Analysis of quality metadata in the GEOSS Clearinghouse[*]

Paula Díaz[1], Joan Masó[1], Eva Sevillano[2], Miquel Ninyerola[2], Alaitz Zabala[2],
Ivette Serral[1] and Xavier Pons[2]

[1]CREAF, Universitat Autònoma de Barcelona; paula.diaz@uab.cat;
joan.maso@uab.cat; ivette@creaf.uab.cat

[2]Universitat Autònoma de Barcelona; eva.sevillano@uab.cat;
miquel.ninyerola@uab.cat; alaitz.zabala@uab.cat; xavier.pons@uab.cat

Abstract

The Global Earth Observation System of Systems (GEOSS) Clearinghouse is part of the GEOSS Common Infrastructure (GCI) that supports the discovery of the data made available by the Group on Earth Observations (GEO) members and participant organizations in GEOSS. It also acts as a unified metadata catalogue that stores complete metadata records, not only about datasets but also for other kinds of components and services. By exploring these records, users often try to find the fit-for-use data. Quality indicators and provenance are included in the metadata and are potentially useful variables that allow users to make an informed decision avoiding to download and to assess the data themselves. However, no previous studies have been made on the completeness and correctness of the metadata records in the Clearinghouse. The objective of this paper is to analyze the data quality information distributed by the GEOSS Clearinghouse. The aim is to quantify its completeness and to provide clues on how the current status of the Clearinghouse could be improved and how useful quality aware tools could be. The methodology used in the current analysis consists in first harvesting of the Clearinghouse and then quantify the quality information found in 97203 metadata records, by using a semi-automatic approach. The results reveal that the inclusion of quality information on metadata records is not rare: 19.66% of the metadata records contain some quality element. However, this is not general enough and several aspects could be

DOI: 10.2902/1725-0463.2012.07.art17

improved. For instance, 77.78% of quantitative measures lack measure units. When quality indicators are not sufficient, the lineage metadata information could be used to mitigate this situation by analysing the process steps and sources used to create a dataset. However, even though lineage is reported in 15.55% of the records, only 1.27% of the cases return a complete list of process steps with sources. This paper also provides indications on what is lacking in the current producer metadata model and, detected a gap in usage or user feedback metadata in GEOSS. Moreover, information extracted from GeoViQua interviews with users indicates that they value informal comments and user feedback on datasets as a complement of the more formal producer-oriented metadata description of the data. Although, many efforts within the scientific community and the Quality Assurance Framework for Earth Observation (QA4EO) group have been invested in describing how to parameterize data quality and uncertainty, we conclude that still extra work can be done to provide complete quality information in the metadata catalogues. In brief, since the GEOSS Clearinghouse references data from the most important agencies and research organizations, the results presented in this paper provide a perspective on how well quality is disseminated in the Earth observation community in general.

Keywords: Quality, GEOSS, metadata, ISO 19115, Earth observation, GCI

1. INTRODUCTION

The Global Earth Observation System of Systems (GEOSS), coordinated by the Group on Earth Observations (GEO), is a public infrastructure that interconnects a diverse and growing array of data, instruments and systems to allow monitoring and forecasting changes in a global environment (GEO, 2005). The GEOSS architecture task AR-07-01 (GEO, 2009) initialized the Interoperability Process Pilot Project, where the Components and Services Registry (CSR), the Standards and Interoperability Registry (SIR), the Clearinghouse and a Web Portal were designed and prototyped to promote the discovery of geospatial resources and the interoperability among diverse geospatial services (Bai et al, 2009). These four components are now operational parts of the GEOSS Common Infrastructure (GCI). The GEO Portal is broadly used by the scientific community when dealing with representations and models of the Earth System. Additionally, GEOSS gives support to policy-makers, resource managers and many other experts in their daily work with Earth observation data (Ollier et al, 2009). The AR-07-01 task has been recently replaced by the IN-05-C1 GEOSS Design and Interoperability (GEO, 2012) in the new GEOSS implementation plan.

The GCI is subject to continuous development and improvement in its functionalities, 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 (Fellous and Béquignon, 2010).

