons), and when available, 2D charts, intercomparison maps, or quality/error flags on the user's query result list, metadata description or 3D globe. Peer-review inclusion requires features such as decoupling of metadata results from where they are stored, editing metadata (with the compulsory regard to access rights), external linking of metadata with related reviews and comments (*e.g.,* free-text comment, data rating, user details –identity, company/agency, contacts-, automatic recording of date/time comments, interactive links) or creating direct links between Portal and source database.

The Broker, as presented in EuroGEOSS in Madrid (January 2012), is designed to handle the diversity of services and standards across disciplines and thus implements components that account for quality and tagging, discovery, access, semantics and composition. Another relevant feature would be the versatility needed to adapt to evolving standards, using Web Processing Service (WPS) protocol and implementing extensions if necessary.

1.3 *Exploring the GCI*

This graph depicts the growth of the registered resources in the GCI from 2010 up to November 2011:

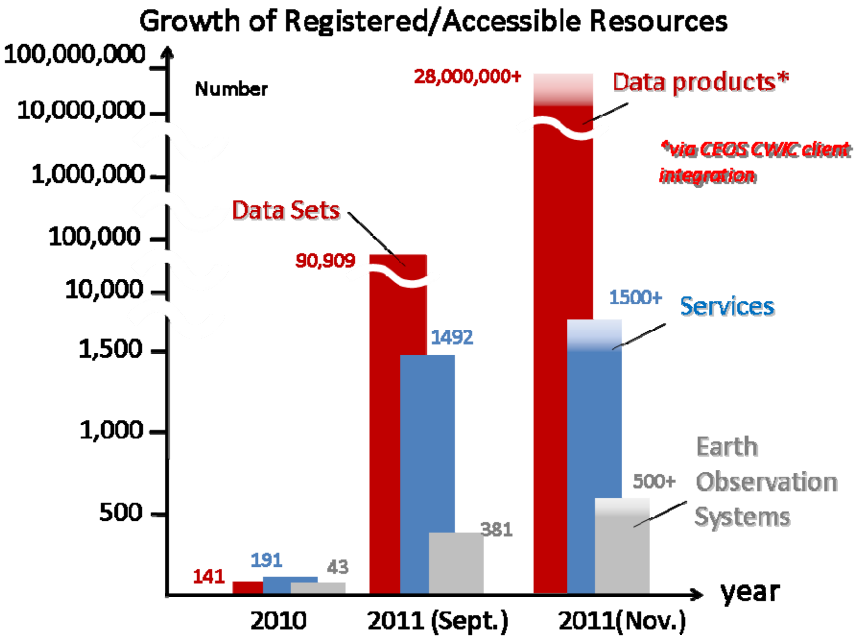


Figure 4. Evolution of resources in the GCI from 2010 to 2011. (Architecture and Data Committee ADC 2011)

*CEOS Committee on Earth Observation Satellites
CWIC CEOS Working Group on Information Systems and Services WGISS Integrated Catalogue initiative

The first relevant fact which draws attention is the huge increase in the number of data products, jumping from 141 in 2010 to 28,000,000 in 2011. The increment in services is of lesser importance yet nonetheless still significant, growing from 191 in 2010 to over 1500 in 2011. Lastly, Earth Observation systems also increase from 43 in 2010 to over 500 in 2011. Whether these numbers are exact or remain to be updated, the general trend is undoubtedly valid, providing a quick and certain impression on the dynamic evolution of the Clearinghouse contents and, by extension, of the GCI, GEOSS, and probably all geospatial dataset storing systems. The significant growth in the number of GCI datasets and services is due to the connection of around ten capacity catalogues to the GCI through the Broker in November 2011, as was explained in the previous section 1.2, accessible through the GEO Portal.

At the start of our analysis, the GCI was connected only with the Clearinghouse catalogue, representing the bar group in the middle of the chart above (*i.e.*, 2011 Sept.). Thus, in our study, the methodology applied enabled the extraction of the total number of 97203 metadata records available in October 2011 in the clearinghouse, resembling the total number of metadata records in this GCI component between September and November of 2011.

1.4 *User requirements surveys in GeoViQua*

The GCI infrastructure is the link between data producers and users. In GeoViQua, users have been given a special weight and a specific work package (WP2) has focused on revealing, retrieving, classifying and analysing user requirements as regards data products quality, visualization, search, discovery and access.

User requirements have been analysed and classified in a formal platform provided by Fraunhofer, GeoViQua's coordinator of this work package, currently for internal use. The requirements have been clustered into several areas: Feedback, Intercomparison, Provenance and Purpose, Quality and Uncertainty, Quality-aware Search, Standards and Communication and Visualization. Interesting conclusions have been derived, which are very valuable both for the evaluation of quantitative measurements of quality indicators and the so-called soft knowledge, a more qualitative evaluation of quality that has proved to be relevant to users.

In this respect, Aston University presented this summary in the Quality Assurance for Earth Observation (QA4EO) meeting (Oxford, United Kingdom, October 2011):

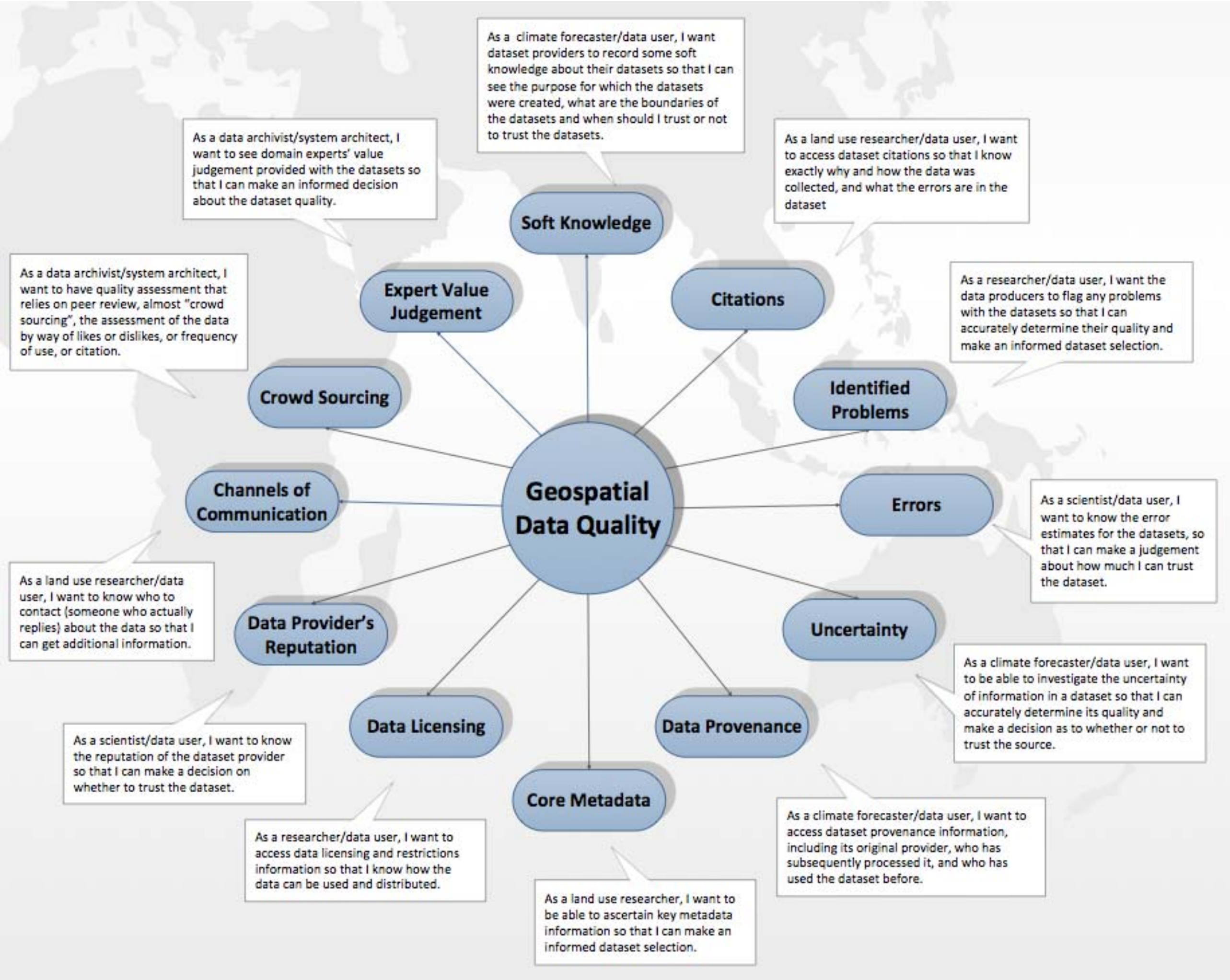


Figure 5. Synopsis of results related to the analysis of user requirements performed in WP2. Extracted from the poster presented in the QA4EO by ASTON University (Oxford, United Kingdom, October 2011) (**Lush et al. 2011**)

Users have identified that the following elements are usually missing or incompletely recorded:

-provenance
-citation
-licensing information

Other interesting elements for users are:

-comments on the overall quality
-errors, identified problems, uncertainty
-potential use

This means, ultimately, anything that could help in evaluating fitness-for-use of the data.

Users seek information on data providers, valid contact details, recommendations, feedback, peer-reviews, side-by-side comparison functionality, features from which to decide whether to trust or not to trust the dataset or be able to determine to what extent the data can be used. They are interested in assessing potential use, adapted to the quality of the data. Due to the relevant importance of the qualitative fitness-for-purpose information for users, it was decided to adapt our quality analysis to include the usage information embedded in the metadata records. In this respect, other solutions related with free comment platforms and metadata evaluation are analyzed in order to enable user feedback functionalities.

As for the quantitative measurements of quality indicators, GeoViQua has analysed the existing quality measurements based on the schema ISO 19115:2003 (**ISO-TC211 2003**), 19115-2:2009 (**ISO-TC211 2009**) and ISO 19139:2007 (**ISO-TC211 2007**) that are also currently taken into account in the most commonly used geospatial data products and services metadata standards. GeoViQua is in the process of developing a Quality Model in which quality encoding would propose an extension of quantitative measurements based on QA4EO approach, as well as the inclusion of the non quantitative quality components. To this end, UncertML contributions (http://www.uncertml.org/) are being integrated together with some other contributions (**Yang et al. 2011**, in preparation). In addition, GeoViQua's Quality Model is complemented with a feedback model that might comprise likely rating, usage report, intercomparison, expert review, external feedback, metadata overriding and GEO Label items.

As will be stated in the next sections, the implementations of certain quality components have not been so far foreseen in the standards and indeed are hardly implementable within the current Earth Observation systems architecture. Moreover, the relevance of user feedback was pointed out in the EuroGEOSS "Advancing the vision for GEOSS" congress (Madrid, Spain, January 2012), where some initiatives related to the GBIF catalogue were shown as an example of enactment of user ratings.

1.5 *Existing/Scanning Quality encodings*

Currently, several standards are used in geographic information data documentation which is realized by filling metadata records. Thus, ideally, defining features of the data products and services are explicitly recorded within a standardized encoding structure. For geospatial datasets, the most widely used standards include the Dublin-Core Metadata Initiative (DCMI), the Federal Geographic Data Committee (FGDC CSDGM) and, predominantly in the GEOSS GCI, the ISO 19000 series standards. In particular, the ISO 19115:2003 (**ISO-TC211 2003**) is related to data products, the ISO 19119:2005 (**ISO-TC211 2005**) refers to services, and the ISO 19139:2007 (**ISO-TC211 2007**) contains a specific implementation in XML format. This format and some others are supported by the GCI architecture enabling the extraction of metadata records from the Clearinghouse. The standards records include quality elements organized in a hierarchic model.

GeoViQua has carefully revised the standards and detected gaps in quality assessment. For instance, contributions from UncertML and other statistical suggestions as regards quality information documentation in metadata records were already discussed in Deliverable 7.1, and other activities in the project. These are being implemented in a proposal of a quality information model under development in WP3 and WP6 that will account for the identified breaches by embedding quantitative and qualitative quality statements. Other issues are related to harmonization, multidisciplinary and multilingual interoperability, implementation and architecture supporting. The definition of standards, thesauri, protocols and development of supporting tools is a hot topic highlighted in the EuroGEOSS "Advancing the vision for GEOSS" (Madrid, Spain, January 2012). Emphasis on quality awareness in the system was pointed out by GeoViQua participants. On the whole, efforts should be coordinated. The derived actions should equally take into account the dynamic evolution of geospatial data products, discovery and retrieval services, interactive user feedback and interfaces, features encoding spectrum and the embracing engine in GEOSS infrastructure, including visualization, search and intercomparison components.

1.6 *Ongoing activities*

In the third Project Technical Board meeting held at University of Reading facilities in Reading (United Kingdom) GeoViQua activities were revised in diverse split sessions in which partners reported in progress made and planning for the ensuing stages, coordinating efforts and interactions. Previous and subsequent telecons reinforced the links between the different work packages, represented in the GeoViQua components model (**Figure 6**) and in practice translated into prototypes of quality-aware search, visualization, intercomparison, elicitation and encoding components. A core element is the Quality Model.

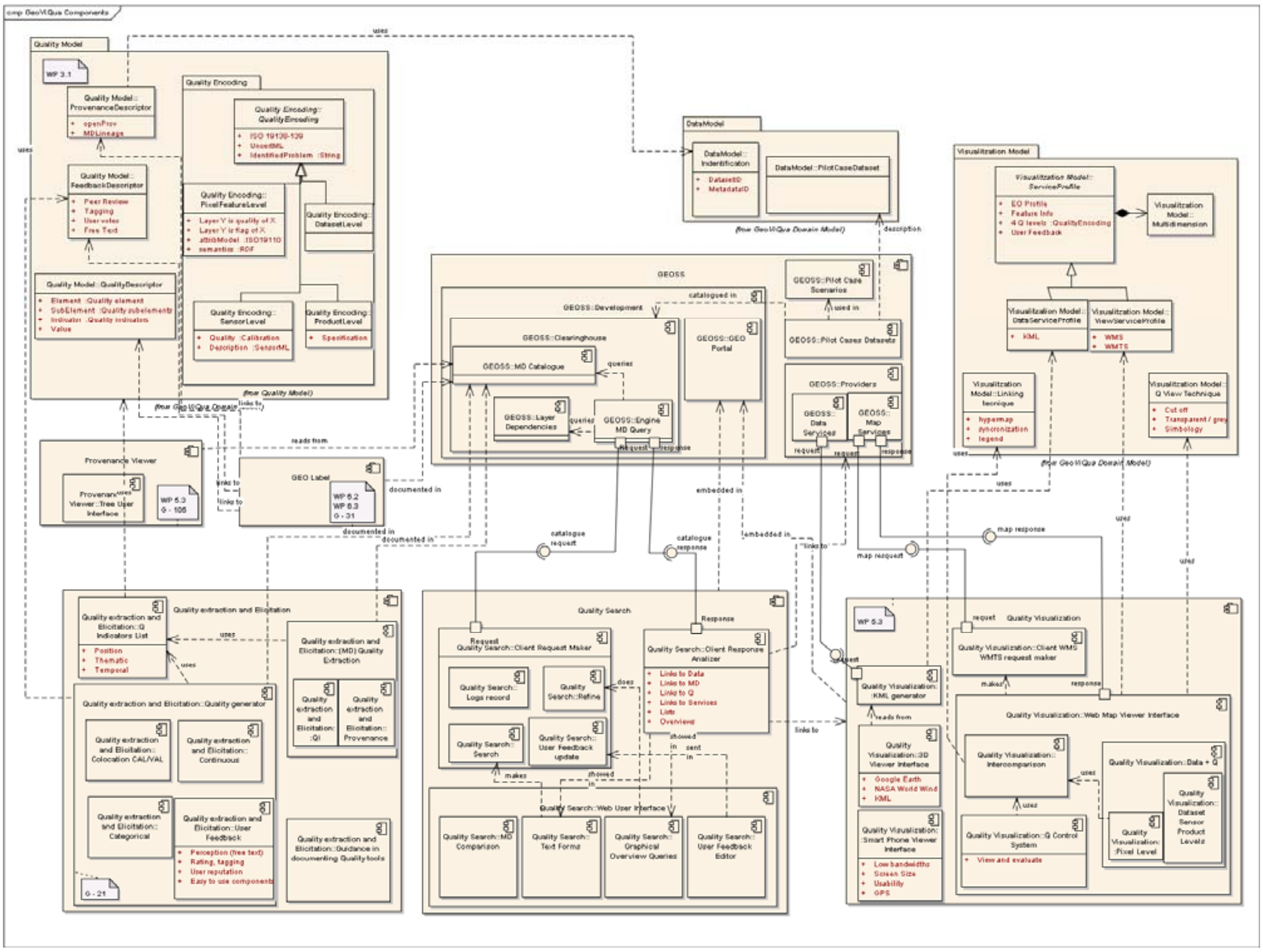


Figure 6. GeoViQua components

In the scope of GeoViQua, other activities are being considered, in for instance the response to the call for participation in the AIP-5, participations in GEO tasks, other possible collaborations in OWS-9, QA4EO tutorials and discussions. GeoViQua partners are actively contributing to several Earth Observation working groups, discussions, and promoting dissemination of results (*e.g.,* EuroGEOSS conference, UncertML, QA4EO meeting, and others).

Related to the issues in this deliverable, the feasibility of a methodology for massive datasets and quality information analysis in GEOSS is being evaluated, including consideration of formats and interoperability, data, metadata and products harmonization, in line with the activities started in AIP-3, associated to GEOSS architecture.

Networking is considered essential in the rapidly evolving GEOSS structures and coupled projects. Moreover, the recent EuroGEOSS "Advancing the vision for GEOSS", although not completely focused on quality so far, provided a discussion platform of relevant topics directly affecting GeoViQua, such as semantics, interoperability, cataloguing, harvesting, multidisciplinarity, metadata recording benefits, to name a few. Participants were presented with the results so far obtained in the exploration of the GCI Clearinghouse quality information and in several discussions the concern for quality assessment and retrieval (*e.g.,* quality-aware WMS, WPS, WFS, WCS) was expressed.

Users and providers interfaces, including adequate feedback mechanisms were also highglighted.

At present, several deliverables are about to conclude, covering User requirements (D2.1 User requirements document, D2.2 Technical requirements document), exploration of quality components (D3.1 Metadata extraction quality component and D7.2 Report on current GCI and identification of issues in integration with GEOSS). In addition, significant progress has been made in the Quality Model component development, to which a Feedback Model is likely to be connected.

This document describes:

-Framework of WP7 in GeoViQua (section 1)
-GEOSS infrastructure and some numbers (section 1)
-User requirements input (section 1)
-Quality assessment and tracking (section 2)
-Quality standards and quality measurements (section 2)
-Quality extraction from metadata (section 3)
-Results of the exploration of quality information in the GCI Clearinghouse (section 4)
-Discussion and corrective measures (section 5)
-Conclusions (section 6)

A common glossary has been annexed in order to ease understanding of frequently employed terms as defined in the context of GeoViQua as a whole, and thereby appearing in this document (also acting as acronyms list). In this sense, a detailed GeoViQua dictionary (http://twiki.geoviqua.org/twiki/bin/view/GeoViQua/GeoviQuaTerminology) is being implemented in GeoViQua's site, which can also be consulted.

2. QUALITY APPROACH AND ENCODING

2.1 *Quality approach and tracking procedures*

Earth Observation data sources are ideally created following quality assessment procedures, resulting in quality estimates and other related indicators. When this occurs, what is called "producer quality information" is generated. This kind of quality information is needed to allow users comparing and deciding about data fitness for a purpose.

However, in practice, systems providing methods to distribute, show and exploit this producer quality information in a standard and interoperable way are rarely used.
Moreover, producer quality indicators must be combined with user feedback and other soft knowledge about the usability of datasets. In this sense, GeoViQua has carried out a complete study on user requirements that, surprisingly or not, has provided relevant conclusions: users demand not only quantitative quality measurements, but also require other qualitative quality components, such as provenance traceability, references, experts judgement or peer-review ratings.

Thus, GeoViQua is in the process of providing a quality metadata model (developed in task 6.1 jointly with WP3) in order to be considered as an extension to the current metadata model in ISO 19115:2033 (**ISO-TC/211 2003**), with the purpose of making them available to users and experts, producing more reliable studies about Earth systems and their dynamics, and in the end including the possibility of tagging spatial information by means of a specific quality label: the GEO-Label.