The GEOSS Clearinghouse is a metadata catalogue service that harvest the catalogues registered in the CSR and integrates the other components and services registered in the CSR, to facilitate a single entry point for data discovery and access (Christian, 2008). This current study explores the real content of quality metadata in the Clearinghouse within the current architecture. This effort ideally will result in a major contribution to the GEOSS infrastructure, as the results of the current analysis highlights the strengths and weaknesses in the GEOSS metadata. This is a state of the art needed to evaluate the information providers make available in the Clearinghouse, consequently to design the quality components development; in particular the “Quality elicitation mechanisms” and “Delivery of solutions to end users”, settled in the GEOSS 10-Year Implementation Plan (GEO, 2005).

Earth observation data sources are ideally elaborated following quality assessment procedures that gives quantitative values and conformance results (referred as ‘quality measures’ in this paper), resulting in quality estimates (referred as ‘quality indicators’ in this paper), alongside the lineage of the data and the estimates conform to the “producer quality information”. This methodology is summarized in a set of guidelines and recommendations edited by the Quality Assurance Framework for Earth Observation (QA4EO) group (Teillet and Chander, 2010). The quality information is needed to allow users deciding about data fit-for-use (Goodchild et al, 2007; van der Wel et al, 1994). Despite all the geographic data tools available to users, there are still issues to be fully addressed, such as the process needed to assure the quality of the information provided (Craglia et al, 2008). Furthermore, quality metrics are relevant in computing the fitness for use of the resource described in metadata records (Tolosana et al, 2006), as is also expected. Data quality is a difficult notion to define precisely and it has different meanings to different communities. ISO 9000 defines quality of a product 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”.

The studies that cover general metadata assessment are based on a wide range of methods and criteria. Although, the need for creating a common theory for metadata assessment has been expressed widely (Stvilia and Gasser, 2008), we consider this a difficult tasks to achieve in general. The complexity lies in the nature of the quality assessment, as it will have a different methodology depending on which standards are been used and what is the subject of the data. It is worth to keep in mind that only few papers have been published on quality of geospatial metadata and almost nothing on quality of the geospatial quality metadata. For that reason, most publications come from the bibliographic world and mainly based on Dublin Core (Bruce and Hillmann, 2004; Margaritopoulos et al, 2008; Stvilia and Gasser, 2008). The ISO 19115 (ISO-TC211, 2003) geospatial metadata is far more complex, therefore, it has to be adapted. However, some commonalities are maintained to analyse and compare catalogues. Some previous studies carried out in this matter highlight various methods to assess consistency, accuracy and relevance (Moen et al, 1998); correctness by conformance to a set or rules (Margaritopoulos et al, 2008); completeness and relation to the cost of improvement (Stvilia and Gasser, 2008); logical consistency and coherence, timeliness and accessibility (Bruce and Hillmann, 2004). Different methodologies on quality metadata assessment have in common the retrieving of selected records, the creation of a database and the quantification of more or less elaborated statistical results.

In our case, these methodologies were adapted to the geospatial quality metadata records. Our methodology focuses on the content of the entities of data quality information (DQ_DataQuality) in ISO 19115, for geographic information metadata. In the ISO vocabulary, an entity is a set of metadata elements describing the same aspect of data. These entities are the quality element (DQ_Element), which carries a quality indicator obtained by a quality measure; and the lineage (LI_Lineage). Additionally, we have also considered the usage information (MD_Usage).

According to ISO 19113 (ISO-TC211, 2002), quality indicators (DQ_Element) can be classified in five classes and fifteen indicators. The five different classes are:

- Completeness. Presence and absence of features, their attributes and relationships;
- Logical consistency. Degree of adherence to logical rules of data structure, attribution and relationships;
- Positional accuracy. Accuracy of the position of the features;
- Temporal accuracy. Accuracy of the temporal attributes and temporal relationships of the features;
- Thematic accuracy. Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of the features and their relationships.

Additionally, each quality element can have one or more measure methods and the indicators are expressed either by a numerical value, a conformance declaration with a methodology, or a per pixel value image, *i.e*. a coverage grid instead of an overall result; the latter introduced in ISO 19115-2 (ISO-TC211, 2009).