Within GeoViQua, information about data quality is extracted from metadata (GCI and EUROGEOSS Broker sources), compilation of pilot cases datasets, validation processes with in-situ sensors, provenance information and user feedback. In this particular deliverable, the demonstrated results refer exclusively to the analysis of the quality information currently included in GEOSS clearinghouse metadata. The corresponding deliverables and results-dissemination activities will account for other progress made within the project in this respect.

Partners in GeoViQua agree that a thorough control of the testing dataset is essential for the success of the tasks to be undertaken. This conclusion is extended to scientific work with geospatial datasets. This is to say, a thorough understanding of the data peculiarities, including provenance, traceability and, if available, a certain control of associated quality measures must be carefully watched in the error propagation chain assessment. It should definitely be initially granted by data providers in control of data generation in first instance. Without this input, obtaining solid results and a comprehensive validation in the testing of datasets and associated quality data might not be feasible.

The next section describes the basic characteristics of the quality model specified in the standards. As aforementioned, some of the elements that are demanded by users are difficult to implement in such a metadata structure: for instance, feedback, or calibration/validation results that are generated after metadata creation and harvesting. The flow between users/clearinghouse/user providers should be well established, easily accessible and clearly defined in order to avoid confusion or ambiguous semantics that would lead to misunderstandings. The EuroGEOSS Broker has to account for multidisciplinary standards, components and services integration, semantics and tagging, and (of special interest for GeoViQua), quality elicitation, queries, accessibility and visualization.

2.2 Standards description

Metadata can be documented following diverse standards, which in addition are in constant evolution. This poses great challenges to users not familiar with metadata styles, data providers, and, inside GEOSS, the harvesting processes infrastructures. This alone explains part of the difficulty of carrying out massive data analysis.

An immediate corrective measure would be the generalization of a Quality Model. In fact, GeoViQua is directing efforts in a design of such a Quality Model that would take into account quality and visualization issues, interoperability and user requirements.

In particular, these efforts are focused in the Quality Model component, led by WP3 and WP6, User requirements, but also receiving inputs from other workpackages: WP4, WP5 and WP7 in task 7.1.

In our view, standards are needed and useful because a standard facilitates implementation and provides obvious advantages, allowing for communication and comparison, minimizing data integration, permitting data operations, and other applications. Nevertheless, multidisciplinary standards are implicitly diverse, and this is unavoidable within the SBAs and CoPs in GEOSS. Moreover, as already mentioned, standards are dynamic, necessarily subjected to revisions and therefore in constant evolution. This fact must be taken into account by the GEOSS GCI, and more specifically by its components.

As regards standards, metadata contents of three standards can be harvested directly by the Clearinghouse: DCMI, FGDC CSDGM, and ISO. More specifically, the ISO series related to geographical information is the 19000. Among this series, ISO 19115:2003 (**ISO-TC211 2003)** describes a UML metadata model, with elements, entities and sections. While it is developed as a model, ISO standards are open enough to still accepting diversity of formats in the development of the metadata itself.

In this sense, ISO 19139:2007 (**ISO-TC211 2007)** contains the specifications for writing or translating the UML model in ISO 19115:2003 (**ISO-TC211 2003)** in XML format. In other words, ISO 19139:2007 (**ISO-TC211 2007)** can be regarded as the realization of ISO 19115:2003 (**ISO-TC211 2003)** in a XML format. Thus, other formats (*e.g.*, excel files, external documents) must be converted to ISO19139:2007 (**ISO-TC211 2007)**. In brief, it could be stated that full metadata formats in XML are recognized.

Also relevant to our study, **ISO 19119:2005** (**ISO-TC211 2005)** is the specific development of the ISO standard for Services.

In this concrete study it was decided to fully analyse the ISO quality information, as this is the schema followed by the common data model adopted in the GCI, and consequently in the EuroGEOSS Broker. Nonetheless, surprisingly, metadata records not in accordance with ISO have not been found. The current metadata quality extraction is focused on a better fitness for the future structure of the GCI: the Quality Broker will be the component operating in that sense in the GEO Portal. Other tasks, such as task 3.1, are fulfilling the review of other standards and formats.

Structure of the XML model in ISO 19115:2003

Some basics on the structure of the XML model of ISO 19115:2003 (**ISO-TC211 2003)** will ease comprehensiveness of the methodology and results in the present study. In this UML metadata model, the schema is based on a hierarchic level structure in which elements hang between each other into lower levels in which classes and descriptors are defined.

The analysis of the metadata content is based on the data quality **"DQ_DataQuality"** which is hanging from the **MD_Metadata**. This is to say, all the quality elements and measurements and other aspects we are analysing should be contained in the ISO 19115:2003 (**ISO-TC211 2003**) metadata and thus we should be able to extract this information from the metadata records in xml and extend the analysis to other catalogues.

Areas of analysis within this structure are:

-**Scope**. Following the scope schema in ISO 19115:2003 (**ISO-TC211 2003)**, it is referred to Lineage and DQ_Elements. Within its class, scope includes level (scope code), extent and level description. The first one, scope code is the most relevant for task 7.1 in GeoViQua.

-**Lineage**. It must be remarked that in the latest versions of ISO 19115:2003 (**ISO-TC211 2003)** and its updated version (*i.e.,* extended versions 19115-2:2009, ISO 19157), lineage is not hanging any more from DQ_Data Quality. Instead, it directly hangs from MD_Metadata quality, and thus LI_Lineage is at the same level as DQ_Data Quality. Maybe this fact eases the implementation of a Quality Model including a “non quantitative quality component”, the so-called “soft knowledge” currently under discussion in GeoViQua.

-**Report** is expected to provide quality elements, including quality measurements, and in lower levels, the corresponding quantitative and qualitative values, quality evaluation and results. A detailed insight into results shows that there are two kinds of results: conformance and quantitative results. Conformance results refer to statements of having passed (or not) a certain quality evaluation test which the data provider considers appropriate enough to demonstrate the fitness for purpose of the dataset. Quantitative results of quality measurements can be described as the statistic error or as a measure description, both preferably with an assigned value and units to allow for comprehensiveness. There are five quality classes in ISO 19115:2003 (**ISO-TC211 2003**): Positional accuracy, Completeness, Logical consistency, Temporal accuracy and Thematic accuracy.

As will be seen in the discussion section, these classes and this quality component contemplated in the standard quality indicators are also subject of other GeoViQua’s tasks (*e.g.,* revision in Deliverable 7.1, tasks in WP3 and WP6 related to quality elicitation and encoding, Quality Model, including UncertML and quality vocabulary).

-**Usage:** The entity MD_Usage explains the specific uses of the data. It is embedded in the ISO including several items that GeoViQua’s surveys identified as relevant for users. For instance, MD_Identification/resourceSpecificUsage, which relates to the specific applications for which the resource was used or MD_Usage/specificUsageuserDeterminedLimitations. The latter is a very useful item although due to the fact of exclusively enabling a free text domain it is in practice hard to use in an automatic evaluation (**Růžička 2008**). In addition, usage specifies resource constraints, yet still uses a free text domain which is difficult to evaluate automatically.

The following figure extracted from the ISO 19115:2003 (**ISO-TC211 2003)** standard shows the hierarchic metadata structure and in particular the section in which the quality information extraction is based:

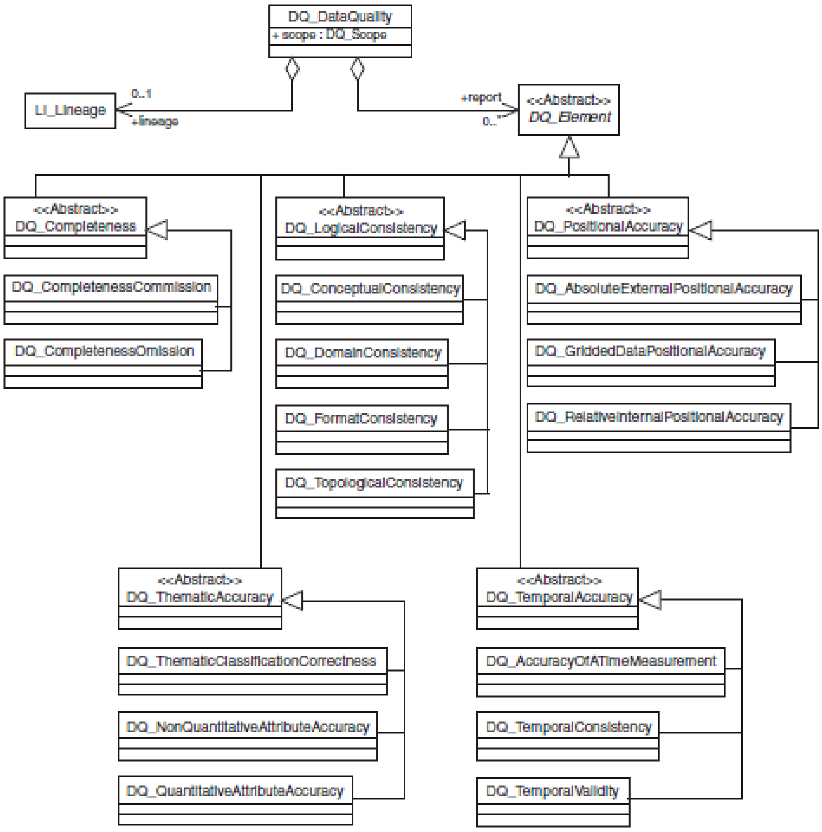


Figure 7.

Note the quality indicators (*i.e.,* completeness, thematic accuracy, logical consistency, temporal accuracy, and positional accuracy) hanging under the epigraph DQ_Data Quality/Scope. Scope is then divided into the Lineage section that will be detailed later in this section, and the DQ_Elements, containing the quality indicators, and elements. In a lower level, not shown in this figure, quality measurements are included, all of these inheriting from DQ_elements and the corresponding quality indicators and elements.

In higher levels of this hierarchic structure we have considered the possibility of having two DQ_DataQuality entities, which can represent the complete consideration of all possible results. In lower levels, we have considered up to 16 quality indicators and, for each quality indicator, the first measure method is considered, and the corresponding result can in turn be expressed by a numerical value, a conformance declaration with a methodology, or even a per pixel value (*i.e.,* a coverage grid instead of an overall result, explained in ISO 19115-2:2009 extension –**ISO-TC211 2009**-). In this sense, it must be noted that this per-pixel value is a highly desired feature in quality information, especially for Earth Observation data, as has been identified in the user requirements and expert usage analysis already performed in GeoViQua and external surveys (*e.g.,* users interviews, GEO Label questionnaire).

Finally, in terms of user requirements, the standard specifications contain a MD_Usage class. This element is intended for a brief description of ways in which the product is being used or has been used. However, user requirements fulfilment might need a mechanism that works the other way round. This is to say, an implementation of user feedback, not just producer records, where potential or real users, can express their opinion based on experience. However, it's neither necessary nor appropriate to include this kind of information in the metadata records. For instance, citation is already embedded in the standard, but as a definition of how the data product should be cited when used as a data source. In contrast, user requirements about citation include how many times ,when and where the dataset has been cited by an acknowledged source. This means a complementary sort of citation referring to specialized journals, peer-review, expert judgment and releases. In other words, a type of user feedback in a direction that is not envisaged in the current standard model, at least in metadata content post-edition.

Thus, our exploration of the quality indicators does not include such elements. As a relevant conclusion after analysing the user requirements in GeoViQua (see the summary poster presented in the QA4EO in Oxford, related to the so-called quality soft knowledge shown in the introduction) and the inputs from WP2 activities, the mismatch between these and the quality statements provided by standards like ISO 19115:2003 (**ISO-TC211 2003**) should be highlighted. In this sense, as was aforementioned, there is a high demand for qualitative quality components. The analysis carried out for Deliverable 7.2 has different facets. It would be interesting to quantify the presence of these non quantitative elements in the Metadata XML structure when available and implemented as required; the target of proposing a Quality Model in which user requirements and feedback would be quantitatively and qualitatively addressed should not be missed.

2.3 Quality indicators and measures

Quality measures analysed are supported by ISO standards, although within GeoViQua the Quality Model under development is likely to be complemented with statistical descriptions which conform to the UncertML dictionary. This action is recommended after revision of results of activities in WP2, related to user requirements identification associated to quality information (*e.g.,* Deliverable 2.1 and 2.2 connected to User Requirements for GeoViQua).

In parallel, the study carried out for this deliverable explores the real content of quality data in the GCI available within the current architecture. In the end, the comparison between user requirements and up-to-date quality information content is a valuable input to the Quality Model to be developed in GeoViQua. This effort ideally will result in a major contribution from our project to the GEOSS infrastructure. Translating these issues to GeoViQua's tasks, we are referring to the 'Quality elicitation mechanisms' and 'Delivery of solutions to end users'. Special attention has been paid on this issue, as it constitutes a main pillar in GeoViQua.

First, as was explained in the previous section of this document, lists of typified common quality parameters (defined as "quality elements" in ISO standards –ISO 19113:2002 **ISO-TC211 2002**-) for geographical datasets were revised.

In turn, quality parameters can be quantified using several measures (*e.g.,* root mean square error, error rate, etc.). Following ISO standards, quality parameters and quality indicators correspond to elements and sub-elements, extracted from ISO 19113:2002 (**ISO-TC211 2002**).

Quality measures in this context apply to the actual quantification (*e.g.,* root mean square error, value at 95% confidence level) of quality parameters or indicators, obtained from ISO19114:2003 (**ISO-TC211 2003**) and ISO19138:2006 (**ISO-TC211 2006**). Thus, these three concepts related to quality elements comprise a hierarchy, from quality parameters (aspects), to quality indicators (sub-elements), and finally quality measures (test applied to evaluate data quality elements).

Table 1 summarizes the results of the revision of the abovementioned ISO standards, including some complementary concepts introduced in GeoViQua.

Table 1. Classification and definition of quality parameters and indicators

Quality parameter	Quality indicator	Definition
Completeness	Commission	Excess data present in a dataset
	Omission	Data absent from a dataset
Logical consistency	Conceptual consistency	Adherence to rules of the conceptual schema
	Domain consistency	Adherence of values to the value domains
	Format consistency	Degree to which data is stored in accordance with the physical structure of the dataset
	Topological consistency	Correctness of the explicitly encoded topological characteristics of a dataset
Positional accuracy	Absolute or external accuracy	Closeness of reported coordinate values to values accepted as or being true
	Relative or internal accuracy	Closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true
	Gridded data position accuracy	Closeness of gridded data position values to values accepted as or being true
Temporal accuracy	Accuracy of a time measurement	Correctness of the temporal references of an item (reporting of error in time measurement)
	Temporal consistency	Correctness of ordered events or sequences
	Temporal validity	Validity of data with respect to time
Thematic accuracy	Classification correctness	Comparison of the classes assigned to features or their attributes to a universe of discourse (*e.g.,* ground truth or reference dataset)
	Non-quantitative attribute correctness	Correctness of non-quantitative attributes
	Quantitative attribute accuracy	Accuracy of quantitative attributes

These quality parameters are very typical of geospatial data sets, usually conforming important fragments of the metadata, and allowing users to judge whether the data set is fit for their purpose. However, there is a bias towards vector data within the list.

In general it is expected that all data sets should be logically consistent - if not, it would not make sense to use them. In this sense, there is a difference between discovery time and propagation of the information in data processing. Thus, the difference between quality information required in dataset discovery and evaluation, and the information required during the use of the dataset should be stated clearly.

With regard to the associated quality indicators, it must be noted that these indicators have a singular utility in particular scenarios. For example, if two datasets are to be used in a single project, being aware of only their respective internal accuracy is not sufficient, since the fact of how they are related remains unknown. In that case, it is not possible to combine them in a principled framework. Therefore, it is worth noting that accuracy measures reporting “the 'closeness' of a value to that which is accepted as being true (reality)" are the most useful, since these allow true information interoperability in the sense that the data can be combined with other data, processed and integrated into a coherent decision making framework.

On the current ISO standards, it must be remarked that the ISO 19138:2006 (**ISO-TC211 2006**) is being revised. The ISO 19157 (**ISO-TC211**), currently under development, is likely to provide updated insights in this context.

The quality measures presented above are a small subset of the possible approaches to representing quality of data. GeoViQua is considering the extension of the standards quality model with concepts such as contingency tables, which rely on a whole family of indicators (similar to kappa statistic), or the true skill statistic (**Liu et al. 2011**). The target is that all quantitative, categorical and binary data, are covered by appropriate quality measures. Quality assessment activities are a principal point in the scenario flow diagrams, and these activities will include a diversity of facets, such as quality indicator estimation, recording and propagation at a variety of stages from data providers to user accessibility.

This topic generated complex discussions in GeoViQua, leading to a fusion of ISO standards with solutions to address statistical concerns, agreed in the second Project Technical Board meeting, held in ESA-ESRIN (July 2011 Frascati, Italy). Tasks are being coordinated with WP2, WP3 and WP6, and the corresponding deliverables will further extend this section. For example, deliverable “D6.1 Data Encoding as a best practice paper” represents an updated version of these efforts.

Remarkably, despite the relevance of associated quality to the use of datasets, and the demonstrable improvement in accuracy of results, quality parameters are frequently unknown and not explicit in the metadata. This study analyses facts that shed some light on the issue and eased the establishment of guidelines for several WP. Although the exact causes for the information gaps have not yet been identified, GeoViQua proposes solutions and corrective measures, based on the conclusions of the exploration of the Clearinghouse quality information contents, and the analogous on-going work on other Community Catalogues accessed through GEOSS (under WP3 tasks). From the adoption of QA4EO quality tutorials, and the generalization of the GeoViQua’s Quality Model, together with the GeoViQua quality-aware tools linked to enhanced search, visualization, intercomparison, user feedback, and quality measurements extraction, all areas in GEOSS are tackled.

3. QUALITY EXTRACTION FROM METADATA

3.1. Recollection of metadata: preliminary tests

The clearinghouse metadata content up to November 2011 was automatically extracted and analyzed in terms of their quality information. The clearinghouse follows the Open Geospatial Consortium OGC CSW standard specification, whereas the metadata retrieved follows the ISO 19115:2003 standard (**ISO-TC211 2003)** and is thus encoded in the ISO 19139:2007 standard (**ISO-TC211 2007),** this is to say translating the ISO 19115:2003 (**ISO-TC211 2003)** specifications into XML format files.

One benefit of limiting analysis to standardized specifications is the feasibility of the application of a semi-automatic methodology which had been previously tested on the quality information of the metadata records catalogued in the Spanish regional Spatial Data Infrastructures SDIs (**Díaz et al. 2010**).

In the first stage of the analysis we attempted to carry out a subsampling trial, intended to obtain detailed results by splitting the metadata records into subgroups of data and services. This first classification will be explained in detail in the next section (3.3). Additionally, a secondary analysis was carried out in order to determine the presence or absence of different standards. Searching for the specific word “MD_Metadata” in all the documents, results showed a 100% of metadata records responding to ISO schemas. Therefore the study centered the quality extraction in ISO schemas, which is moreover the schema followed by the EuroGEOSS Broker.

Splitting the sample by standards was unsuccessful due to the lack of truly distinctive keywords for each standard to be applied to a massive sample, moreover not always containing correctly filled in metadata. In order to do these, several xml have been reviewed from which to find certain identifiers for each subsample. For instance, searching “SV_ServiceIdentification” would separate ISO 19119:2005 (**ISO-TC211 2005**) xml from the rest and this string would be considered an ISO 19119:2005 **(ISO-TC211 2005**) standard identifier. Even though the subgrouping was rejected due to the lack of certainty and the high probability of overlapping between subsamples, this task permitted us becoming familiar with the xml contents and provided useful tips for the following stages.

In the end, taking into account the difficulties in the design of an adequate computational methodology, the identified gaps in quality encoding mechanisms and system retrievals from data producers to user feedback it can be concluded that on the whole, massive analysis is a tough issue, and thus it is logical that the subsampling trials did not provide certain outcomes. Moreover, if the methodology is to be extended to the increasing databases numbers in the GCI, subsampling should be discarded at least in the scope of GeoViQua. This circumstance could not have been foreseen in the initial DoW. Considered the huge quantity of metadata records currently available in the GCI, a massive downloading of XML files is not at all recommended: instead, we would suggest the harvesting of the information through CSW standard and XPath language, exploiting the EuroGEOSS Broker which acts as a middleware technology.