The lineage refers to information about the provenance of the dataset, including details of processing applied to it. The usage explains the specific uses of the data, the specific applications for which the resource was used and some determined limitations. The latter is a 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).

New standards, new services and new datasets will be added to the GCI that will allow building more complex applications. These applications, in turn, will enable a better understanding of our environment, combining initial components already operational with new ones (van Zyl et al, 2009). Just as an example, the current GCI has been extended in order to include other existing EO catalogues, exponentially increasing the number of resources linked by GEOSS. The EuroGEOSS broker (Nativi et al, 2009), a service that transparently distributes any query to other catalogues located outside the GCI, is becoming a key component of the GCI infrastructure. The Clearinghouse metadata records can be retrieved following the ISO TS211 schemas (ISO-TC211, 2007), so that the study centres the data quality extraction in ISO standards for geographic information, which are also adopted by the EuroGEOSS broker.

Despite all the efforts done, the quality metadata has not still been exploited and the GEO Portal has so far neither included quality information as search variables, nor a way to easily compare them. Moreover, quality information is useful to find fitting-for-use data by comparing metadata records. This matter is vital to GeoViQua (QUAlity aware VIsualisation for the Global Earth Observation system of systems, http://www.geoviqua.org/), an EC FP7 project that is developing tools to elicit search and visualize quality information in GEOSS (Masó et al, 2010). The project has studied the state of the art in quality metadata currently in the Clearinghouse. This paper exposes the findings of this study and establishes recommendations, as a basis for a future GeoViQua Quality Broker. This projected broker will improve search capabilities within GEOSS by extending the current EuroGEOSS broker component and allowing quality aware queries and results.

In principle, it is possible to define quality at a hierarchy of levels, from a single attribute or measurement of position through entire features, entire layers, and entire seamless product (Devillers et al, 2005). Each of these sets of features has very different data quality characteristics that are difficult to capture in a single data quality indicator according to a metadata standard (Bai et al, 2009). In ISO 19115 each quality element can be associated to a piece of information in the hierarchy using a “ScopeCode”, which, according to ISO 19115 (ISO-TC211,

2003) is a “class of information to which the referencing entity applies”. That is to say, that in the current ISO model providers can inform about quality in different levels: layer, feature or pixel level.

In the current study, the complete Clearinghouse metadata content up to November 2011 was automatically extracted and analyzed in terms of its quality information. The Clearinghouse follows the Open Geospatial Consortium OGC CSW standard (Nebert, 2007), whereas the metadata retrieved follows the aforementioned ISO 19115 standard. Section 2 of this paper explains the quality scope and the methodology; section 3 refers to the analysis of the quality metadata in the Clearinghouse and section 4 closes with the discussion and conclusion.

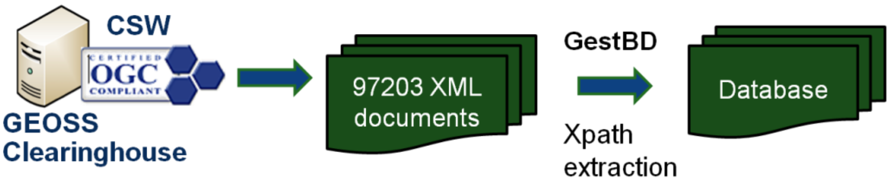
2. SCOPE AND METHODOLOGY

This section describes the methodology used to download and analyze the Clearinghouse metadata. Also, the filtering applied to focus the analysis on the quality metadata indicators in ISO 19115 standard (ISO-TC211, 2003) is described. The aim is to analyze the quality metadata and assess the completeness and the quality of the current content of the Clearinghouse and generate recommendations for improvement. The methodology applied in the current analysis had been previously tested, and therefore enhanced, in the assessment of data quality and metadata records in a regional Spatial Data Infrastructure (SDI) (Díaz et al, 2010).

2.1. Structuring Metadata XML Contents into Databases

In the Clearinghouse, each metadata record refers to a discoverable resource in GEOSS, such as datasets, services, portals, etc. The Clearinghouse is based on the CSW standard in the ISO 19139 profile (ISO-TC211, 2007). CSW protocol defines some queries to get metadata records fitting with the request conditions and applying filtering operations. Nevertheless, the current version of the catalogue standard does not allow easily extracting statistical summaries by metadata elements. The alternative presented consists in performing a massive extraction of quality entities (DQ_DataQuality), including quality indicators (DQ_Element), lineage (LI_Lineage) and usage (MD_Usage), from the metadata records in the Clearinghouse, done by means of the methodology presented in Figure 1.

Figure 1: Flow Diagram of the Extraction of Metadata from the Clearinghouse and Generation of a Database of Selected Items Relevant to Quality Data.



The first step of this methodology consists of a massive downloading of the metadata XML files contained in the Clearinghouse. The several steps are:

- First, the metadata records are harvested by repetitively requesting the single records by file identifier. These records are saved in individual XML files. The total number of Clearinghouse metadata records collected was 97203.
- Second, the precise information required for the analysis is selected, considering the relations amongst entities, the multiplicity and the different ways in which the metadata entities can be related. In this step we prepare the appropriate Xpath sentences, allowing the multiplicity. For instance, in lineage information, a resource can have many sources related to several process steps or viceversa.
- Third, the information contained in the XML files is extracted by requesting it through XPath language using batch files. This information is organized in database tables in which the rows are the XML files that contain quality information, unmistakably identified by their file identifier; and the columns contain the values of the corresponding quality, lineage and usage elements. Once the database tables are generated, the information is analyzed and the results are summarized.

The final database tables contain up to 52332 rows and 50 columns, all the tags extracted are listed in the Table 1. The length of the table corresponds to the number of quality indicators extracted, not corresponding to the metadata records for two reasons. First, not all the metadata records contain quality indicators; and second, a metadata record can contain several quality indicators. Such structure permits an enhanced examination of the information, comparison and statistics calculation, otherwise unfeasible in a massive analysis of XML file folders. The extraction of XML files from the Clearinghouse and the ensuing extraction of quality indicators, performed prior to the main stages of the analysis, were undertaken employing MiraMon modules (Pons, 2002). The quality elements to be analysed were extracted from the XML files using XPath queries. For instance, the scope code is retrieved using the following XPath: “/gmd:MD_Metadata/gmd:dataQualityInfo[0]/gmd:DQ_DataQuality/gmd:scope/gmd:DQ_Scope/gmd:level/gmd:MD_ScopeCode/@codeListValue”.

Table 1: List of the Extracted Tags from the Clearinghouse Metadata, Classified by Entities in ISO 19115.

Report	Lineage	Usage
Scope	Extent description	Description
Scope value	Extent geometrical	Date time
Quality scope description	Extent temporal	Limitation
Element	Extent vertical	Responsible party
Name of measure	Quality scope code	
Measure identification	Statement	
Measure description	Description[1]	
Evaluation method description	Rationale	
Date time	Datetime	
Result	Process responsible	
Specification	Description[2]	
Explanation	Denominator[2]	
Pass	Citation[2]	
Value type	Quality extent[2]	
Value unit	Source extent[2]	
Error statistic	Description[3]	
Value	Denominator[3]	
File identifier[4]	Citation[3]	
	Extent[3]	
	Geometrical element	
	Temporal element	
	Vertical element	

[1] refers to the direct list of processes.
[2] refers to the list of processes describing sources.
[3] refers to the direct list of sources.
[4] belonging to MD_Metadata.