3.2. Selection of quality elements

The next decision focusses on the extracted **quality contents**. It is noteworthy that, in the end, the automatized methods mainly produced results related to data products encoded following the **ISO 19115:2003 (ISO-TC211 2003)** standard. As has been explained, the ISO 19115:2003 (**ISO-TC211 2003)** has a clearly defined UML structure and therefore Xpaths can be tracked to obtain massive extraction of metadata quality elements.

According to the ISO 19115:2003 structure (**ISO-TC211 2003**), services can be introduced following this standard. Actually there is an entity MD_ServiceIdentification hanging from MD_Metadata and MD_Identification, analogously to the MD_DataIdentification. This implies there is a way of documenting services within ISO 19115:2003 (**ISO-TC211 2003**). Fortunately, this is evolving into the specific ISO for services 19119:2005 (**ISO-TC211 2005**), although this is not a generalized practice. In addition, the coexistence of alternative and likewise correct paths for defining services metadata is to be noted. Data and services are separated in the metadata records of XML using the quality scope. In order to quantify the approximate representation of metadata records related with different scopes, differing from datasets, we collected the scope of the quality, in which the services represent a very low result (*i.e.,* 617 records representing 0.63%).

Thus, our metadata records deal strictly with data products (not services) as the extraction of the quality information was carried out by choosing a direct XPath guaranteeing only access to data products under ISO19115:2003 (**ISO-TC211 2003**). Services in our metadata records should be documented under ISO 19115:2003 (**ISO-TC211 2003**), and accordingly it was assumed that services would be still linked to data.

Within this standard, the sections that have been analysed were Data quality Scope, Lineage (LI_Lineage) and Data quality reports (containing quality elements and measures) and usage information (MD_Usage).

As mentioned above, quality indicators are conceptually classified in completeness, logical consistency, positional accuracy, thematic accuracy and temporal accuracy. In lower levels of this hierarchic structure, for each quality indicator, one or more measure method are considered, and the corresponding result can in turn be expressed by a numerical value, a conformance declaration with a methodology, or even a per-pixel value (*i.e.,* a coverage grid instead of an overall result). In this sense, it must be noted that this per-pixel quality information values are highly desirable at the stage where the data is used , as identified in the user requirements and expert usage analysis already performed in GeoViQua and external surveys (*e.g.,* users interviews, GEO Label questionnaire).

3.3 Methodology: structuring metadata XML contents into databases

The massive extraction of quality elements, including provenance and usage, from the metadata records in the GEOSS GCI Clearinghouse was done by means of the following methodology, drawn in **Figure 8**:

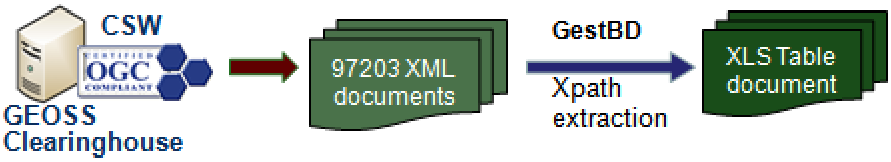


Figure 8. Flow diagram of the extraction of metadata from the Clearinghouse and generation of a database of selected items relevant to quality data.

This method consists of a massive downloading of the metadata XML files contained in the clearinghouse, harvesting the metadata records by repetitively requesting CSW standard XML files.

The second step (section 3.2) consisted of selecting the precise information required for the analysis, taking into account the relations amongst entities, the multiplicity and the different ways in which the information can be described (for instance the lineage information). Afterwards, twenty-five different quality elements were selected for the extraction, considering up to 16 multiple results for each quality element, twenty lineage elements were selected for the extraction considering up to 8 multiple results and, finally, five usage elements were selected considering up to 8 multiple results for each element.

The next step consisted in the extraction of the information contained in the XML files by request through XPath language sentences. The information is then saved into a database in which the target ISO elements are recorded, organized in columns whereas the corresponding metadata files are organized in rows, identified by their specific number.

Once that is done, the database is analyzed and results are summarized. The total number of GEOSS clearinghouse metadata records harvested was 97203.

Preliminary exploration (section 3.1) of ISO 19115:2003 (**ISO-TC211 2003)** standards specifications and corresponding UML model (as specified in ISO 19138:2006 - **ISO-TC211 2006**) in the XML files was undertaken using XML Spy. At this stage, several documents non compliant with the XML structure were identified.
UltraEdit-32 was used to perform massive queries in the metadata records files to determine whether the downloadad XML were following the ISO 19115:2003 standard (**ISO-TC211 2003)** by means of searching for MD_Metadata entity.

The extraction of files from the Clearinghouse and the ensuing extraction of quality components and measurements, which comprises the main stages in the analysis, were undertaken employing Miramon **GestDB** module. GestDB is a database management tool in Miramon software capable of addressing large database files, with restructuring, combining and comparing functionalities. Actually, a specific command in GestDB was designed in order to easily perform the direct extraction of XPath sentences from XML files and implement an automatic organized storage of the extracted information in a DBF file. Thus, GestDB module, in first instance is used for Xpath extraction from XML documents and resulting in a DBF creation containing the extracted quality elements, indicators and lower levels content in the extracted metadata. This information is organized in the DBF database in which the rows are the XML files, unmistakably identified by a number, and in which the columns contain the corresponding quality elements. Such structure permits an enhanced examination of the information, comparison and statistics calculation, otherwise unfeasible in a massive analysis of XML files in the corresponding hierarchic schema. The final DBF contains 52332 rows and 50 columns.

A detailed insight into the sentences used for the extraction of the information would describe the following: XPath 2.0 is an expression language that allows the processing of values conforming to a data model that provides a tree representation of XML documents as well as atomic values such as integers, strings, and booleans, and sequences that may contain both references to nodes in an XML document and atomic values. The result of an XPath expression may be a selection of nodes from the input documents, or an atomic value, or more generally, any sequence allowed by the data model. (**XML Path Language XPath 2.0 Second Edition** http://www.w3.org/TR/xpath20/)

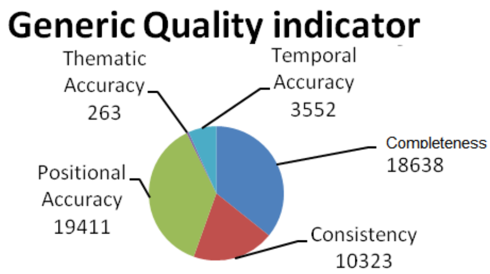
4. RESULTS

4.1 General results

As mentioned above, the overall number of metadata records with quality indicators is 19107 (which represents 19.66% of the total sample, a figure far from the ideal situation explained in section 2.1). To be positive, it must be stated that this number is higher than initially expected, especially when compared to results obtained in previous experiences with other spatial data infrastructures (*e.g.,* IDEC, a regional data infrastructure in Spain) and preliminary trials with the sample in this study.

In turn, the 19107 metadata records contain a total of 52187 quality indicators which results in a mean of 2.7 quality indicators per document. **Table 2** and **Figure 9** show a more detailed analysis of this data, summarizing the quality indicators by classes and subclasses (as classified in Figure A.6 in ISO 19115:2003 –**ISO-TC211 2003**-).

Table 2. Classification of quality indicators in classes and subclasses in the GEOSS Clearinghouse

Generic quality indicator	**Specific quality indicator**	**Number**	**%**
Positional accuracy	Absolute external positional accuracy	17767	34.04
	Gridded data positional accuracy	1364	2.61
	Relative internal positional accuracy	280	0.54
		19411	***37.19***
Completeness	Completeness commission	9815	18.81
	Completeness omission	8823	16.91
		18638	***35.72***
Logical consistency	Conceptual consistency	9454	18.12
	Domain consistency	857	1.64
	Topological consistency	12	0.02
	Format consistency	0	0.00
		10323	***19.78***
Temporal accuracy	Accuracy of a time measurement	2870	5.50
	Temporal consistency	682	1.31
	Temporal validity	0	0.00
		3552	***6.81***
Thematic accuracy	Quantitative attribute accuracy	261	0.50
	Non-quantitative attribute accuracy	2	0.01
	Thematic classification correctness.	0	0.00
		263	***0.51***
	**Total quality indicators**	**52187**	

The results show that, among the generic indicators, positional accuracy and completeness are the ones most widely used, with a quite similar importance (37.19% and 35.72%, respectively), and adding up to a high percent of the total, 72.91%. The third quality indicator in number is logical consistency, reaching around a 20%, whereas the fourth is temporal accuracy, adding up to a 6.81%. Last, not completely unexpectedly, thematic accuracy is below a 1% in the sample of generic quality indicator classes.

Breaking down into specific indicators, not surprisingly, absolute external positional accuracy is the most significant subclass, with a total number of 17767 and explaining most part of the relevance of positional accuracy in the sample of metadata with quality indicators (34.04%). Remarkably, an interesting finding is the 1364 number corresponding to gridded data positional accuracy, which is a much demanded quality indicators both for users and experts in geospatial datasets.

Within completeness, the distribution among subclasses is more homogeneous, as could also have been anticipated (commission with 18.81% in front of omission with 16.91%). Equally rational results the subclasses distribution within logical consistency, in which there is a clear predominance of conceptual consistency, followed by domain consistency. The numbers obtained for topological consistency and format consistency reveal commonly encountered situations: the absence of topological structure in vector layers and a lack of awareness of interoperability. A considerable number of metadata containing conceptual consistency, combined with a very high number of metadata lacking domain, format and topological consistency represent a scarcity of coherence amongst the information declared.

As regards temporal accuracy, accuracy of a time measurement is clearly the dominant subclass, followed by temporal consistency. Instead, temporal validity indicators were not present in the sample. Finally, the thematic accuracy subclass more representative is quantitative attribute accuracy, distinctly in higher numbers than non-quantitative attribute accuracy, whereas thematic classification correctness produced no results in the sample.

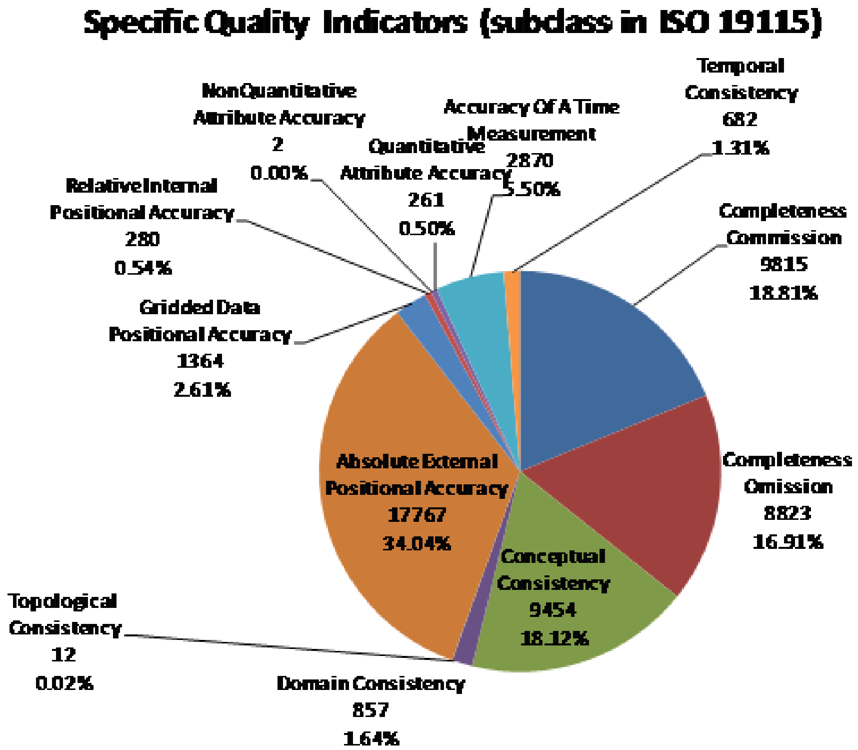


Figure 9. summarizes these numbers enabling a visual comparison of the quality indicators present in the analysed sample, both in numbers and %

Despite the clear predominance of absolute positional accuracy in the quality indicators, the one thing to be highlighted is the diversity of quality indicators used: there is representation of all the 5 generic indicators. These results contrast with the previously published study about the Spanish regional SDIs, in which completeness, consistency and temporal accuracy do not represent more than 5% overall (**Díaz et al. 2010**), while in the GEOSS products analyzed these indicators represent 35.71%, 19.78% and 6.81% respectively.

4.2 Data Quality Scope

The following figure represents the results obtained in the analysis of the scope code list present in the sample. On the left, the UML code list of this feature as specified in the ISO 19115:2003 **(ISO-TC211 2003)** standard is shown. On the right hand side, the summarizing plot of scores in the sample:

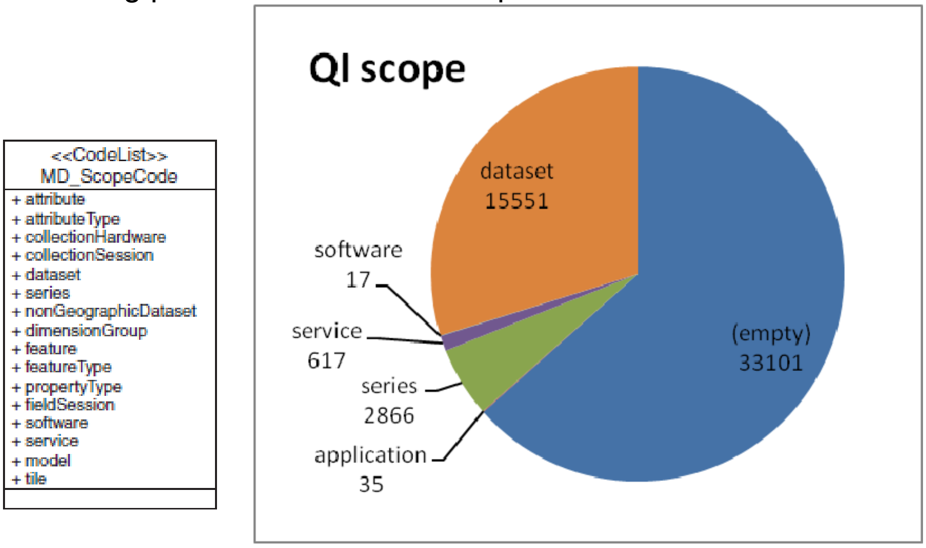


Figure 10. Scope Code list under ISO 19115:2003 (ISO-TC211 2003) and plot of results of the level of detail of data quality from the metadata contained in the Clearinghouse in 2011

The 33101 empty registries (34.05%) in the metadata records of metadata with quality indicators is unfortunately a common finding in studies of this kind. This is explained by the fact that this element is not mandatory in order for the metadata record to be compliant with the ISO standard. Within the metadata with content in this level of the standard specification, “dataset” represents the highest number, with up to 15551 metadata records (15.99%). Other attributes of the scope code list present in the metadata records are series (2866, 2.94%), service (617, 0.63%), “application” (35, 0.36%), and software (17, 0.17%). These results are summarized in **Table 3**.

Table 3. Summary of Scope code results in the metadata records extracted from the Clearinghouse

Scope code	Number of metadata records
(empty)	33101
Application	35
Series	2866
Service	724
Software	32
Dataset	15573
Metatada records with scope	**52331**

Note that “application” is not in the MD_ScopeCode list. Strictly speaking this would be understood as unacceptable under the standard specifications. It could be interpreted as a synonym of software and if so, the results obtained for application considered valid and added to the software category in the results. This could be an example of the semantic topics that need to be addressed in interoperability studies.

4.3 Quality measures analysis

The ISO 19115:2003 (**ISO-TC211 2003**) standard specifications detail corresponding to DQ_results is shown in **Figure 11**.

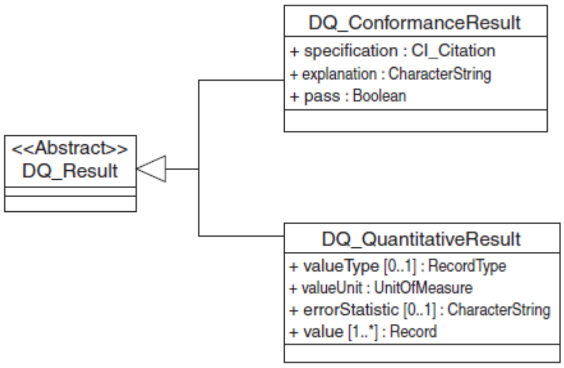


Figure 11. DQ_Results schema under ISO 19115:2003 (ISO-TC211 2003)

This figure explains that quality measures results can be either expressed by a numerical quantitative result or else by a conformance result, which means determining whether the data product is compliant with an acknowledged quality test or specification.

A more detailed analysis in our metadata records, reveals that there are 25944 measures. Sometimes a quality indicator can be expressed in more than one measure. On the other hand, not every quality indicator mentioned has its corresponding measure: for instance, a producer can just describe the quality indicator applied in the dataset without providing any specific results. A measure can be in turn classified in numerical measures (22275-85.86%) and in a conformance declaration to a specification, in the case of the Clearinghouse, the conformance results are mainly referred to INSPIRE flag (3669-14.14%). This can be considered a good sign, implying that providers are following the mandatory directive regarding SDI in Europe.

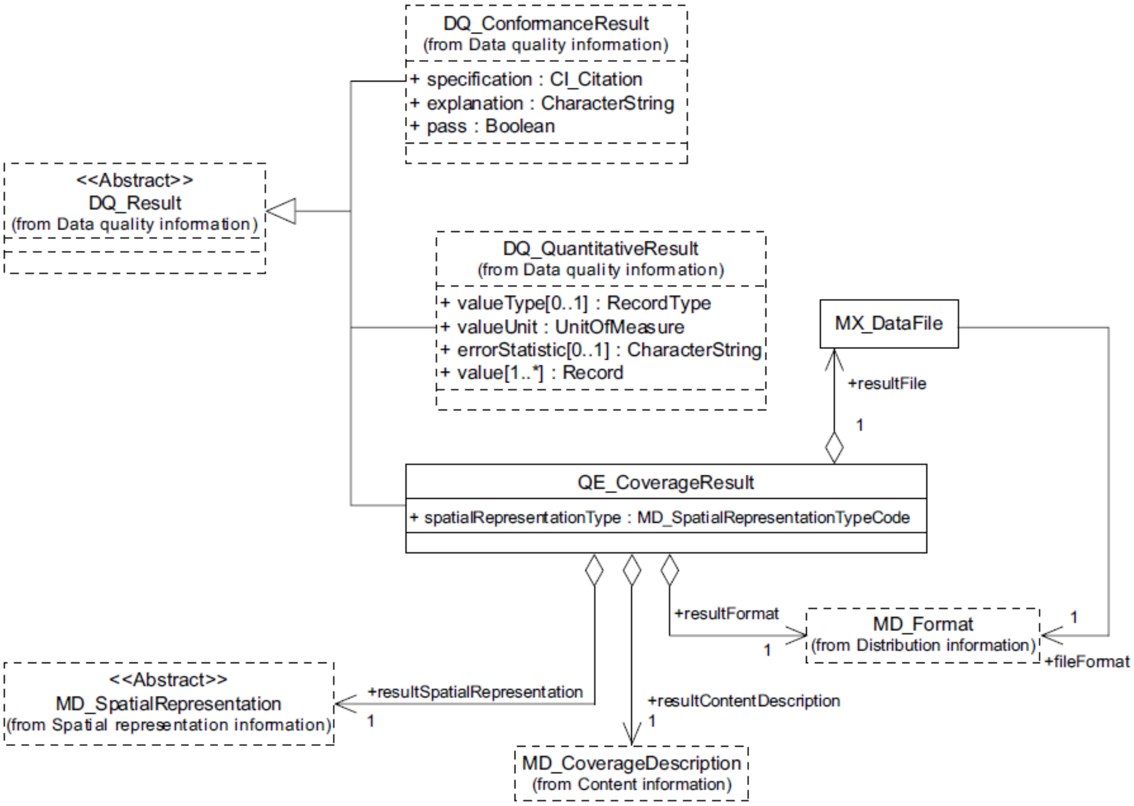


Figure 12. ISO 19115:2003 extension 2 **(ISO-TC211 2009)** for “per pixel” quality

Most interestingly with respect to the Earth Observation datasets as shown in **Figure 12**, the quality measurement results can be expressed in a coverage grid, as specified in the ISO 19115-2 extension **(ISO-TC211 2009)** (5-0.02%), which is highly desirable as was aforementioned previously.