2.2. Quality Extraction from Metadata

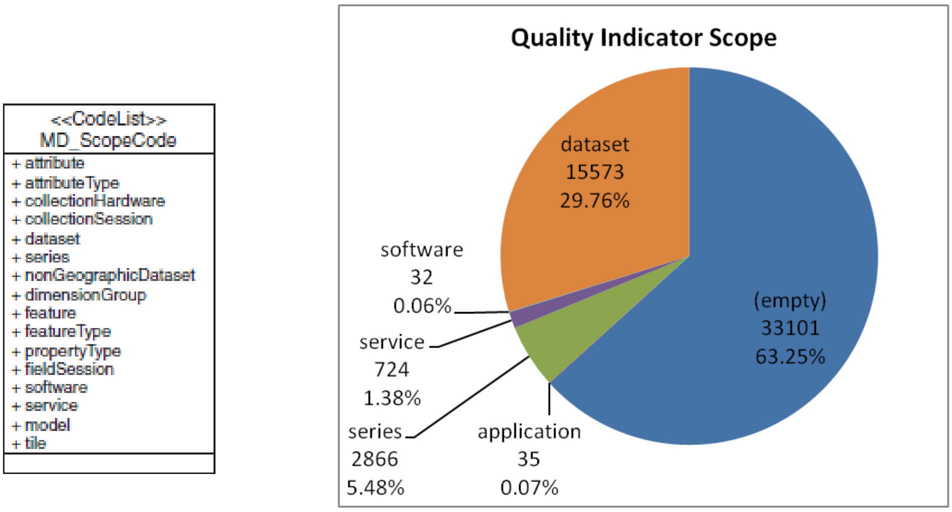
The most relevant decisions are focused on the extraction of quality contents conformant the standard ISO 19115, which represent the complete consideration of all possible quality classes and the lineage. Therefore, twenty-five different elements were selected for the extraction, taking multiplicity into consideration. Up to 16 results for each quality indicators and the first measure method for each quality indicator were compiled. The corresponding indicators were then collected, having either a numerical value, or a conformance declaration with a methodology, or even a per pixel value. In this sense, note that the per pixel value is an increasingly important feature in quality information, especially for

Earth observation data (Cressie and Kornak, 2003) but also in other disciplines (Xiao et al, 2007). Finally, twenty lineage elements were extracted considering a multiplicity of up to 8 and five usage elements were also extracted considering a multiplicity of up to 8 for each metadata record.

2.3. Quality Scope.

The quality described in metadata following ISO 19115 could refer to various hierarchy levels, as described previously. In ISO 19115 the different levels are described as "ScopeCode". The following figure represents the results obtained in the analysis of the scope code list present in the Clearinghouse metadata records. On the left hand side (figure 2a), the UML code list of this feature as specified in the ISO 19115 standard is shown. On the right hand side (figure 2b), the summary plot of scores in the metadata records is represented.

Figure 2: a) Scope Code List under ISO 19115 (MD_ScopeCode). b) Plot of Results of the Detailed levels of Data Quality from the Metadata contained in the Clearinghouse in 2011



The first classification of the quality information is related to the hierarchy of levels, or "ScopeCode" (DQ_Scope), in order to determine the quality scope in the Clearinghouse, and establish an ad-hoc methodology. The 33101 quality scope empty indicators (63.25%) in the metadata records are unfortunately a common finding in studies of this kind (Díaz et al, 2010). This could be explained by the fact that this element is not mandatory in the in ISO 19115 standard. Within the metadata containing scope code, "dataset" represents the highest percentage, with up to 15573 metadata records (29.76%). Other attributes of the scope code list present in the metadata records are: series (2866, 5.48%), service (724, 1.38%), application (35, 0.07%), and software (32, 0.06%).

3. RESULTS ON QUALITY ANALYSIS

3.1. Quality Elements.

Besides the measure methods used to evaluate the quality of the dataset, mentioned earlier in the text, the ISO 19115 describes an entry for the quality measure. This apply to the actual quantification of quality indicator (*e.g.*, root mean square error, value at 95% confidence level), obtained from ISO 19114 (ISO-TC211, 2003) and ISO 19138 (ISO-TC211, 2006). Thus, there are three concepts related to the quality item: quality class, quality indicator and quality measure, *i.e.* test applied to evaluate data quality indicator and the actual measured value. The table 2 summarizes the abovementioned ISO standards quality indicators and its definition. The next figure (Figure 3) is a representation of the ISO 19115 UML schema in which quality elements are structured.