It is to be noted that this extension of the ISO 19115-2:2003 (**ISO-TC211 2003**) has been created after numerous revisions that have been conducted during 2011 by the expert’s metadata groups in ISO. At present, it is an unofficial version of the standard. Indeed, this extension is a very valuable output for gridded data, which in fact is its specific focus. Unfortunately, even though the five coverage results found in the Clearinghouse describe the link to the data file (as it would be the desired result), the link does not work due to the fact of being in a local or non traceable location.

Although a small number in the total picture in the study, the mere presence of these coverage grids is a promising result, showing that first, from the data production level, quality awareness has risen and improved, and second, that the harvesting mechanisms enable accessing the quality information, surpassing several system boundaries and interoperability issues. This fact alone confirms that quality extraction, visualization, intercomparison, and search, as core targets in GeoViQua, can be achieved, and that GEOSS architectural structure is in the right lane towards quality awareness compliance. Nonetheless, an important effort must follow in order to make of this a general practice and significantly increase the frequency of this sound expression of quality results in geospatial datasets.

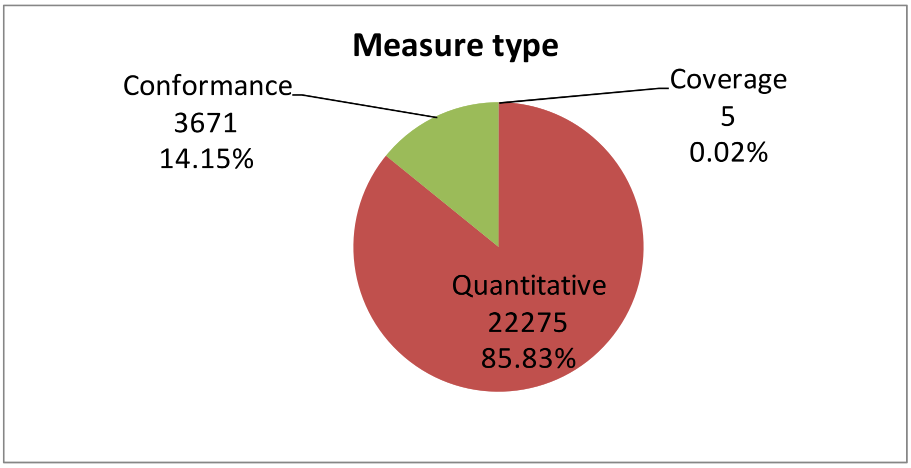


Figure 13. Plot of Results types.

The results related to quality measurements results are summarized in **Table 4**.

Table 4. Summary of DQ_Result, including coverage, quantitative results and conformance result for a quality indicator.

DQ_Result	Number of metadata records	%
QE_CoverageResult	5	0.02
DQ_QuantitativeResult	22275	85.83
DQ_ConformanceResult	3671	14.15
Metatada records with DQ_Results	**25951**	

As for quantitative results, the graph below shows that although there is a quantitative numerical result, in most data products (17325, representing around 77.78%), the units in which the quality measurements quantitative result is provided are missing. Evidently, this is an essential gap identified in the study that requires immediate mending. Not surprisingly, the most common unit of measure is meters (present in 3782 metadata -16.98%-), which is in coherence to positional accuracy being the most frequent quality indicator. Other relevant unit for quality measurements is percentage, representing the 5.24% of the measures.

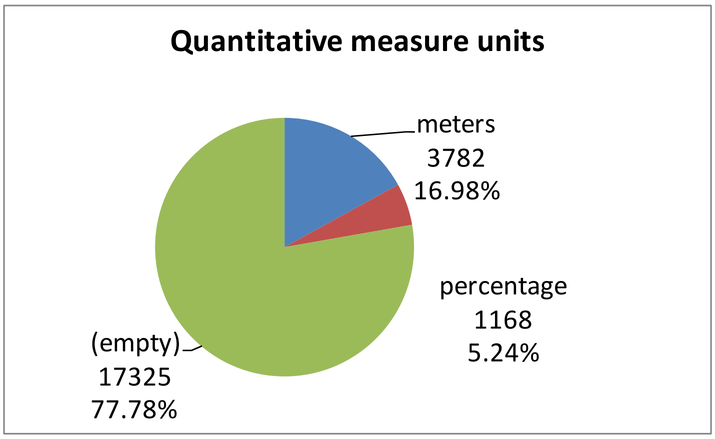


Figure 14. Units in which Quantitative results of quality measurements are given in the metadata extracted from the Clearinghouse

Regarding conformance results, 3361 metadata (91.56%) state the data product is conformant with a certain specification, in contrast with the 128 false statements (3.49%).

One might wonder at the fact that a metadata provider would explicitly state not being compliant with a particular quality assessment methodology, although in any case it is an interesting finding proving quality awareness and attempts of completing the standards. The number of empty files is reduced when compared to the quantitative results.

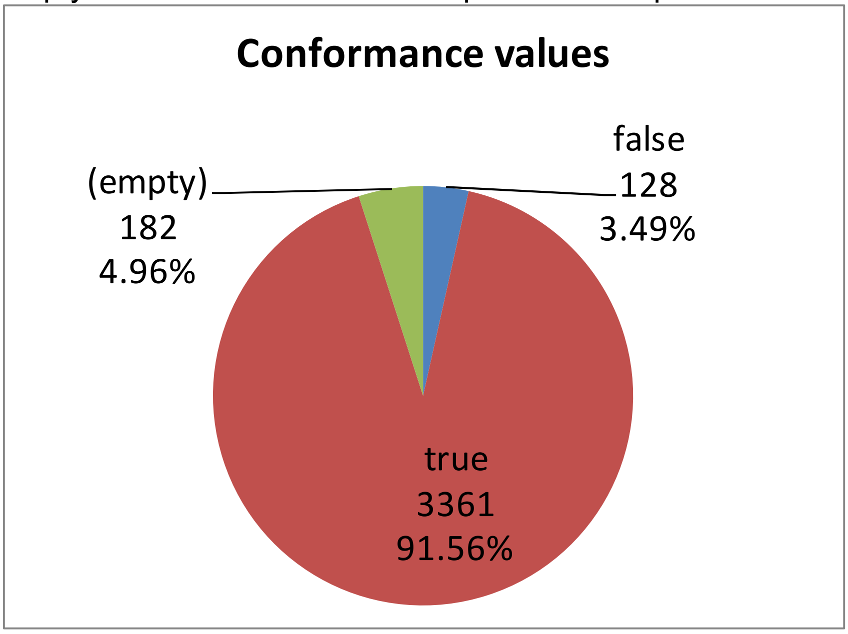


Figure 15. Conformance statements for quality measurements contained in the metadata extracted from the Clearinghouse

In a more detailed analysis related with quality indicators and the subsequent information associated to those -measurement results, units, conformance or related explanations that complement the quality indicator (in fact, all the details explained above)- it was determined whether the quality indicator -quantitative or qualitative)- is well complemented with the required elements to be understandable. The results are summarized in **Table 5** and the graphs bellow, the first for quantitative results (**Figure 16-1**) and the second for qualitative results (**Figure 16-2**).

Table 5.Detailed analysis of quality elements associated to quality indicators.

	TOTAL	Conformance type	Declare conformance	Conformance to specification	Quantitative type	Declare quant. value	Complete quant. value
Absolute External Positional Accuracy	17767	66	66	43	17701	10443	3807
Accuracy Of A Time Measurement	2870				2870	1435	553
Completeness Commission	9815				1081	1081	1077
Completeness Omission	8823	2	2	2	87	87	87
Conceptual Consistency	9454	684	684	684			
Domain Consistency	857	857	677	657			
Gridded Data Positional Accuracy	1364	1364	1364	1364			
Non Quantitative Attribute Accuracy	2	2	2	2			
Quantitative Attribute Accuracy	261				256	255	31
Relative Internal Positional Accuracy	280				280	280	280
TemporalConsistency	682	682	682	682			
Temporal Consistency	12	12	12	12			

Figure 16-1. Graph containing detailed information about quantitative indicators and subsequent information

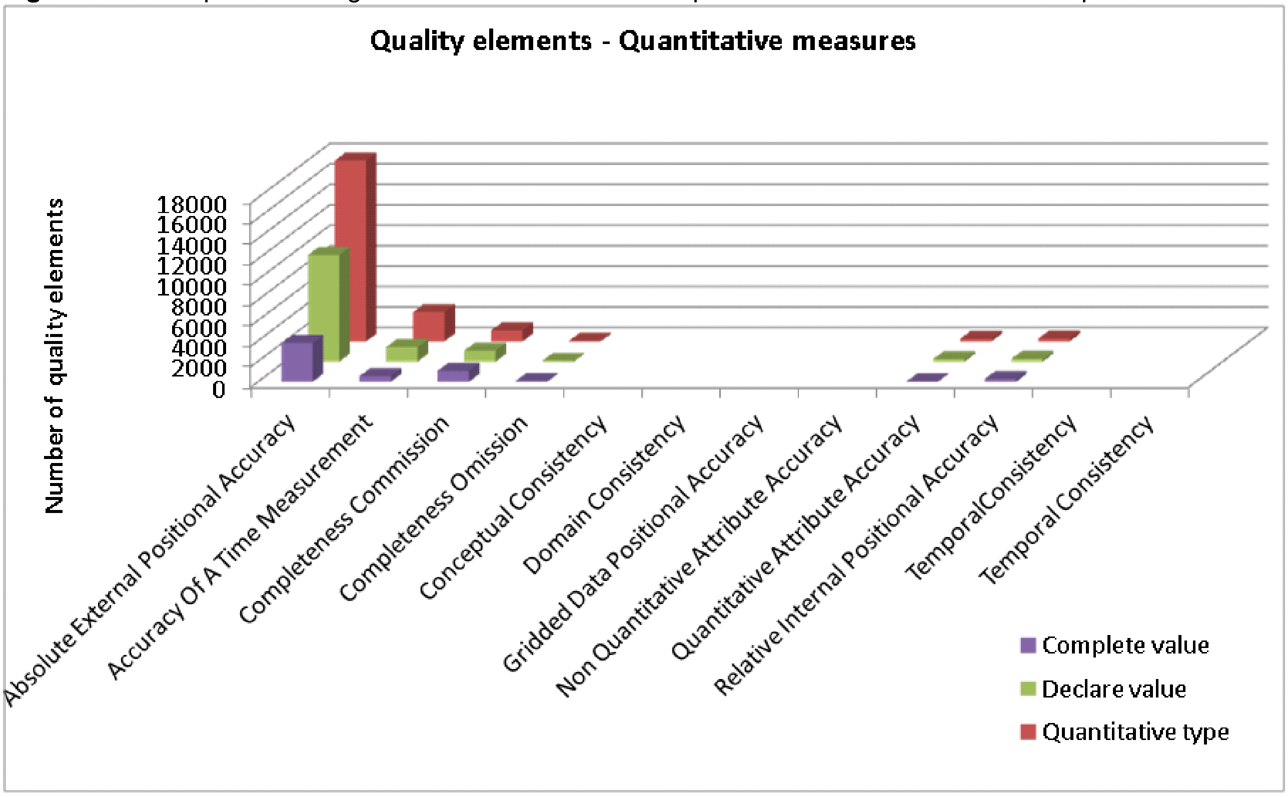
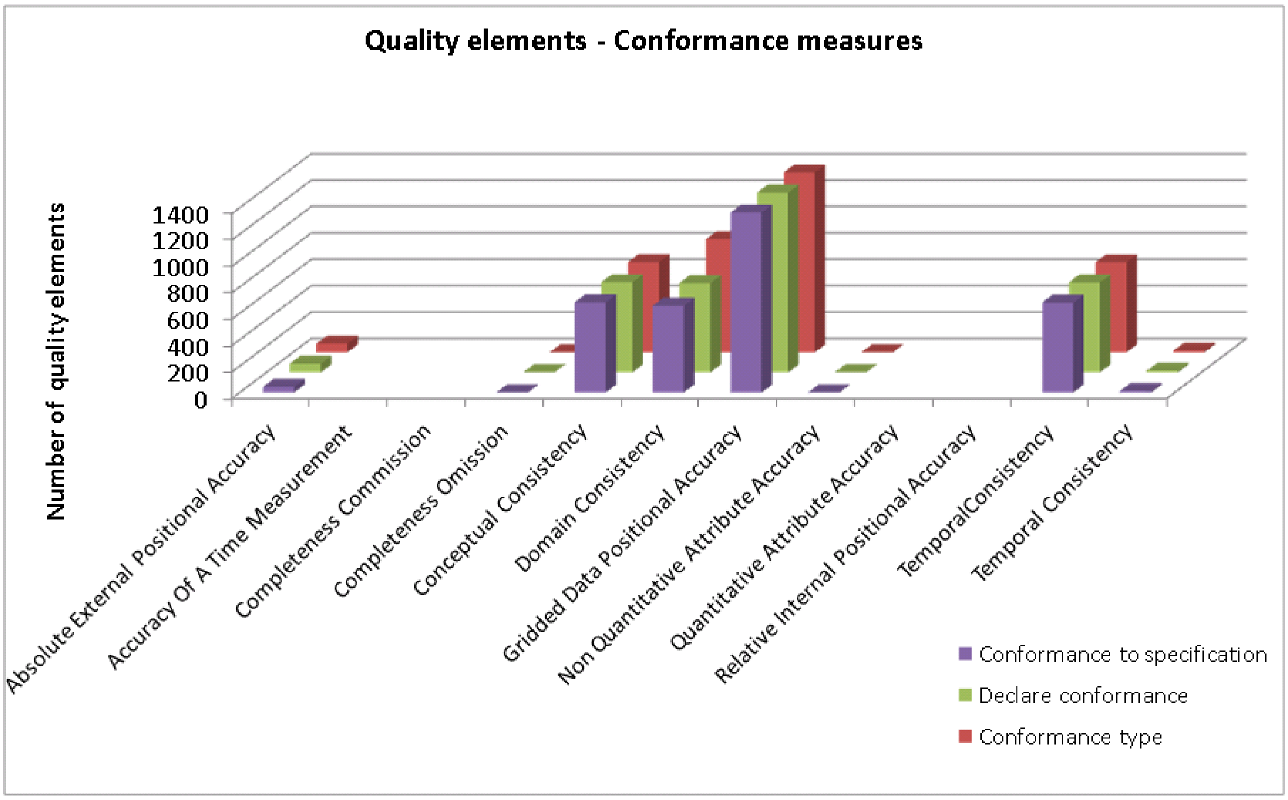


Figure 16-2. Graph containing detailed information about the qualitative indicators and subsequent information



In both figures the quality indicators classified in 3 levels of concretion are represented. The red bar represents all the indicators declaring the result type, either quantitative (Figure 16-1) or qualitative (Figure 16-2). As can be appreciated in the figures, it is the highest bar. The green bar represents the number of indicators containing the indicator type and also a unit (when quantitative) or instead an indication of whether the dataset passes the conformance or not (when qualitative). In all cases the results are lower than those obtained for the red bar. Finally, the blue bar represents the number of indicators containing the name of the measure (or else an explanation), a value and a unit (if quantitative) and a name or description of the measure, if it passes the conformance and the name of the specification (if qualitative). It can be concluded from these figures that the most complete information (for example - "root mean square error with 1 pixel of error") is obtained in very few cases.

4.4 Lineage

The ISO 19115:2003 (ISO-TC211 2003) specifications hierarchic model for Lineage is shown in **Figure 17**. As was explained, Lineage is descends from Scope. Lineage in turn embeds sources and process step. In addition, there is an alternative way to reach process step through Source. This means, process step can directly hang either from Lineage or from Lineage through Source in which case the sources implied in the corresponding process step should be specified. This last option would imply a complete record on provenance. This feature is also a highly demanded user requirement.

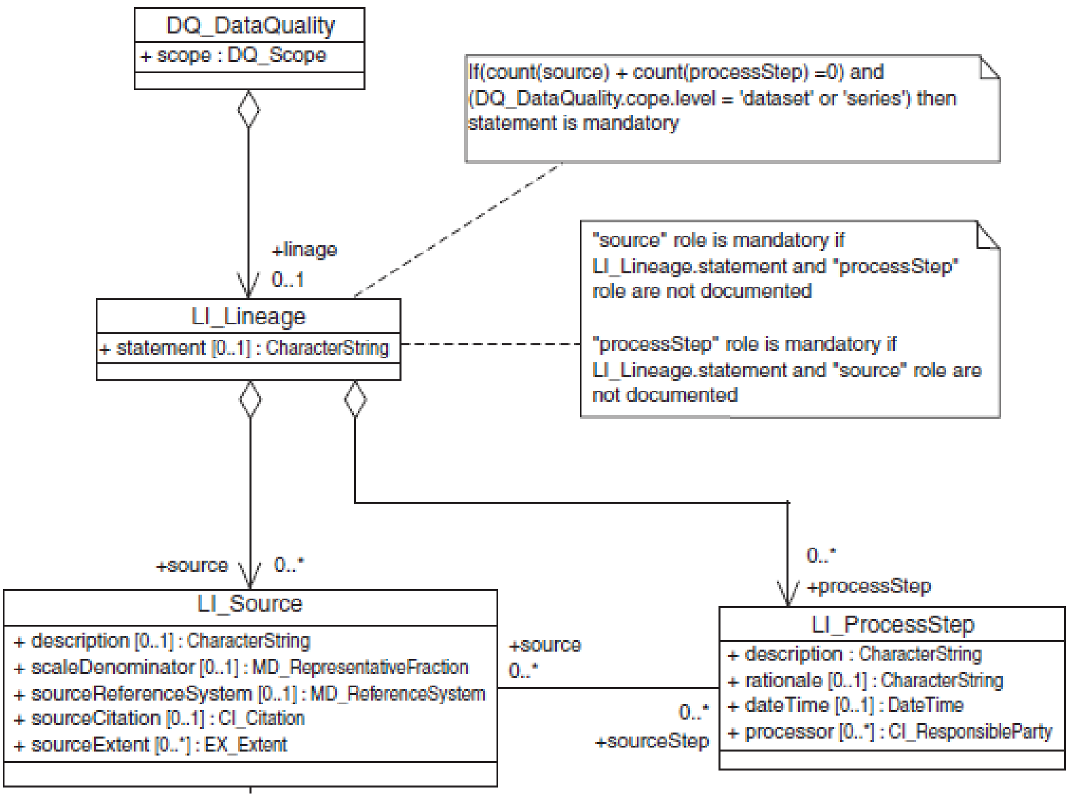


Figure 17. Lineage schema under ISO 19115:2003 (ISO-TC211 2003), including Source and ProcessStep.

Referring to Lineage, in the Clearinghouse there are 5851 (6.02%) metadata records containing a direct list of the data sources, 9261 (9.53%) metadata records containing a direct list of the processes, and 1226 (1.26%) metadata records that links *data sources* to each *process* (complete provenance).

Separately analysing sources and processes, GeoViQua has obtained the following results:

-5851 metadata records (6.02%) contain a direct list of the data sources, including 1798 (1.85%) with temporal extent

Table 6. Summary of source elements contained in the metadata records extracted from the Clearinghouse

Source Elements	**CITATION**	**6578**
	DESCRIPTION	**5777**
	TEMPORAL ELEMENT (EXTENT)	**3805**
	SCALE DENOMINATOR	**2070**
	VERTICAL ELEMENT (EXTENT)	**0**
	GEOGRAPHICAL ELEMENT (EXTEN	**0**
Metadata with source elements		**5851**
Of which	MD with Temporal element	1798

Note: 5851 is not the sum of citation + description + temporal element.The same metadata record might contain several source elements and thus the total number of metadata records with source elements is not the sum of the source elements (*i.e.,* source elements might br in the same MD document).

As for sources, this is the simplest way to mention that this resource was derived from previous (re)sources, at the same time giving credit to the registries (*i.e.,* provided attribution, and eventually some trust on them). Moreover, this distribution enables some form of descriptive quality report based on sources. In this case, if quality indicators are not provided for the dataset, the quality indicators from sources can still provide a clue of the quality.

9261 metadata records (9.53%) contain the direct list of the processes, containing either processes or processes linked to sources. Among these, 8035 (8.26%) strictly inform about MD_ProcessStep exclusively, without mentioning sources. There are 292 (0.30%) identified with date.

This provides information on the exact list of processes execution and the order of these processes. It must be stated however that following this structure, if quality indicators are not provided for the dataset, it is difficult to infer resource quality when only a process list is available.