Figure 3: Metadata Report UML Model in ISO 19115

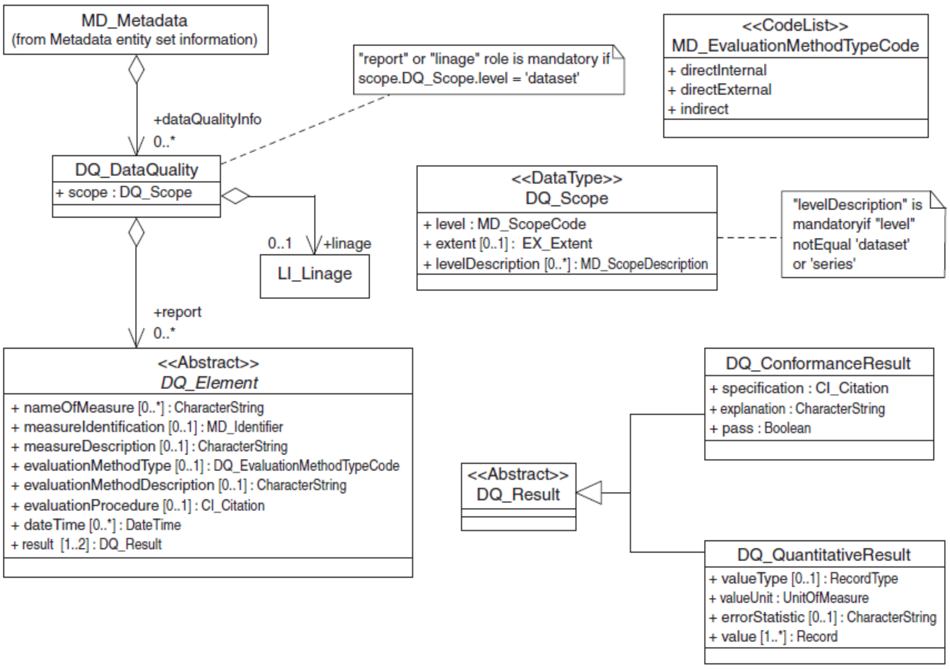

MD_Metadata
(from Metadata entity set information)
"report" or "linage" role is mandatory if scope.DQ_Scope.level = 'dataset'
<<CodeList>>
MD_EvaluationMethodTypeCode
+ directInternal
+ directExternal
+ indirect
+dataQualityInfo
0..*
DQ_DataQuality
+ scope : DQ_Scope
0..1
+linage
LI_Linage
<<DataType>>
DQ_Scope
+ level : MD_ScopeCode
+ extent [0..1] : EX_Extent
+ levelDescription [0..*] : MD_ScopeDescription
"levelDescription" is mandatoryif "level" notEqual 'dataset' or 'series'
+report
0..*
<<Abstract>>
DQ_Element
+ nameOfMeasure [0..*] : CharacterString
+ measureIdentification [0..1] : MD_Identifier
+ measureDescription [0..1] : CharacterString
+ evaluationMethodType [0..1] : DQ_EvaluationMethodTypeCode
+ evaluationMethodDescription [0..1] : CharacterString
+ evaluationProcedure [0..1] : CI_Citation
+ dateTime [0..*] : DateTime
+ result [1..2] : DQ_Result
DQ_ConformanceResult
+ specification : CI_Citation
+ explanation : CharacterString
+ pass : Boolean
<<Abstract>>
DQ_Result
DQ_QuantitativeResult
+ valueType [0..1] : RecordType
+ valueUnit : UnitOfMeasure
+ errorStatistic [0..1] : CharacterString
+ value [1..*] : Record

Table 2: Classification and Definition of Quality Elements by ISO 19115.

Quality Class	Quality indicator	Definition
Positional accuracy	Absolute or external positional 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 positional accuracy	Closeness of gridded data position values to values accepted as or being true
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
	Topological consistency	Correctness of the explicitly encoded topological characteristics of a dataset
	Format consistency	Degree to which data is stored in accordance with the physical structure of the dataset
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	Quantitative attribute accuracy	Accuracy of quantitative attributes
	Non-quantitative attribute correctness	Correctness of non-quantitative attributes
	Thematic classification correctness	Comparison of the classes assigned to features or their attributes to a universe of discourse (*e.g.,* ground truth or reference dataset)

The quality indicators presented above are a small subset of the possible approaches to represent quality of data and there is a bias towards vector data, given that raster data is unrepresented. Concepts such as contingency tables, which rely on a whole family of indicators (similar to kappa statistic), or the true skill statistic can also be considered (Liu et al, 2011). These quality indicators are commonly used in geospatial datasets, usually conforming important fragments of the metadata, and allowing users to judge whether the dataset is fit for their use. Nevertheless, the quality not only informs the user about the fitness for use of the data in certain fields of application, but also enables the user to interpret the results from a data processing analysis (Donaubauer et al, 2008). Quality indicators are optional, but can appear more than once (multiplicity represented by "(0..*)" in the table 3), so that, the number of quality indicators and measures can be higher than the metadata records with quality indicators.

Table 3: Summary of the Analysis of the Clearinghouse Metadata Quality Data Content

Total metadata records	97203
Metadata records with quality information of any kind	87491
Metadata records with quality elements	19107
Total quality indicators (0..*)	52187
Total quality measures (0..*)	25944

3.1.1. Quality Elements Analysis

In this section of the analysis we focus on data quality indicators (DQ_Elements ISO elements). These quality indicators, as represented in table 2, are classified in positional accuracy, completeness, logical consistency, temporal accuracy and thematic accuracy.

The overall number of metadata records with quality indicators is 19107, which represents 19.66% of the total metadata records, a number far from the ideal situation. Nevertheless, it reflects that some members of the Earth observation community are sensitive to this information and know how to communicate it.

In turn, the 19107 metadata records contain a total of 52187 quality indicators which results in a mean of 2.7 quality indicators per record. Table 4 shows a more detailed analysis of this data, summarizing the quality indicators by classes (as classified in figure A.6 in ISO 19115 (ISO-TC211, 2003)).

Table 4: Classes of Quality Indicators in ISO 19115 and the statistics corresponding to the GEOSS Clearinghouse.

Quality class	Quality indicator	Number	%
Positional accuracy	Absolute or external positional accuracy	17767	34.04
	Gridded data positional accuracy	1364	2.61
	Relative or internal positional accuracy	280	0.54
		19411	***37.19***
Completeness	Commission	9815	18.81
	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 data quality indicators	**52187**	

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

Breaking down into quality indicators, and 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 metadata records with quality indicators (34.04%). Remarkably, an interesting finding is the 1364 number corresponding to gridded data positional accuracy.

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%). The results within logical consistency subclasses are not equally distributed, 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.

Regarding temporal accuracy, accuracy of a time measurement is clearly the dominant subclass, followed by temporal consistency. Temporal validity indicators were not present in the metadata records. Finally, the most representative thematic accuracy subclass is quantitative attribute accuracy, in distinctly higher numbers than non-quantitative attribute accuracy, whereas thematic classification correctness produced no results in the metadata records.