Table 7. Process step elements contained in the metadata extracted from the Clearinghouse

Process Elements	**DESCRIPTION**	**12914**
	PROCES RESPOSIBLE	**1800**
	DATE AND TIME	**437**
	RATIONALE	**15**
Metadata with process elements		**9261**
Of which	MD with Date and time element	292

Note: The same metadata record might contain several process elements and thus the total number of metadata records with process elements is not the sum of the latter.

-Finally, considering the complete path from Lineage to Process_Step through Sources, although it is the most appropriate way to inform about processes and sources, in our opinion, 1226 metadata records (1.26%) have been identified in the Clearinghouse. Thus, these records imply a more complete provenance tracking, providing a list of processes execution and the order of these processes, how and when the data sources were used. In the end, with this information it is possible to infer which sources represent a higher influence over quality and the associated final result.

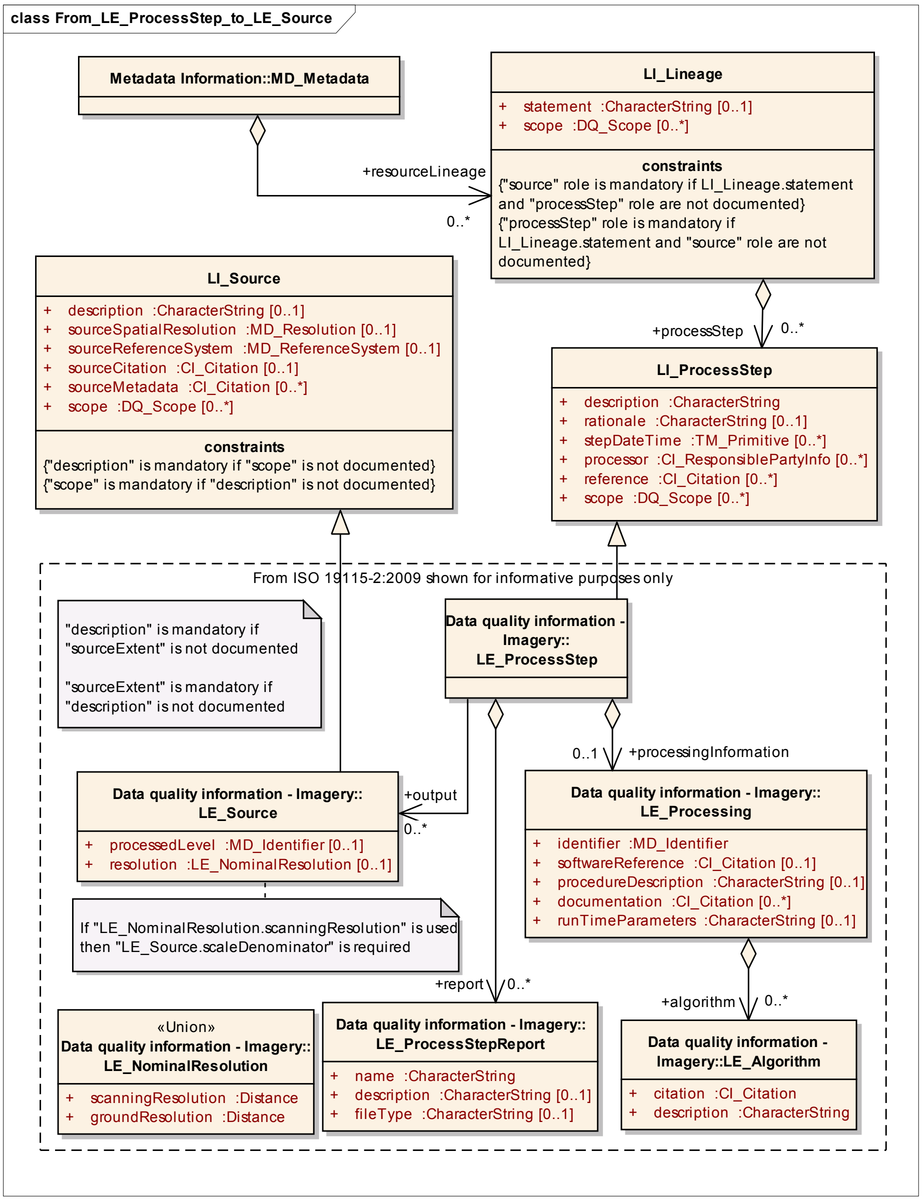


Figure 18. ISO 19115-2:2009, Lineage extension of ISO 19115:2003 (**ISO-TC211 2003**)

The main features in **Figure 18**, could be summarized as follows: LI_ProcessStep includes a LE_Processing that contains a runTimeParameters attribute allowing for the description of the exact list of parameters used in the execution. This path enables the recording of the exact information needed in the execution of the process and its reproduction. In addition, there is a citation of the used algorithm (LI_Algorithm). In this form, it is possible to exhaustively evaluate the quality of the resulting product, should the uncertainties that sources have in their metadata (sourceMetadata citation in LI_Source) be known.

Table 8. Summary of results of metadata following the complete provenance path including linking process steps and sources

Process Elements	CITATION	2094
	DESCRIPTION	0
	SCALE DENOMINATOR	0
+ Sources	SOURCE EXTENT	0
Metadata with process elements		1226

4.5 Usage

In addition, there is one small entry intended for enabling user feedback in the current ISO-19115:2003 (**ISO-TC211 2003**), which is MD_Usage. This class allows for a brief description of ways in which the resource is currently being used or has been used for by different users. This fact is sometimes known to data providers by means of established reporting channels enabled for users. **Figure 19** shows the standard specifications for MD_Usage.

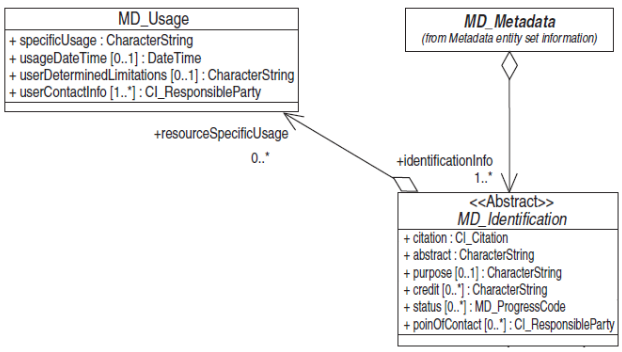


Figure 19. MD_Usage schema under ISO 19115:2003 (ISO-TC211 2003), showing the detail of the "zero to many" correspondence between MD_Identification (related to the dataset) and the MD_Usage (related to the use of the dataset).

Additonally, it can contain a brief description of procedures of using data, the list of resources, the date and time of the first use or range of uses of the resource, and also limitations of use, determined by the user for which the resource is not suitable. Moreover, the usage contains the identification with the person and organization using the resource.

Although in the analysis performed there have been 1133 records (1.17%) identified containing usage information, only the mandatory specificUsage and userContactInfo elements were described. Moreover, all records were provided by the same institution, which in the end produces a very peculiar scenario, likely not representative.

Nonetheless, due to the weight of user feedback within GeoViQua, especially after inputs received from WP2, it is interesting to make some notes on this section of the ISO standard. In the discussion section, the contrast between the existing and user-requested quality information will be emphasized, especially as regards the so-called non-quantitative quality components, and the items defined in the standards. MD_Usage contains SpecificUsage, which is a brief description of the resource and/or resource series usage, UsageDateTime determines date and time of the first use or range of uses of the resources and/or resource series, UserDeterminedLimitations refer to applications, determined by the user for which the resource and/or resource series is not suitable, and finally, UserContactInfo provides a place for the identification of and means of communicating with person(s) and organization(s) using the resource(s).

A very brief outline of the results explained in this section is shown in **Table 9**.

Table 9. Summary of the analysis of the Clearinghouse metadata quality data content

Total metadata records	97203
Metadata records with quality indicators	19107
Total quality indicators	52187
Total quality measurements	25944

Thus, the analysis of the Clearinghouse carried out proved that quality indicators are present in the GEOSS GCI, in diverse flavours. This is not a common occurrence when analysing spatial data infrastructures, and was not at all the case in the previous study performed for a Spanish regional SDI.

Lineage information is rich in many documents. In fact, some of the documents presented over 100 entries in either Source or ProcessStep.

Regarding quality measurements reported, quality coverage results (per pixel quality assessment) for quality measurements are scarce, though existent. Another inconvenient fact related to this issue is that the link to the coverage results is not directly accessible.

In addition, usage examples, although within a peculiar scenario, have been identified.

5. DISCUSSION

In the previous sections some issues have been already discussed. Here, we proceed to summarize some points and add further facets to the discussion.

There exists a set of standards and specifications for data quality in Earth observation, which includes ISO standards, W3C standard and some specifications like UncertML.

Some of the standards are intended for general use in multidisciplinary fields. W3C specifically deals with provenance, recognizing it as a relevant facet of data quality information. Provenance may provide trust cues for the data by recording information on the actors and processes used to derive the data, and could include an annotation that would serve as a standard for user feedback. The forthcoming ISO 19157 (ISO-TC211 under development) conveys a consistent and standard manner of reporting quality information including guidelines for evaluation procedures of quality information, replacing ISO 19113:2002 (**ISO-TC211 2002**) -Quality Principles-, ISO 19114:2005 (**ISO-TC211 2005**) -Quality Evaluation Procedures- and ISO 19138:2006 (**ISO-TC211 2006**) -Data Quality Measures-.

Also to be taken into account, we would like to recall the fact that geospatial standards may conform to formats other than XML. Even ISO 19115:2003 (**ISO-TC211 2003)** in fact is only translated to XML by the realisation under ISO 19139:2007 **(ISO-TC211 2007)**. In this sense, the Clearinghouse harvesting protocols incorporate stylesheets, which enable translation to XML from other formats. This harmonisation of metadata formats explains that the sample that has been subjected to the exploration is entirely extracted as XML metadata format.

Our study reveals that approximately 20% of the records contain some reference to quality indicators formalized following ISO 19115:2003 (**ISO-TC211 2033**) in XML form (*i.e.,* at least a DQ_Element is provided). This is evidence that many resources in the Clearinghouse are following the QA4EO principles and particularly, QA4EO-QAEO-GEN-DQK-002: the need to assign to all data / information products a Quality Indicator (QI). However, although this information is provided to the users, no effort has been made to make it more visible in the GEO Portal. In GeoViQua's view, QA4EO implies that the value of a quality indicator should be linked to the descriptions of procedures involved in the obtention of this value. Strictly speaking, the QA4EO provenance and lineage are related to the quality indicator itself and thus do not refer to the lineage of the dataset *sensu stricto* (*i.e.,* processes to derive the dataset from sources). Moreover, the ISO standard does not cover the provenance of a quality indicator.

The classification of the metadata records by scope detailed in section 4.2 explains the percentage of different quality "ítems" that are present in the Clearinghouse, and provides the argument to stick to datasets in the analysis.

The GEOSS clearinghouse metadata study done by GeoViQua reveals that some datasets accumulate more than one thousand processes and sources in their metadata records and that provenance information is provided in 12% of the resources. In general, provenance descriptions are mainly textual and consist of links to the local data that was used when sources are mentioned (*i.e.,* using local file system notations). This means that provenance as it is currently provided does not supply the link to distributed data resources in catalogues. The origin of this drawback is that provenance is generated automatically by the software in local computers accessing local files and not in a distributed environment. Once the metadata is made public, redirection of local files to public resources in catalogues (when available) was not taken care of.

Another relevant conclusion revealed in the analysis is that the more detailed the quality information (i*.e.,* lower levels in the hierarchic model of the standard), the fewer the number of metadata records with content. This is important because it might imply that either the data provider has not filled in the metadata records to the ultimate level of detail or the information has been implemented in formats that don't allow for a complete harvesting process. It could also imply that the value or unit are not available and this would mean that the information in the higher levels of the hierarchic model is useless (*e.g.,* having stated that there is an RMS for the dataset, with no value and no unit, is ultimately like not having the RMS at all). An option in this case would be contacting the data provider, but this is neither always feasible nor reasonable in massive search or massive analysis of quality, or even in dataset search, comparison and selection. In this sense, it is not possible to ascertain whether the obtained results for the red bars for instance (Figure 16) involve this background or not.

As the ISO19115:2003 (**ISO-TC211 2003**) "MD_Usage" class can only handle simple user comment, an appropriate user feedback model need to be further developed to support ranking / rating mechanism and blog-based discussions.

The lack of quality indicators thwarts the value of most datasets forcing users to make decisions on usage without well-grounded and rational justifications. Unjustified assumptions and problems with robust error propagation assessment are some general consequences of this habitual fact. By contrast, the presence of a quality indicator can inform of a global measure of quality without implying a detailed examination of a dataset. However, despite the standardisation bodies' work towards establishing core metadata elements and enforcing good metadata practices, dataset providers do not always follow standards, and may leave metadata records incomplete.

The surveys conducted in GeoViQua identified that geospatial data users are extremely interested in quality information records. At present, users find metadata records are typically incomplete with a lot of essential data omitted. Effective data quality evaluation cannot be performed under such circumstances. Though difficult to implement in the current standard metadata records, other desirable quality information are data providers', peer-reviews and users' comments on overall quality of the data product, explicitation of data errors, or potential usage, provenance and citation. Indeed, geospatial data users and experts stated that they rely heavily on peer feedback and recommendations when selecting a dataset, and that they contact their peers to obtain suggestions on what datasets are most suitable and are of good quality. Likewise, licensing information shouldn't be missing as it usually is. Experts stated that they typically require information about dataset providers, and in particular, valid contact details. The reputation of data providers was identified as a key factor in dataset selection.

On the whole, discovery, comparison, selection of datasets and feedback are more difficult with the current practices, particularly taking into account the huge increase of data products and locations.

In summary, from the data production at the beginning of the metadata contents flow, a proper fulfilment of the standards specifications translated into complete metadata recording is not a trivial matter. Actually, most of the assumptions that are made in order to enable massive analysis are based on an understanding that metadata will have been properly filled, which is not in the least certain. On the contrary, what has been identified for certain is a generalized practice of not completing the specifications especially in the quality classes, and the delivery of incorrect content; for instance not accomplishing with the specified allowance for the corresponding field (*e.g.*, introducing an "application" character string where the code list only admits "software", "service", "dataset"…).

It is highly desirable that GeoViQua should contribute somehow to the harmonisation of standards with its proposed Quality Model through participation in quality tutorials and guidelines for data providers.

In short, all these circumstances make a massive analysis easily unfeasible.

5.1 Corrective actions proposal

As was shown in **Figure 4**, with the inclusion of capacity catalogues that the EuroGEOSS Broker integrates, the Clearinghouse does not contain the complete set of metadata records available from the GCI. GeoViQua proposes to extend the evaluation of quality elements in metadata records to other capacity catalogues in order to find out how much quality information could be extracted and derive conclusions related to the harmonization of metadata records in the EuroGEOSS Broker without losing quality information in the flow. To date, it is hard to design an exhaustive methodology allowing for the extraction of all quality elements following the hierarchic structure of the quality model under the ISO specifications in an extra-large massive sample. Also, due to the architectural design of the evolving GEOSS GCI, it is even more difficult to track whether the quality information is lost in the data processing flow along the way between the GCI components to the Broker or from the Broker. Although these issues are being accounted for in the new architecture, at present it is difficult to foresee in advance how to coordinate efforts and define a consistent quality information analysis methodology valid for the next few years.

For instance, currently, Web-accessible folders are a simple way to provide access to a list of resources that can be directly downloaded using hyperlinks. This way, producers skip the need to set up and maintain a separate metadata catalogue and a data service. This feature, although useful from other points of views (*e.g.,* exchange of information, quality information flow, etc.), can hinder the metadata harvesting process by the GEOSS Clearinghouse or the EuroGEOSS Broker. Regarding formats, datasets in HDF, netCDF, TIFF (with additions), and JPEG2000 formats support embedded metadata. GeoViQua is considering the feasibility of creating tools that would enable an automatic extraction of quality parameters from these formats and generate independent metadata records, thus easing the harvesting processes. As anticipated in section 1.6, this activity would be within the scope of of section 2.3.2 “Data, Metadata and Products Harmonization” from the DHWG started in GEOSS AIP-3. In fact, the GeoViQua Quality Model could also be an important component in such a tool. In this sense, it must be remarked that the separation between data and metadata that has been adopted for practical reasons in metadata catalogues is an artificial one that needs to be overcome by using only one identifier for data and metadata and additionally a link between the two if required. GeoViQua is aware of the implications of this matter in the implementation of the proposed Quality Model.

There is a need to provide an encoding and recommendations on how to provide provenance information in such a way that GEOSS resources can point to other GEOSS resources. Even if a producer wants to fix provenance metadata to link to public resources in GEOSS, the main obstacle to achieving this purpose is the lack of a clear schema for a global and persistent identifier for each GEOSS resource. In addition, the Clearinghouse catalogue adds to the problem by providing new identifiers to pre-existing resources (*i.e.,* with previously assigned identifiers in the source catalogues). An approach to overcome this challenge can be found in the experimental Portal GEOSSBack (www.ogc.uab.es/geossback), where the GEOSS Clearinghouse identifier is used to link user comments to GEOSS resources. Nonetheless, this strategy is only valid in the existing GCI architecture for the Clearinghouse. Thus, in this sense, at present, the EuroGEOSS Broker is using its own solution for a coherent global identifier and GeoViQua is considering an extended proposal in this line for the AIP-5.

As described previously, citation information, user feedback, discovered issues, and other soft knowledge quality information are crucial. Thus, these statements should be treated with the same priority as producer data quality information. However, non-quantitative quality components are not explicitly supported in the most widely used standards in GEOSS (*i.e.,* ISO 19157 and ISO 19115 series). While ISO 19157 describes the quality elements and associated descriptors in detail, the handling of consumer data quality information is less discussed. In fact, under the ISO specifications, only the “Lineage” class is explicitly treated as ‘soft’ knowledge quality information, and falls directly under the “Metadata” class. This implies that quality scope, measure, evaluation, and result, related to this ‘soft’ knowledge quality information cannot be explicitly applied and consistently described.

In turn, in ISO 19115:2003 (**ISO-TC211 2003**) and ISO 19115-1, the “Citation” class is mainly used for identification of the dataset (*i.e.*, to appropriately reference a dataset). This is to say, it does not explicitly support the purpose of "how often and which published papers or researchers have used that dataset.“ Although “additionalDocumentation” of “Citation” type could be used in this sense, at present it seems that “additionalDocumentation” is just an attribute of an “Identification” class not compatible with the latter purpose. Moreover, currently, the “Citation” class does not include a Digital Object Identifier attribute for referencing modern digital scientific publications.

GeoViQua proposes to extend the quality information model based on the ISO standards by means of a quality component for data quality that would also allow characterization of ‘soft knowledge’ quality information (*e.g.*, including publication and citation, user feedback, lineage and discovered issues). A prototype system allowing users to provide their comments on datasets is the GEOSSBack Portal (http://www.ogc.uab.es/GEOSSBack). In turn, WP2 outcomes are leading to the development of a Feedback model that would be connected to the GeoViQua Quality Model.

In terms of the implementation of the proposed quality information model, GeoViQua will develop a schema by extending the existing schema provided by ISO 19139. To assist in encoding uncertainty without loss of valuable detail, GeoViQua recommends the use of the UncertML 2.0 dictionary and associated schema. UncertML is based on concretely-defined elements which represent a number of distributions, summary statistics, realisations and samples. Inheritance is required in order to enable storing metadata items and quality information at the optimum level in a hierarchic model. GeoViQua’s Quality Model would be totally applicable in this respect. Remarkably, no recommendation is stated on the current standards regarding inheritance rules.

As for distribution, the OGC WMS provides a standard protocol for requesting geo-referenced map images from one or more distributed geospatial databases. However, WMS specifications do not apply for the data quality aspect. In GeoViQua, WP4 is approaching this issue by proposing a quality-enabled WMS (*i.e.,* WMS-Q) and some guidelines in how to use the quality information model to encode the quality information in such Web services. Likewise, quality awareness for the Keyhole Markup Language (KML) would be an asset and could be considered in the scope of GeoViQua.

As has been stated, components in the GEOSS GCI architecture are evolving. The GEO Portal seems to be the one interface for the user from which resources are searched, discovered, intercompared, visualized and/or accessed. In this sense, it is to be noted that users' requirements will vary at different stages in the usage of the data. Users will generally be less interested in high-granularity accuracy information at the discovery stage than in the intercomparison or usage stages, but will still need to know of the existence and coverage of that detailed quality information on order to assess fitness-for-purpose. Therefore, all levels of quality information should be discoverable through the GEO Portal. Ultimately, considering the whole product lifecycle from producers to usage and user feedback, quantitative quality indicators appropriate for quantitative variables and also categorical variables complemented by the so-called soft quality knowledge (*e.g.,* peer-review, expert judgement, citation, provenance) must be taken into account. Thus, GeoViQua proposes the use of XSLT transformations to make quality elements (including quality indicators, lineage, and usage) more visible in ISO metadata records in the GEO Portal. Additionally, a metadata comparison tool could make it possible to compare quality indicators between two or more metadata records coming from different datasets.

Another identified issue in the analysis of the Clearinghouse metadata records is that despite the employment of quality indicators, the name and description of the corresponding measure used to derive the indicator are in fact rarely well described. Problems can occur due to the lack of semantic definitions of quality measures. For example, the value of a quality measure can be textual, categorical, or numeric. Both system and operators need to properly understand the meaning of some quality measures such as "uncertainty at 90% significance level". In order to address this need, GeoViQua is considering a Quality Measure Dictionary aimed at defining vocabularies for quality measures with associated semantic annotations that integrate UncertML concepts and vocabularies. The use of such a dictionary can effectively provide a knowledge base of quality measures hence hiding the scientific details to the application developers. The dictionary is expected to be used in quality information discovery and usage (*e.g.,* quality visualisation).

In the improvement of visualization and comparison of data products and services in the GEO Portal a full and correct submission of metadata is required from the producer. To help providers to assess the completeness of their metadata records in general and on quality aspects in particular, GeoViQua proposes the extension of the rubric system (http://www.ngdc.noaa.gov/metadata/published/NOAA/NESDIS/CLASS/iso). In this context, in the EuroGEOSS "Advancing the vision for GEOSS" symposium (Madrid, Spain, January 2012), the need of implementation of quality tutorials was highlighted. In fact, data providers are at present not certain that taking the time and effort of completing metadata records will benefit data distribution, discovery, access or user choice. In the symposium, GeoViQua and other working groups related to GEOSS agreed on the importance of defining clear paths in the links between users and providers, with a special emphasis on issues relating to the quality of quality indicators (*i.e.,* meta-quality, producers rewards in fulfilling metadata records, users accessibility, quality information flow, user feedback, extension, harmonization and generalization of standards).

Finally, GeoViQua is aimed at producing tools for GEOSS. Ongoing activities in WP4, WP5 and WP3 are leading to prototypes respectively for search, visualization, and quality elicitation components in coordination with the evolution of the new GCI architecture. A core component in the intended functionalities is the GeoViQua's Quality Model connected to a User Feedback model. However, quality elicitation and visualization tools will be useful only when components and in particular quality information is available, produced, validated and recorded following quality standards, in adequate formats that enable propagation, feedback and information flow within GEOSS architecture, from producers to users and *vice versa*. At present, the results of our study point out the lack of completeness as regards data quality in metadata records, at least in the sample extracted from the GCI Clearinghouse. Unfortunately, with our experience in the Earth Observation field, we believe this conclusion reflects general practice and that there is still a long road ahead in the improvement of quality information recording, producing and usage.

In this context, the Best Practices wiki within the GEOSS GCI components is a promising initiative. GeoViQua is planning to gather recommendations on, for example, the choice of resources identifiers which will be global, consistent and compatible with GEOSS needs.

5.2 Dissemination and ongoing activities

These conclusions have been recently presented in the 3rd PTB Meeting held in the University of Reading (Reading, United Kingdom) and as part of the dissemination activities in GeoViQua, two recent presentations were made in the EuroGEOSS symposia "Advancing the vision for GEOSS" conference in Madrid (25-27th January 2012):

-GEOSS Quality Framework and Quality Information Encoding (**Xiaoy Yang et al. 2011**)
-Analysis of the Quality Metadata in GEOSS Clearinghouse (**Sevillano et al. 2011**)

Additionally, GeoViQua is directly participating and/or in direct contact with staff engaged in other parallel activities within GEOSS related projects as was explained in section 1.6 (*e.g.,* UncertML, QA4EO, AIP-5, GEO Spain, OWS-9). Regarding AIP-5, from the producer's point of view, QA4EO principles can be summarized in two main objectives:
-provide quality indicators
-provide traceability of quality indicators

GeoViQua is considering the possibility of reusing a LI_Lineage element to support a full traceability and additionally the soft-knowlede extension of the GeoViQua Quality Model is needed in the implementation of references to existing publications and peer reviews of the data and a better usage description (MD_Usage). The GeoViQua Quality Model could therefore contribute both to QA4EO and AIP-5 actions.

On the consumer's side, complaints have been stated about the completeness of the descriptions of datasets that result in relying more on users' comments than in producer's metadata. Thus, the feedback metadata model has to be developed and in under consideration in the AIP-4 activities dealing with quality. The expectations are to include user comments, ratings, use case descriptions as well as consumer details (related to the common user identification system activity in GEOSS). The feedback model should be simple and agile to adapt to the specific requirements of the social networks systems for the GEO community in the near future.

GeoViQua presented the results of this study to the AIP and forwaded some of the corrective measurements presented above (section 5.1) (*e.g.,* XSLT transformations, metadata comparison tool, extension to the rubric system in the producer's assessment of completeness of metadata, metadata extraction tool for diversity of data formats, data and metadata identifiers). GeoViQua also proposes the extension of the study to other capacity catalogues in the GCI and surely conclusions will be applied to ongoing activities.

Finally, some issues of this deliverable related to the OWS-9 deal with Cross-Community Interoperability (CCI) and OWS innovations (*e.g.,* fit-for-purpose quality, communication of quality metadata at the product, layer and pixel level, trust mechanisms), in coordination with the associated provenance information for OGCServices in the CCI and Provenance thread of OWS-9. For instance, tasks in GeoViQua related to provenance encoding, WPS-Q allowing for a semi-automated client supported interactive data registration, semantic and intercomparison rules, queries and visualization could be disseminated and performed according to the OWS-9.

5.3 Next steps

GeoViQua is now entering an active stage in the development of quality-aware tools after analysis of the quality components in both products and services in GEOSS, existing metadata standards, implementation and GCI architecture, and last but not least, providers' and users' requirements.

Outcomes of our exploring analysis have been discussed in the previous sections of this document and further assessed in the corresponding deliverables. In this context, as stated in the ISO 19157, meta-quality is information about the quality of the quality information **(ISO-TC211**). Particular aspects require further investigation in the GeoViQua Quality model to support meta-quality, including:

- a mechanism to represent the lineage of quality information,
- a link to the reference data used in the quality evaluation, accompanying the quality metadata with producer caveats on applicability and reliability of the quality information,
- supporting user feedback on the data quality statements themselves in addition to user feedback on the data.

The ultimate target is to promote first the availability of quality information, propagation through the processes, and once this is achieved, elicit users' trust in quality information. In GeoViQua's view, meta-quality would be most usefully provided in a form that describes quality as exhaustively as possible, explaining how the quality indicators were generated, and thus enabling their potential reproduction.

As mentioned in the introduction of this deliverable, GeoViQua is adapting the methodology applied in this study in order to extend the analysis to some of the capacity catalogues currently linked to the GCI. WP3 is engaged in defining the approach to make feasible the extension of the present study to the new architecture of the GCI in GEOSS. Other federated previously-existing Earth Observation catalogues are in the process of being linked to the GCI, as demonstrated by the exponential increase in the number of resources currently linked to GEOSS since 2010.

The EUROGEOSS Broker (**Nativi et al. 2009**) is a key component in the new GCI infrastructure, constituting the service that transparently distributes any query to the GCI. GeoViQua is actively contributing to this Broker by incorporating quality requests over distributed data and the Broker is core to WP4 activities.

As detected in GeoViQua surveys, visualization of spatial data is a very important process for end users in the evaluation of data suitability for specific targets. The most commonly available simplistic visualisation tools often do not provide enough quality information for a user to make a decision. In addition, users need an adequate interface enabling the discovery of new datasets. The associated quality information is extremely important in this context. Indeed, quality is associated with the data product in itself but it also involves a usage component which is essential in determining the appropriateness of the intended use – in other words, the intercomparison of available datasets as regards a fitness-for-purpose search (**van der Wel 1994**).

Accordingly, GeoViQua tasks will develop adequate graphical representation of metadata quality parameters that in the end would assist the user in the exploring of data collection structure and patterns and, thus, enable appropriate data selection. Due to expanding capabilities and resources, geo-searches usually produce multiple results. In this frame, users may become lost in the process of objectively and subjectively determining which datasets are best suited for them in terms of quality. For instance, in WP4 some tasks are referred to search filtering based on conformance to certain quality standards, such as a GEO Label criterion compliance (also under study in GeoViQua), or quality indicators accessibility (*e.g.,* quality statistical charts, coverage results, quality representations through symbolization, user feedback and other techniques are to be explored in the project). The aim is that search results in GEOSS produce relevant quality information adapted to users' needs, so the latter can effectively rank search results by quality metrics.

Analogously, data visualization of quality information shall be integrated in the structure, either through existing standards and most probably through the extended version of standards that is being defined in the GeoViQua Quality Model. In the end, these tools would be accessible for users through the evolving version of the GEO Portal, and ideally through mass-market and free tools (*e.g.,* Google Earth–style map tools, 3D viewers and applications for mobile devices are under consideration).

6. CONCLUSIONS

This document describes the GEOSS GCI structure, focusing on the application of a methodology that enables the exploration of the quality information in GEOSS resources harvested by the Clearinghouse in October 2011. Due to the constant evolution of the GCI components and huge increase in geospatial resources, the analysis should be further extended. However, the amount of information and the state of the art in quality information availability and delivery make an exhaustive assessment hardly feasible. The methodology to be applied in the analysis of federated catalogues is thus to be revised.

As for results of the analysis performed for the Clearinghouse, it is demonstrated that the documentation of quality indicators and lineage is far from universal in the Earth observation data arena. Nonetheless, the up to date status revealed in the exploration of the GCI points out that the current documentation is a sufficient starting point for the development of tools for exposing and exploiting quality data already present in GEOSS catalogues.

More specifically, it can be remarked that a variety of quality indicators are used, providing both numerical and conformance results. However, the name and description of the corresponding quality indicator measure used to derive the indicator are rarely described. This fact highlights the need for a well described quality measurements thesaurus.

Finally the way usage information is provided clearly indicates that this mechanism is not the right solution for user feedback and in addition it is unknown by many users.

In GeoViQua, corrective approaches are focused in the definition of a Quality Model, based in the widely-used ISO standards and embedding appropriate extensions that account for the detected biases. These extensions include a formalization of the soft knowledge quality indicators and measurements demanded by users, the use of UncertML to complement ISO19138:2006 (**ISO-TC211 2006**) measures to encode statistical quantitative results, links to a Feedback Model, to implementation issues, to interoperability mechanisms and to services in the GCI architecture.

The EuroGEOSS Conference “Advancing the vision for GEOSS” showed that quality of geospatial products is a hot topic demanding immediate attention in GEOSS and related projects, including GeoViQua. In addition it has been recognized that data quality has many facets and the whole lifecycle has to be taken into account when addressing this issue.

In summary, in this deliverable, data quality concepts have been discussed, a methodology of massive dataset analysis was successfully applied in the exploration of quality information in the GEOSS GCI Clearinghouse, and the results have been analysed in the context of ongoing activities within GeoViQua. Existing standards and specifications on data quality, including practical implementation and users’ demands, are reviewed in this document and in other deliverables in GeoViQua. Our findings have led to the proposal of corrective measures, such as the development of an integrated Quality Model linked to a Feedback Model, and search, visualization and intercomparison tools for GEOSS. These components are definitely required provided the already large (and quick increase expectations) numbers of datasets in Earth observation. Actually, some of the conclusions as regards quality issues may as well have a wider impact in other e-science disciplines.

7. GLOSSARY

-AIP5 stands for the **5th phase of the GEOSS Architecture Implementation Pilot** (AIP-5). The leadership corresponds to the Open Geospatial Consortium, Inc. (OGC®).

AIP-5, as a continuation of AIP-4, will improve access to GEOSS datasets that support the "Critical Earth Observation Priorities" that have been identified by the GEO User Interface Committee. It will increase the use of these data by building on the accomplishments of prior AIP phases. AIP-4, concluded in December 2011, aimed to:

- Increase on-line access to "Critical Earth Observation Priorities Data Sources"
- Ensure datasets are discoverable through the GEOSS Common Infrastructure
- Demonstrate effectiveness of general and specialized software tools for using data.

GeoViQua is considering responding to the call for the new AIP-5 with a coordinated effort where our project could maybe contribute in a clearer way.

-Architecture and Data Committee (ADC). The Architecture and Data Committee supports GEO in all architecture and data management aspects of the design, coordination, and implementation of the Global Earth Observation System of Systems (GEOSS) for comprehensive, coordinated, and sustained Earth observations.

Objectives:

- Enable GEO, based upon user requirements and building on existing systems and initiatives, to define the components of GEOSS, and to converge or harmonize observation methods, and to promote the use of standards and references, intercalibration, and data assimilation.
- Enable GEO to define and update interoperability arrangements to which GEO Members and Participating Organizations agree to adhere, including technical specifications for collecting, processing, storing, and disseminating shared data, metadata and products.
- Enable GEO to facilitate data management, information management, and common services, and will help to promote data sharing principles in support of the GEO Plenary for the full and open sharing and exchange of data and information, recognizing relevant international instruments and national policies and legislation.

-Committee on Earth Observing Satellites (CEOS): Established in 1984, the Committee on Earth Observation Satellites (CEOS) coordinates civil space-borne observations of the Earth. Participating agencies strive to enhance international coordination and data exchange and to optimize societal benefit. Currently, 50 members and associate members made up of space agencies, national, and international organizations participate in CEOS planning and activities.

The purpose of the Web Services Interest Group is to serve as a forum for exchange of technical information and lessons-learned experience about Web services and other WWW / Internet related software technologies. In order to facilitate construction of interoperable service infrastructures, the Interest Group closely observes and interacts with OGC and ISO standardization activities and OGC testbeds (e.g., WCS, CS-W, and WMS).

The Web Services Interest Group is encouraging all members to participate in the CEOS WGISS Integrated Catalogue (CWIC) initiative. CWIC is attempting to utilize OGC and ISO standards to facilitate cross-agency, cross-information system data discovery and direct data access.

The CWIC pursues the following objectives:

-To establish CWIC as a Web service interface for EO data discovery and access
-The development of what would hopefully become a “standard” interface for inventory search, results and granule-level data access
-Eventually to evolve to replace CWIC with member agencies employing the “standard” interface directly to their inventory and data systems

WGISS is one of three subgroups supporting the Committee on Earth Observing Satellites (CEOS). CEOS addresses coordination of the satellite Earth Observation (EO) programs of the world's government agencies, along with agencies that receive and process data acquired remotely from space.
The Working Group on Information Systems and Services (WGISS) promotes collaboration in the development of systems and services that manage and supply EO data to users world-wide.

-Community of Practice (CoP). A Community of Practice (CoP) is a user-led community of stakeholders, from providers to the final beneficiaries of Earth observation data and information, with a common interest in specific aspects of societal benefits to be realized by GEOSS implementation.

-**CSR. Components and Services Registry**.

The GEOSS Components and Services Registry provides a formal listing and description of all the Earth observation systems, data sets, models and other services and tools that together constitute the Global Earth Observation System of Systems. These various components are being interlinked using standards and protocols that allow data and information from different sources to be integrated. The components and services listed on the Registry can be searched and explored by decision-makers, managers and other users of Earth observations via the GEO Portal.

The success of the GEOSS Components and Services Registry will depend on the commitment of GEO Members and Participating Organizations to input and update their registration details on a regular basis. Registration involves completing a standardized form on-line.

-Catalog Services for the Web (CSW):

The OGC Catalog Service defines common interfaces to discover, browse, and query metadata about data, services, and other potential resources.

Web Catalog Service includes several profiles including Catalog Service - Web.

Web Map Service (WMS)
Web Feature Service (WFS)
Web Coverage Service (WCS)

-Data quality: As it is stated in GeoViQua’s terminology section, ISO 9000 defines quality as the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs, degree to which a set of inherent characteristics fulfils requirements. Since in Earth observation each data item might be used for many purposes, some of which are not foreseeable in advance, and provided data quality itself has different facets, GeoViQua has analysed the standards and related works approaching data quality both from the producer data quality elements and consumer data quality elements, addressing and extending ISO19113 (**ISO-TC211 2002**), ISO 19115 (**ISO-TC211 2003**), ISO 19157 and contributions from UncertML (http://www.uncertml.org/) in a proposal of Quality Model under development in WP3 and WP6. Quality indicators would include completeness, positional accuracy, temporal accuracy, thematic accuracy, lineage, meta-quality and user feedback (*e.g.,* citation, peer-review, etc.). In this sense, GeoViQua project considers the whole lifecycle of quality information.

-EuroGEOSS Broker: To develop a pan-European information system, data and information available at the national levels should be interoperable since many of the issues are trans-border issues. Ultimately, the same needs are recognized at a global scale. EuroGEOSS is addressing interoperability through focus areas that originated as different specialties. For effective decision making, information from diverse scientific domains need to be readily available.

EuroGEOSS is addressing an approach using a Broker utilizing software approaches to facilitate interoperability while minimizing the impact on existing services. The steps for Broker development are:

- Extend the traditional SOA approach
- Enable multi-disciplinary interoperability
- Implement Broker, Mediation, and simple chaining services
- Make use of the GI-cat technology

An initial operating capability has been implemented. A wide range of services is supported by the Broker. These include discovery/catalog services, data download services. These are primarily based on OGC standard interfaces.

Enhancements to the initial operating capability are planned during the next year. These include: Ontology for semantic enabled discovery to support natural language use; Web 2.0 services and the incorporation of models for conversion of data to information for decision making.

Federated/capacity catalogues:

- **CWIC** (see CEOS above): The Web Services Interest Group is encouraging all members to participate in the CEOS WGISS Integrated Catalog (CWIC) initiative. CWIC is attempting to utilize OGC and ISO standards to facilitate cross-agency, cross-information system data discovery and direct data access.

- **Enviro GRIDS/WaterML**: The Group on Earth Observation Systems of Systems (GEOSS) is building a data-driven view of our planet that feeds into models and scenarios to explore our past, present and future. EnviroGRIDS aims at building the capacity of scientist to assemble such a system in the Black Sea Catchment, the capacity of decision-makers to use it, and the capacity of the general public to understand the important environmental, social and economic issues at stake. EnviroGRIDS will particularly target the needs of the Black Sea Commission (BSC) and the International Commission for the Protection of the Danube River (ICPDR) in order to help bridging the gap between science and policy.

http://envirogrids.net/index.php?option=com_content&view=article&id=1&Itemid=3
portal.opengeospatial.org/files/?artifact_id=21743

- **EuroGEOSS Broker**: the EuroGEOSS project adopted a Brokering Approach to implement multi-disciplinary interoperability and lower entry barriers for both users and data Providers. According to such an approach, users and data providers are not asked to implement any specific interoperability technology but to continue using their tools and publishing their resources according their standards as much as possible. In keeping with the Brokering Approach, EuroGEOSS has developed a Brokering framework (*i.e.,* a set of Brokers) to bind the heterogeneous resources published by the Data Providers, adapting them to the tools commonly utilized by the users. EuroGEOSS developed a Broker for each main interoperability functionality, namely: Discovery Broker, Access Broker, Semantic Broker. Actually, the EuroGEOSS Broker is a Brokering framework.

http://www.eurogeoss.eu/broker/Pages/AbouttheEuroGEOSSBroker.aspx

- **European Environmental Agency-Spatial Data Infrastructure (EEA-SDI):** Within the EEA strategy for 2004-2008, one of the priorities was to build an integrated spatial information system to support more effectively the development of EEA products and services. An operational spatial platform for geo-referenced data flows, modelling and analysis, will gradually develop towards a shared European integrated spatial information system in line with the INSPIRE (Infrastructure for Spatial Information in Europe) and GMES (Global Monitoring for Environment and Security) initiatives. Building the infrastructure is defined in the EEA management plan as a cross-cutting activity.

http://www.ekoi.lt/base_gis/docs/recom/EEA_GIS_Guide_200506.pdf

EEA Guide to Geographical Data and Maps (version 1.3)

- **Ground European Network for Earth Science Interoperations (GENESI)**: The GENESI-DR and GENESI-DEC projects are funded by EC FP7 programme intended to demonstrate possible new features. GENESI-DR and GENESI-DEC allow the real time discovery, access, and use (visualize, process on demand…) of space, in situ and other data types located in distributed and federated repositories/catalogues. GENESI respects the data providers' data policy: user authorisation is performed only once by data providers while the automatic user identification, for subsequent data access, is performed by GENESI services. The Earth Science Data Discovery and Access services are contributions provided by the GENESI-DR and GENESI-DEC infrastructure, in anticipation of similar capabilities for later integration in the GEO Common Infrastructure (GCI).

portal.opengeospatial.org/files/?artifact_id=38175

- **Global Biodiversity Information Facility (GBIF)**: The Global Biodiversity Information Facility was established by governments in 2001 to encourage free and open access to biodiversity data, via the Internet. Through a global network of countries and organizations, GBIF promotes and facilitates the mobilization, access, discovery and use of information about the occurrence of organisms over time and across the planet.

http://www.gbif.org/

- **Instituto Nacional de Pesquisas Espaciais (INPE)**: the Brazilian Space Agency fosters science and technology in earth and space context in order to offer products and regular services in benefit of the country. INPE's Data products and services have added to the community catalogues that should be accesible through the EuroGEOSS Broker.

http://www.inpe.br/ingles/institutional/about_inpe/mission.php

- **IODE Ocean Data Portal (ODP)**: is aimed at facilitating and promoting the exchange and dissemination marine data and services. This is achieved through a network of interconnected National Oceanographic Data Centres (NODCs) which can provide seamless access to collections and inventories of marine data from these data centres. The Ocean Data Portal provides the full range of processes including data discovery, access, and visualization. The ODP supports the data access requirements of all Intergovernmental Oceanic Commission IOC programmes areas.

http://www.ioc-cd.org/index.php?option=com_oe&task=viewDocumentRecord&docID=326

- **NASA Global Change Master Directory (GCMD):** The GCMD holds more than 25,000 Earth science data set and service descriptions, which cover subject areas within the Earth and environmental sciences. The project mission is to assist researchers, policy makers, and the public in the discovery of and access to data, related services, and ancillary information (which includes descriptions of instruments and platforms) relevant to global change and Earth science research. Within this mission, the directory also offers online authoring tools to providers of data and services, facilitating the capability to make their products available to the Earth science community. In addition, citation information to properly credit data set contributions is offered, along with direct links to data and services. As an integral part of the project, keyword vocabularies have been developed and are constantly being refined and expanded. These vocabularies are also used in other applications within the broader scientific community. Users may perform searches through the Directory's Website using controlled keywords, free-text searches, map/date searches or any combination of these. Users may also search or refine a search by data center, location, instrument, platform, project, or temporal/spatial resolution.

http://gcmd.nasa.gov/

- **NASA Goddard Space Flight Center (GSFC) DAAC:** The GSFC DAAC provides access to data products related to the following disciplines: upper atmosphere, atmospheric dynamics, global biosphere, geophysics. EOSDIS is an integral part of NASA's Earth Observing System (EOS). It is the robust distributed system that processes, archives, and manages Earth science satellite and field data, and distributes these data to a diverse global user community. Science data products (including over 700 from missions preceding EOS) are available from the DAACs, with each DAAC responsible for distributing data from specific disciplines. All data products are fully supported with documentation and technical user support.

http://daac.gsfc.nasa.gov/

- **NOAA Unified Access Framework Catalog (NOAA UAF):** NOAA has developed a dataset integration capability that leverages several standards: netCDF, a recently adopted OGC standard, which provides the abstract data model, software libraries and a persistent binary format the Climate and Forecast (CF) metadata conventions; the OPeNDAP protocol for Web transport of data subsets; THREDDS XML catalogs, providing a distributed topology connecting data suppliers; and an OGC compatibility layer that provides access to the grids through WMS and WCS. This resource contains ISO metadata that was harvested from the NOAA wide THREDDS catalog. Visit the NOAA Unified Access Framework Website for more information. http://geo-ide.noaa.gov/

http://ferret.pmel.noaa.gov/uafwaf/

- **One Geology**: OneGeology is an international initiative of the geological surveys of the world. This ground-breaking project was launched in 2007 and contributed to the 'International Year of Planet Earth', becoming one of their flagship projects. Thanks to the enthusiasm and support of participating nations, the initiative has progressed rapidly towards its target - creating dynamic geological map data of the world, available to everyone via the Web. We invite you to explore the Website and view the maps in the OneGeology Portal.

http://www.onegeology.org/

- **Sea Data Net**: SeaDataNet is a standardized system for managing the large and diverse data sets collected by the oceanographic fleets and the automatic observation systems. The SeaDataNet infrastructure network and enhance the currently existing infrastructures, which are the national oceanographic data centres of 35 countries, active in data collection. The networking of these professional data centres, in a unique virtual data management system provide integrated data sets of standardized quality on-line. As a research infrastructure, SeaDataNet contributes to build research excellence in Europe. SeaDataNet infrastructure was implemented during the first phase of SeaDataNet - SeaDataNet project, grant agreement 026212 was an Integrated Infrastructure Initiative of the EU Sixth Framework. The second phase of SeaDataNet (SeaDataNet 2 project, grant agreement 283607, started on October 1st, 2011 for a duration of 4 years) aims to upgrade the present SeaDataNet infrastructure into an operationally robust and state-of-the-art Pan-European infrastructure for providing up-to-date and high quality access to ocean and marine metadata, data and data products by:–setting, adopting and promoting common data management standards–realizing technical and semantic interoperability with other relevant data management systems and initiatives on behalf of science, environmental management, policy making, and economy.

http://www.seadatanet.org/

- **South African Earth Observation Strategy (SAEOS)**: The strategy captures South Africa’s response to the Global Earth Observation Systems of Systems (GEOSS). SAEOS coordinates the collection, assimilation and dissemination of earth observation data establishing an overarching national framework for Earth observation in South Africa.

http://www.esastap.org.za/esastap/pdfs/space05_2011_pres01.pdf

-Federated Earth Observation (FedEO): The Federated Earth Observation Missions (FedEO) Pilot is a collaborative, open standards development, supporting collective requirements of governments with the knowledge and capabilities of academic, commercial, and non-commercial solutions providers. This pilot is sponsored by ESA, the European Space Agency.
The Federated Earth Observation Missions (FedEO) Pilot will provide a broad international venue for operational prototyping and demonstration of Earth Observation (EO) requirements. The FedEO Pilot will

apply and refine OGC specifications relevant to EO, test and validate OGC specifications in a business context, and provide feedback regarding their ability to improve access to and application of earth observation data and services. Among other aspects, the FedEO Pilot will build on the experiences made with workflow description and execution in interoperable environments as performed as a part of the last major Testbed OGC Web Service Interoperability Program phase 4 (OWS-4 Workflows).

-**GCI. Global Common Infrastructure** in GEOSS (Global Earth Observation System of Systems). The GCI consists of Web-based Portals, a clearinghouse for searching data, information and services, and registries containing information about GEOSS components, standards, and best practices.
The goals of GCI in 2011 seek a facilitated access to datasets and other resources supporting the Earth Observation Priorities identified in UIC report, enable discovery or and facilitate access to data-core contributions.

To achieve convincing and demonstrable benefits at the GEO-VIII Plenary, a short term action is underway (ADC, GCI providers and others) to collaborate bi-laterally with a few data providers to identify datasets supporting the Critical Earth Observation Priorities.
To mutually adapt interfaces to enable effective data discovery and access (‘fewer clicks to the data’), better usability/user experience.

-**Ground European Network for Earth Science Interoperations (GENESI)**: The GENESI-DR and GENESI-DEC projects are funded by EC FP7 programme intended to demonstrate possible new features. GENESI-DR and GENESI-DEC allow the real time discovery, access, and use (visualize, process on demand…) of space, in situ and other data types located in distributed and federated repositories/catalogues. GENESI respects the data providers’ data policy: user authorisation is performed only once by data providers while the automatic user identification, for subsequent data access, is performed by GENESI services. The Earth Science Data Discovery and Access services are contributions provided by the GENESI-DR and GENESI-DEC infrastructure, in anticipation of similar capabilities for later integration in the GEO Common Infrastructure (GCI).

-**GEO. Group on Earth Observations.**

GEO is an intergovernmental organization working to improve the availability, access, and use of Earth observations to benefit society. It is composed of a voluntary partnership of 148 governments and international organizations, launched in response to calls for action by the 2002 World Summit on Sustainable Development and by the G8 (Group of Eight) leading industrialized countries. GEO is coordinating efforts to build a Global Earth Observation System of Systems (GEOSS).
http://earthobservations.org/about_geo.shtml.

-**GEO Label**

The Science and Technology Committee committed itself to developing the concept of a voluntary GEO Label. This was intended to:
-encourage scientists, researchers, and others to contribute their data and systems to GEOSS by offering an accepted voluntary label that provides recognition that their contribution is valued by the GEO community.
-differentiate components, data and products delivered through GEOSS and provide a “trusted brand” to GEOSS users; member governments may base their decisions on data/products of such contributions.
-highlight the importance of GEOSS to those previously unaware they were reliant on this initiative for their data or product.
The active task states that such a label should assist the user to assess the scientific relevance, quality, acceptance and societal needs of the components. These parameters clearly contain a mix of objective and subjective assessments.

GeoViQua follows the evolution of this task and actually participates with some people involved in the GEO Label group, linked to other institutions and projects connected to this activity (*e.g.,* EGIDA project). Specifically, members from ASTON, CNR, CREAF and UAB, led by ESA, constitute the GEO Label teamwork in GeoViQua. As regards this deliverable, some contributions from GeoViQua will come from WP7 (*e.g.,* task 7.4 might include testing of GEO Label features, gathering user requirements, design of surveys, etc.).

-**GEO Portal.** The GEO Portal is a Website that provides convenient access to the full range of GEOSS data and information.

Operated by the European Space Agency (ESA) and the Food and Agriculture Organization (FAO) of the United Nations (UN), it provides a Web-based interface for searching and accessing data, information, imagery, services and applications. It connects users to a variety of databases, services and Portals that provide reliable, up-to-date, integrated and user-friendly information – vital for the work of decision-makers, managers and other users of Earth observations.

The content available via the Portal continues to expand at a rapid rate and promises to reach a critical mass in the near future.

-**GEOSS. Global Earth Observation System of Systems.**
GEOSS builds on national, regional, and international observation systems to provide coordinated Earth observations from thousands of ground, in situ, airborne, and space-based instruments. GEO is focused on enhancing the development and use of Earth observations in nine SBAs: Agriculture, Biodiversity, Climate, Disasters, Ecosystems, Energy, Health, Water, and Weather.

-**GeoViQua**. GeoViQua stands for a FP7 project related to Quality-aware Visualisation for the Global Earth Observation System of Systems (GEOSS). The focus is on adding quality specifications to spatial data in order to improve reliability in scientific studies and policy decision-making. To achieve our targets, our team is collecting pilot cases spread over the whole Earth Observation chain.

Work packages in GeoViQua:

-WP1: Project management
-WP2. Requirements for data quality visualisation
-WP3: Data quality elicitation mechanisms
-WP4: Enhanced geo-search components
-WP5: Quality-aware visualisation components
-WP6: Delivery of solutions to end users
-WP7: Pilot case studies
-WP8: Dissemination and Capacity building

The **Description of Work (DoW)** is the reference document which contains a complete account of GeoViQua project and related work packages.

-Infrastructure for Spatial Information in Europe (INSPIRE):

The INSPIRE directive came into force on 15 May 2007 and will be implemented in various stages, with full implementation required by 2019.

The INSPIRE directive aims to create a European Union (EU) spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe.

A European Spatial Data Infrastructure will assist in policy-making across boundaries. Therefore the spatial information considered under the directive is extensive and includes a great variety of topical and technical themes.

INSPIRE is based on a number of common principles:
- Data should be collected only once and kept where it can be maintained most effectively.
- It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
- It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.
- Geographic information needed for good governance at all levels should be readily and transparently available.

Easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used.

-**Interoperability**: Following **O'Brien and Marakas (2005)** interoperability is defined as being able to accomplish end-user applications using different types of computer systems, operating systems, and application software, interconnected by different types of local and wide area networks. It includes syntatic, semantic, format, software components, and open standards, as relevant issues in GeoViQua and extendedly in GEOSS.

-Lineage: Tracking data sources and processing done to them is becoming increasing important as scientists seek to define trends and unexpected changes in the environment. Keeping track of data transformations and processing, generally termed *lineage*, is an important role of high-quality metadata. The ISO metadata standard provides a simple lineage model based on *sources* which are either used or produced in a series of *process steps*. This model can be helpful in many cases despite its simplicity. Sources and process steps are linked together to describe the lineage of a resource.

FGDC and ISO source descriptions have several important differences.

-The FGDC source model is based on citations to scientific papers. These citations do not include contact information (commonly an email address or a URL) which limits their usefulness in many cases.
-The scope of the references to the citations is limited to the metadata lineage section. This makes it difficult to unambiguously identify sources across multiple records.
-The FGDC standard allows the description of the temporal extent of a source, but not the spatial extent.
All of these limitations are overcome in the ISO lineage model.

-**Metadata.** This term refers to data complementary information, providing details about data features. In the case of geographic information data (*e.g.,* remote sensing scenes, categorical maps resulting from the application of digital classification techniques to remote sensing scenes, etc.), metadata would cover for instance identification aspects, discovery date, accuracy, distribution conditions, horizontal reference system, etc. There are several metadata standards defining metadata contents (*e.g.,* ISO19115:2003).

-OGC. Open Geospatial Consortium.
The Open Geospatial Consortium, Inc. (OGC®) funded in 1994 is an international consortium of more than 420 companies, government agencies, research organizations, and universities participating in a consensus process to develop publicly available geospatial standards, GIS data processing, or data sharing. OGC Standards empower technology developers to make geospatial information and services accessible and useful with any application that needs to be geospatially enabled. Visit the OGC Website at http://www.opengeospatial.org. Among the most well-known standards, the Web Map Server (WMS) that enable remote sensing images and maps visualisation, the Web Coverage Service (WCS) that is providing downloadable images servers, or the Web Feature Service Interface Standard (WFS) that provides and interface allowing requests for geographical features across the Web using platform-independent calls, can be cited. Some of the OGC standards (*e.g.,* WMS) have been accepted as ISO standards.

-Partners

GeoviQua is supported by the participation of 10 partners, including several universities, research centres, companies and the European Space Agency, from different European countries (France, Germany, Italy, Netherlands, Spain, and United Kingdom).

- 52° North GmbH (Germany): GeoViQua acronym: 52N

52° North GmbH (52N) is an SME that forms the back office and service centre of an international research and development network aiming at fostering the process of innovation in the field of geoinformatics. This network is driven by leading research organizations and individuals of the international GIS community. Partners participate in the development of innovative technologies and their transformation into practical solutions. Principal participants and shareholders of the 52° North GmbH are: con terra - Gesellschaft für Angewandte Informationstechnologie mbH, Münster (Germany), International Institute for Geo-Information Science and Earth Observation (ITC), Enschede (The Netherlands), Institute for Geoinformatics in University of Muenster (Germany), and Environmental Systems Research Institute, ESRI Inc., Redlands (California, USA).
The participants have a long and outstanding record in the domain of geoinformatics, spatio-temporal modelling, spatial data infrastructures, software architectures, and standardisation processes. 52° North and its participants are actively involved and leading the OGC standardisation process for Web based geoprocessing (Chair of the OGC WPS 2.0 Standards Working Group), sensor Web enablement (Chair of the SOS 2.0 OGC Standards Working Group) and digital rights management (Chair of the OGC GeoRM 1.0 Standards Working Group) as well as in INSPIRE drafting teams.

- Aston University (UK): GeoViQua acronym: AST

Aston University is a technology and application focussed University with areas of international excellence. Within Aston University, the School of Engineering and Applied Sciences (SEAS) has roughly 80 members of academic staff working in areas including Computer Science, Sustainability and Mathematics. Within SEAS, the world-class Neural Computing Research Group (NCRG) has for many years been at the forefront of developments in machine learning. The developing Knowledge Engineering Group (KEG) works within the areas of knowledge representation and management, database integration and interoperability, and software engineering. The Sustainable Environments Research Group (SERG) focuses on the application of GIS and modelling methods to address broad sustainability issues including conservation planning, renewable energy and disease monitoring.

- Commissariat a l'Energie Atomique (CEA). LSCE -Laboratoire des Sciences du Climat et de l'Environnement (France). GeoViQua acronym : CEA

The Institute LSCE (Laboratoire des Sciences du Climat et de l'Environnement) is a joint research unit of the Centre National de la Recherche Scientifique (CNRS-INSU), the Commissariat à l'Energie Atomique (CEA), (two major funding agencies in France) and the University of Versailles-Saint Quentin. For this project, CNRS will not participate and CEA alone will carry out and be responsible for the work on this project through LSCE personal. The LSCE is part of the Institut Pierre Simon Laplace (IPSL) in Paris and covers numerous fields of research related to climate and environment. The experience and qualification of LSCE researchers for the Global Carbon Project comprise a unique expertise in atmospheric composition monitoring, development of process-based models over land and ocean, and inversion methods to quantify sources and sinks of greenhouse gases using atmospheric, space-borne, and in-situ observations. The LSCE researchers also played a leading role in the development of Earth System modelling in France. The organizational experience of the LSCE in the European research area is reflected through its active participation in several EU funded projects, GEMS-IP, CARBOEUROPE-IP, QUANTIFY, CARBOOCEAN, CARBOAFRICA and NITROEUROPE-IP. The LSCE coordinates GEOMON-IP dealing with groundbased atmospheric observation complementary to satellite, and IMECC and ICOS dealing with infrastructure carbon cycle measurements. The LSCE is strongly involved in the GMES core services for land, atmosphere and ocean.

- Consiglio Nazionale delle Ricerche (CNR). IMAA - Istituto di Metodologie per l'Analisi Ambientale (Italy). GeoViQua acronym: CNR.

The CNR-IMAA (Institute of Methodology for Environmental Analysis) of the National Research Council of Italy (CNR) is a research Institute, whose research concerns the study of the atmosphere and of the Earth's surface, and of human impacts on these, using remote sensing and environmental and geophysical monitoring. It is a centre of competence of the Italian Civil Protection Department of the Prime Minister's cabinet. The Earth and Space Science Informatics Laboratory (ESSI-Lab) of CNR-IMAA was established to facilitate the effective and seamless provision of Earth and Space resources to Information Society applications. ESSI-Lab research focuses on the application of information and communications technologies to manage, share and harmonize Earth and Space Science data, information, services and knowledge in the framework of Geospatial Information technologies and infrastructures. A major research activity concerns crosswalks between the Earth and Space Science and Geospatial Information communities, investigating interoperability solutions. CNR-IMAA is a member of the OGC (Open Geospatial Consortium).
CNR-IMAA has specific knowledge and experience on international framework and standard related to Earth Science Systems resources access and interoperability.

- CREAF: Centre for Ecological Research and Forestry Applications (Spain). GeoViQua acronym: CREAF

The Centre for Ecological Research and Forestry Applications (CREAF) is a public research institution that was created in 1987 by the Generalitat (Autonomous Government) of Catalonia, the Autonomous University of Barcelona (UAB) the Institute of Catalan Studies (IEC) and the University of Barcelona (UB), to promote basic and applied research in terrestrial ecology. Some of these contributions are the publication of numerous scientific papers in international academic journals and the development of numerous scientific methodologies and technological tools such as the GIS&RS software MiraMon.

- ESA: European Space Agency (France): GeoViQua acronym: ESA

The European Space Agency (ESA) is Europe's gateway to space. Its mission is to shape the development of Europe's space capability and ensure that investment in space continues to deliver benefits to the citizens of Europe and the world. ESA's purpose shall be to provide for, and to promote, for exclusively peaceful purposes, cooperation among European States in space research and technology and their space applications, with a view to their being used for scientific purposes and for operational space applications systems.
ESRIN, known as the ESA Centre for Earth Observation, is one of the five ESA specialised centres situated in Europe. Earth Observation data has grown in importance as more and more international and national agencies recognise the many uses to which it can be put. Satellites for Earth observation keep a constant watch over the Earth and the data they provide help to safeguard the planet in which we live. Since 2004, ESRIN has been the headquarters for ESA's Earth Observation activities.

- Fraunhofer Institut Graphische Datenverarbeitung -IGD- (Germany): GeoViQua acronym: FRAUN

The Fraunhofer Gesellschaft (FhG) is the leading organization of applied research in Germany, undertaking contract research on behalf of industry, the service sector and the government. At present, the organization maintains 80 research establishments at 40 locations throughout Germany with around 13,000 employees. The Fraunhofer Gesellschaft will contribute to the GIGAS Project through Fraunhofer Institute for Computer Graphics Research (IGD). FhG focuses on the development of product prototypes (hard- and software) and the realisation of concepts, models, and solutions for computer graphics and its adoption to specific application requirements. The work is rounded off by object-oriented basic research projects and the realization of single devices and computer graphics systems with pilot character (*e.g.,* virtual and augmented reality, mobile computing, and security technologies). The tasks of the Graphic Information Systems (GIS) department at FhG are research, development and services in the field of Graphic Information Systems, especially on 3D GIS, location based services and spatial data infrastructure technologies. The activities cover research as well as end user and producer-relevant services.

- S[&]T Corporation (The Netherlands): GeoViQua acronym: S&T

S&T Corporation (S&T stands for Science & Technology) is an SME conducting high-tech projects and consultancy related to technical software engineering and analysis. S&T started in 2000 and has grown to a diverse group of 30 scientists and engineers. Typical customers and participants are the European Space Agency (ESA), TNO, EADS, ASTRIUM, VEGA, Siemens, SNECMA, Logica and others. S&T has a broad experience in building data quality monitoring, control and visualisation systems for various EO-missions, such as ENVISAT, GOCE, and Radar Altimetry. Typically these systems involve extraction of data from both instrument raw data and processed product flows, processing the data into quality data, conditional handling of EO data based on data quality analysis results, and visualisation and reporting data quality to users. The systems also allow quality engineers to analyse these data interactively in order to perform in-depth investigation.
In addition to its quality monitoring software development activities, S&T is currently involved in various standardisation projects for data quality, including investigating application of QA4EO to its software. S&T aims will contribute the lessons learnt from the implementation of industrial data quality monitoring systems and from its current quality information standardisation activities to the GeoViQua project.

- UAB: Universitat Autònoma de Barcelona (Spain): GeoViQua acronym: UAB

The Universitat Autònoma of Barcelona (UAB) was founded in 1968 and is 2nd in the national ranking of universities (2004-2005) and recently has been recognized as a campus of international excellence by the Ministries for Education (MEC) and for Science and Innovation (MICINN) of the Spanish Government. The research group in the Department of Geography aims to propose new algorithms, methodologies and tools for GIS, Remote Sensing, Cartography and land dynamics. UAB has broad experience in image processing of remote sensors, having worked with images of low, medium and high spatial resolution from both satellite and airborne sensors. The members of the research group have published about 150 scientific papers and obtained a "Research, innovation and development" award from the Catalan Government.

- University of Reading (UK): GeoViQua acronym: UREAD

The University of Reading is ranked as one of the UK's 10 most research intensive universities and one of the top 200 universities in the world, with a world-class reputation for teaching, research and enterprise. The University's research in environmental science is rated among the very highest in the UK, and in 2006 the University received the Queen's Anniversary Prize for Higher Education for exceptional contributions to Meteorology. The University enjoys very strong scientific and technical links with the UK Met Office and the European Centre for Medium-Range Weather Forecasting. The UK National Centre for Earth Observation (NCEO) is coordinated from the University of Reading.
The Reading e-Science Centre, formed in 2003, is an interdisciplinary group of scientists and technologists at the University that aims to apply modern computing techniques to environmental science. The ReSC is known internationally for its work in applying OGC-based techniques to problems of environmental data sharing and visualisation. Its open-source ncWMS software (an implementation of the Web Map Service protocol for multidimensional gridded data) is used by research groups, government agencies and private industry around the world.

-**Provenance**: Provenance is the capacity to annotate the sources and processes that were used in the creation of a resource (*e.g.,* datasets and operations involved in the creation of a particular dataset). It provides traceability and trust in the resource by knowing its origins and giving credit to the processors of the resource. If it is complete and has access to the same sources and processes, it potentially allows recreating the resource. It is considered a component of the quality description of the dataset.

-Quality Assurance Framework for Earth Observation (QA4EO): QA4EO has been endorsed by CEOS as a contribution to facilitate the GEO vision for GEOSS. This can only be achieved through the synergistic use of data derived from a variety of sources (satellite, airborne and in situ) and the coordination of the resources and efforts of the GEO members.

QA4EO was established to facilitate interoperability of GEO systems. QA4EO is based on the adoption of guiding principles, which are implemented through a set of key operational guidelines derived from best practices, for implementation by the GEO community. Although these guidelines were originally developed to meet the needs of the space community, they have been written with the aid of national metrology institutes of the UK and the USA and, where appropriate, are based on best practices of the wider non-EO community. They should therefore be readily adoptable by all GEO communities as a top-level framework that can subsequently be translated and implemented to serve each specialist need.

-Quality. The definition contained in the ISO multilingual glossary states quality is the degree to which a set of inherent characteristics fulfils requirements (ISO 9000:2005).
It is recommended that quality conformance levels based on user requirements refer to the data quality elements defined in this standard.

Quality parameters (in bold) and associated quality indicators (dashes)

Completeness
Completeness is defined as the presence and absence of features, their attributes and relationships. It shall be described by applicable data quality elements from the following list:

- commission: excess data present in a dataset;
- omission: data absent from a dataset.

Logical consistency
Logical consistency is defined as the degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical). If these logical rules are documented elsewhere (for example in a product specification) then the source should be referenced (for example in the data quality evaluation). It shall be described by applicable data quality elements from the following list:

- conceptual consistency: adherence to rules of the conceptual schema.
- domain consistency: adherence of values to the value domains.
- format consistency: degree to which data is stored in accordance with the physical structure of the dataset.
- topological consistency: correctness of the explicitly encoded topological characteristics of a dataset.

Spatial accuracy
Spatial accuracy is defined as the accuracy of the position of features in relation to Earth. It shall be described by applicable data quality elements from the following list:

- absolute accuracy: closeness of reported coordinate values to values accepted as or being true.
- relative accuracy: closeness of the relative, spatial positions of features in a dataset to their respective relative, spatial positions accepted as or being true.
- gridded data position accuracy: closeness of gridded data spatial position values to values accepted

as or being true.

Thematic accuracy
Thematic accuracy is defined as the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships. It shall be described by applicable data quality elements from the following list:

- classification correctness: comparison of the classes assigned to features or their attributes to a universe of discourse (*e.g.,* ground truth or reference dataset).
- non-quantitative attribute correctness: measure of if a non-quantitative attribute is correct or wrong.
- quantitative attribute accuracy: closeness of the value of a quantitative attribute to a value accepted as or known to be true.

Temporal quality
Temporal quality is defined as the quality of the temporal attributes and temporal relationships of features. It shall be described by applicable data quality elements from the following list:

- accuracy of a time measurement: closeness of reported time measurements to values accepted as or known to be true.
  NOTE: Time measurement may be either a defined point in time or a period.
- temporal consistency: correctness of the order of events.
- temporal validity: validity of data with respect to the format and calendar specified for the dataset.
  NOTE March 33 is an example of invalid data where the specified format is ISO 8601 compliant.

Usability
Usability is the degree of adherence to a specific set of data quality requirements.
Usability shall be used to describe specific quality information about a dataset's adherence to a particular application or requirements.
If the other data quality elements listed in this International Standard do not sufficiently address a component of quality, Usability shall be used.
NOTE For example, with this element, a data producer can show for one dataset, with quantitative elements, how it fits different identified usages. This element may be used to declare the conformance of the dataset at a particular specification.

-Quality measures: A data quality measure is a test applied to evaluate data quality elements. Quality measures in GeoViQua are based on ISO19138 (ISO-TC211 2006) standards.

-**Societal Benefit Area (SBA).**

The Societal Benefit Areas (SBAs) are nine environmental fields of interest around which the Global Earth Observation System of Systems (GEOSS) project is exerting its efforts. These include: Agriculture, Biodiversity, Climate, Disasters, Ecosystems, Energy, Health, Water, and Weather around which a preliminary hierarchical vocabulary has been created. Each area has some related fields (fifty-eight in total) of interest addressed as subcategories that specialise the goals of the entire project, the focus of the subcategories being mainly the relation between environmental issues and human activity and health.
One of the aims of GEOSS is to implement a proper system of earth monitoring and to render information deriving from this process available to a global range of users.

-Spatial Data Infrastructure (SDI): a spatial data infrastructure is a data infrastructure implementing a framework of spatial data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way. A further definition is given in **Kuhn (2005)**:

An SDI is a coordinated series of agreements on technology standards, institutional arrangements, and policies that enable the discovery and use of geospatial information by users and for purposes other than those it was created for.

-Standards:
OGC(R) standards are technical documents that detail interfaces or encodings. Software developers use these documents to build open interfaces and encodings into their products and services. These standards are the main "products" of the Open Geospatial Consortium and have been developed by the membership to address specific interoperability challenges. Ideally, when OGC standards are implemented in products or online services by two different software engineers working independently, the resulting components plug and play, that is, they work together without further debugging.

Metadata standards are requirements which are intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users. To achieve this common understanding, a number of characteristics, or attributes of the data have to be defined, also known as metadata.

- **Dublin-Core Metadata Initiative (DCMI):** The Dublin Core Metadata Initiative is an open organization supporting innovation in metadata design and best practices across the metadata ecology. DCMI's activities include work on architecture and modeling, discussions and collaborative work in DCMI Communities and DCMI Task Groups, global conferences, meetings and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and best practices.

- **Federal Geographic Data Committee (FGDC)**: The Federal Geographic Data Committee is an interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis. This nationwide data publishing effort is known as the National Spatial Data Infrastructure (NSDI). The NSDI is a physical, organizational, and virtual network designed to enable the development and sharing of this nation's digital geographic information resources. FGDC activities are administered through the FGDC Secretariat, hosted by the U.S. Geological Survey.

- **International Organization of Standardization (ISO)**: ISO is a network of the national standards institutes of 163 countries, one member per country, with a Central Secretariat in Geneva, Switzerland, that coordinates the system. ISO is a non-governmental organization that forms a bridge between the public and private sectors. On the one hand, many of its member institutes are part of the governmental structure of their countries, or are mandated by their government. On the other hand, other members have their roots uniquely in the private sector, having been set up by national partnerships of industry associations. Therefore, ISO enables a consensus to be reached on solutions that meet both the requirements of business and the broader needs of society.

ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. ISO 19115:2003 is applicable to the cataloguing of datasets, clearinghouse activities, and the full description of datasets, geographic datasets, dataset series, and individual geographic features and feature properties.

ISO 19119:2005 identifies and defines the architecture patterns for service interfaces used for geographic information, defines its relationship to the Open Systems Environment model, presents a geographic services taxonomy and a list of example geographic services placed in the services taxonomy. It also prescribes how to create a platform-neutral service specification, how to derive conformant platform-specific service specifications, and provides guidelines for the selection and specification of geographic services from both platform-neutral and platform-specific perspectives.

ISO/TS 19138:2006 defines a set of data quality measures. These can be used when reporting data quality for the data quality subelements identified in ISO 19113. Multiple measures are defined for each data quality subelement, and the choice of which to use will depend on the type of data and its intended purpose. The data quality measures are structured so that they can be maintained in a register established in conformance with ISO 19135. It must be remarked that ISO/TS 19138:2006 does not attempt to describe every possible data quality measure, only a set of commonly used ones.

ISO/TS 19139:2007 defines Geographic MetaData XML (gmd) encoding, an XML Schema implementation derived from ISO 19115:2003.

ISO 19157:2011 Under development., revises ISO 19115:2033 and some other standard specifications.

-**UncertML**: UncertML is a conceptual model and XML encoding designed for encapsulating probabilistic uncertainties. http://www.uncertml.org/.

-**User requirement**: In GeoViQua user requirements is related to requirements mainly described from a (end-) user perspective, as opposed to a technical or implementers' perspective. The requirements have been collected by different means, in particular user interviews and state of the art analysis, and subsequently integrated, reviewed and validated collaboratively. They represent a unified view of the systems GeoViQua intends to build as seen from a requirements perspective.

-User Interface Committee (UIC): The User Interface Committee engages users in the nine SBAs in the development, implementation, and use of a sustained GEOSS that provides the data and information required by user groups on national, regional and global scales.

The User Interface Committee has a specific goal to address cross-cutting issues by coordinating user CoPs, ensuring continuity and avoiding duplication.

The objectives of the UIC are:

-enable GEO to address the needs and concerns of a broad range of user communities in developing and developed countries, across issues and trans-disciplinary needs, with a particular focus on fostering new or less organized communities.

-enable GEO, in the implementation of GEOSS, to engage a continuum of users, from producers to the final beneficiaries of the data and information

-Facilitate linkages and partnerships between established CoPs and new groups or organizations interested in collaborating.

-W3C: PROV-DM is a data model for provenance for building representations of the entities, people and activities involved in producing a piece of data or thing in the world. PROV-DM is domain-agnotisc, but with well-defined extensibility points allowing further domain-specific and application-specific extensions to be defined. It is accompanied by PROV-ASN, a technology-independent abstract syntax notation, which allows serializations of PROV-DM instances to be created for human consumption, which facilitates its mapping to concrete syntax, and which is used as the basis for a formal semantics. **http://www.w3.org/TR/prov-dm/**

-Web Coverage Service (WCS): The OGC Web Coverage Service Interface Standard defines a standard interface and operations that enables interoperable access to geospatial "coverages" [http://www.opengeospatial.org/ogc/glossary/c]. The term "grid coverages" typically refers to content such as satellite images, digital aerial photos, digital elevation data, and other phenomena represented by values at each measurement point.
A WCS describes discovery, query, or data transformation operations. The client generates the request and posts it to a Web feature server using HTTP. The Web feature server then executes the request. The WCS specification uses HTTP as the distributed computing platform, although this is not a hard requirement.

There are two encodings defined for WCS operations:

-XML (amenable to HTTP POST/SOAP)
-Keyword-Value pairs (amenable to HTTP GET/Remote procedure call)

Data may be available in several formats, such as DTED, GeoTIFF, HDF-EOS, or NITF.

Several types of data layers are supported:

-series of points, such as locations of samples
-regular grid of pixels or points, which might represent a photo
-set of segmented curves, often used for road paths
-set of Thiessen polygons, used to analyse spatially distributed data such as rainfall measurements
-triangulated irregular network (TIN), often used for terrain models

Ranges of information may be attached to locations, such as average wind speed or yield by crop type.

-**Web Feature Service (WFS):** The OGC Web Feature Service Interface Standard provides an interface allowing requests for geographical features across the Web using platform-independent calls. One can think of geographical features as the "source code" behind a map, whereas the WMS interface or online mapping Portals like Google Maps return only an image, which end-users cannot edit or spatially analyze. The XML-based GML furnishes the default payload-encoding for transporting the geographic features, but other formats like shapefiles can also serve for transport. The OpenGIS GML Simple Features Profile is designed to both increase interoperability between WFS servers and to improve the ease of implementation of the WFS standard.
The OGC membership defined and maintains the WFS specification. There are numerous commercial and open source implementations of the WFS interface standard, including an open source reference implementation, called GeoServer. A comprehensive list of WFS implementations can be found at the OGC Implementing Products page [2].
The WFS specification defines interfaces for describing data manipulation operations of geographic features. Data manipulation operations include the ability to:

- get or query features based on spatial and non-spatial constraints
- create a new feature instance
- delete a feature instance
- update a feature instance

The basic Web Feature Service allows querying and retrieval of features. A transactional Web Feature Service (WFS-T) allows creation, deletion, and updating of features.

A WFS describes discovery, query, or data transformation operations. The client generates the request and posts it to a Web feature server using HTTP. The Web feature server then executes the request. The WFS specification uses HTTP as the distributed computing platform, although this is not a hard requirement.

There are two encodings defined for WFS operations:

- XML (amenable to HTTP POST/SOAP)
- Keyword-Value pairs (amenable to HTTP GET/Remote procedure call)

In the taxonomy of Web Services, WFS is best categorized as a non-RESTful RPC type service.

-Web Map Service (WMS): is a standard protocol for serving georeferenced map images over the internet that are generated by a map server using data from a GIS database. The specification was developed and first published by the Open Geospatial Consortium in 1999.
WMS specifies a number of different request types, two of which are required by any WMS server:

-GetCapabilities - returns parameters about the WMS and the available layers
-GetMap - with parameters provided, returns a map image

Request types that WMS providers may optionally support include:

-GetFeatureInfo
-DescribeLayer
-GetLegendGraphic

-Web Processing Service (WPS): The OpenGIS Web Processing Service Interface Standard provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services, such as polygon overlay. The standard also defines how a client can request the execution of a process, and how the output from the process is handled. It defines an interface that facilitates the publishing of geospatial processes and clients' discovery of and binding to those processes. The data required by the WPS can be delivered across a network or they can be available at the server.

8. REFERENCES

Díaz P., Masó J., Guimet J., "Comparative Quality Assessment of Metadata. Two Regional SDI case studies", Proceedings of INSPIRE Conference, Krakow, June 2010

GEO (2005) "The Global Earth Observation System of Systems 10-Year Implementation Plan" www.earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf

ISO/TC211 (2002) ISO 19113:2002, Geographic information-Quality principles, ICS: 35.240.70

ISO/TC211 (2003) ISO 19114:2003 Geographic information-Quality evaluation procedures, ICS: 35.240.70.

ISO/TC211 (2005) ISO 19114:2005 Geographic information-Quality evaluation procedures, ISO 19114:2003/Cor 1:2005

ISO/TC211 (2003) ISO 19115:2003, Geographic information-Metadata

ISO/TC211 (2009) ISO 19115-2:2009, Metadata pre-standard extension for imagery and gridded data

ISO/TC211 (2005) ISO 19119:2005, Geographic information-Services, ICS: 35.240.70

ISO/TC211 (2008) ISO 19119:2005/Amd 1:2008, Extensions of the service metadata model, ICS: 35.240.70

ISO/TC211 (2007) ISO 19139:2007, Geographic information-Metadata-XML schema implementation

ISO/TC211 (2006) ISO/TS 19138:2006, Geographic information-Data quality measures, ICS: 35.240.70

ISO/TC211 (under development) ISO/DIS 19157, Geographic information-Data quality, ICS: 35.240.70

Kuhn, W. (2005) "Introduction to Spatial Data Infrastructures". Presentation held on March 14th 2005.

Liu C.,White M., Newell G. (2011) "Measuring and comparing the accuracy of species distribution models with presence-absence data", Ecography 34: 232-243

Lush V., Lumsden J., Bastin L., Cornford D. (2011) "User Perspectives on Geospatial Data Quality", QA4EO, Oxford, October 2011

Parr D.A., Lu Y. (2010) "The landscape of GIScience Publications 1997-2007: An empirical investigation with latent semantic analysis", Transactions in GIS 14 (5): 689-708.

Masó J., Serral I., Pons X. (2011) "GEOVIQUA: a FP7 scientific project to promote spatial data quality usability: metadata, search and visualization", 7th International Symposium on Spatial Data Quality, Coimbra, 2011

Nativi S., Craglia M., Bertrand F., "EuroGEOSS: building inter-disciplinary interoperability for the global community", EGU General Assembly, Vienna, pp. 3440, 2009

O'Brien J., Marakas G. (2005) "Introduction to Information Systems", McGraw-Hill/Irwin; 13 edition, ISBN 0073043559

Růžička J. (2008) "ISO 19115 for GeoWeb services orchestration" Geoinformatics FCE CTU 2008, Volume 3, ISSN 1802-2669

Sevillano E., Díaz P., Ninyerola M., Masó J., Zabala A., Pons X. (2011) "Analysis of the quality metadata in GEOSS Clearinghouse", EuroGEOSS "Advancing the vision for GEOSS", Madrid 24-27th January 2012

van der Wel F.J.M., Hootsman M.R., Ormeling F. (1994) "Visualisation in Modern Cartography" Pergamon, London

Yang X., Blower J.D., Cornford D., Lush V., Masó J., Zabala A., Nüst D. (2011) "GEOSS Quality Framework and Quality Information Encoding", EuroGEOSS "Advancing the vision for GEOSS", Madrid 24-27th January 2012

Yang X., Blower J.D., Cornford D., Zabala A., Masó J., Bastin L., Lush V., Díaz P. (2012) "An integrated view of Data Quality in Earth Observation" (in preparation)