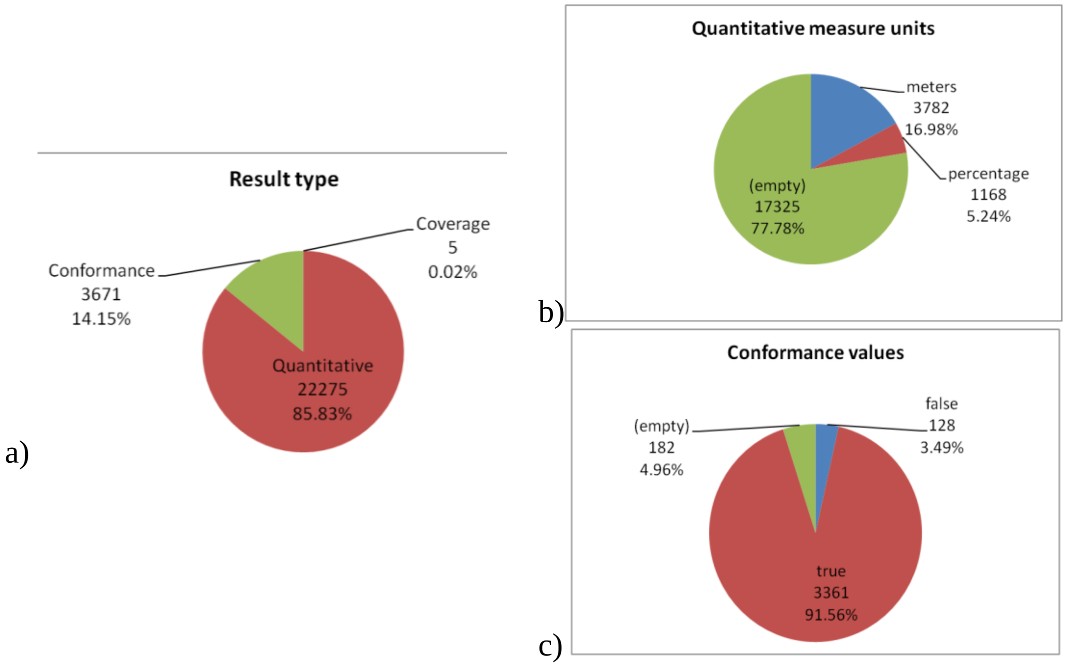
3.1.2. Quality Measures Analysis

In the ISO 19115 standard specification quality measure values can be either expressed by a numerical quantitative measure or by a conformance measure, which means determining whether the data product is compliant with an acknowledged quality test or specification. In turn, these indicators can be described using measures and values (*e.g*. “root mean square error with 1 pixel of error”).

A more detailed analysis in the metadata records reveals that there are 25944 measures (table 3). Sometimes a quality element can be expressed in more than one measure. On the other hand, not every quality element mentioned has its corresponding measure; for instance, a producer can just describe the quality element applied in the dataset without providing any further information. As figure 4a shows, the measures can be classified in quantitative measures (22275-85.83%) and in a conformance declaration to a specification (3671-14.15%). In the case of the Clearinghouse, the conformance measures are mainly referred to INSPIRE directive (3669-14.14%). This can be considered a good sign, implying that some European providers are following the mandatory directive regarding SDI in Europe.

Most interestingly, with respect to the Earth observation datasets, is that quality measure 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. Unfortunately, even though the five coverage results were found in the Clearinghouse, no valid link to the quality distribution file is provided.

Figure 4: a) Summary of Result Types. b) Units for Quantitative Values. c) Conformance Statements

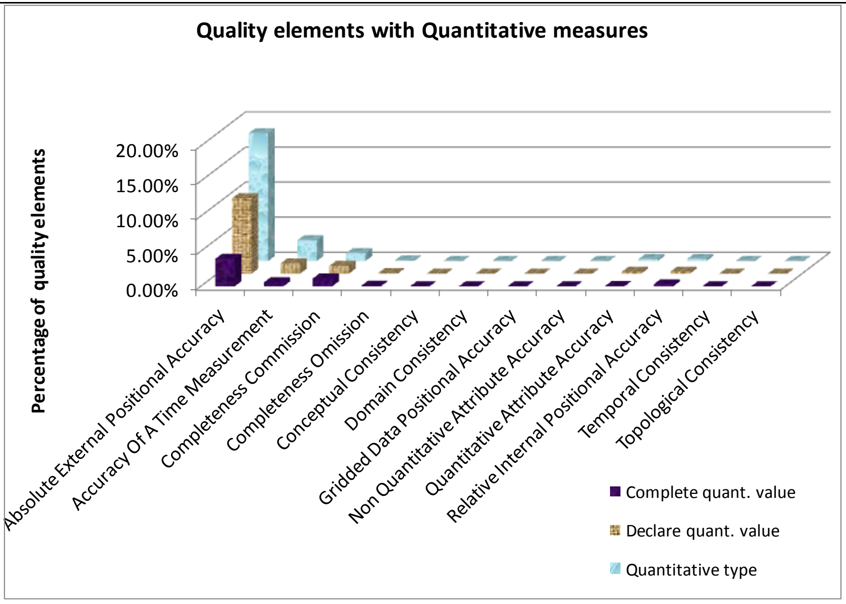


Going deeper into quantitative values, the figure 4b shows that although there is a quantitative numerical value, in most data products (17325, representing around 77.78%), the units in which the quantitative values are provided are missing (see complete value category in figure 5a). This is an essential gap identified in the study that requires 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. Another relevant unit is percentage, having 1168 measures and representing 5.24%.

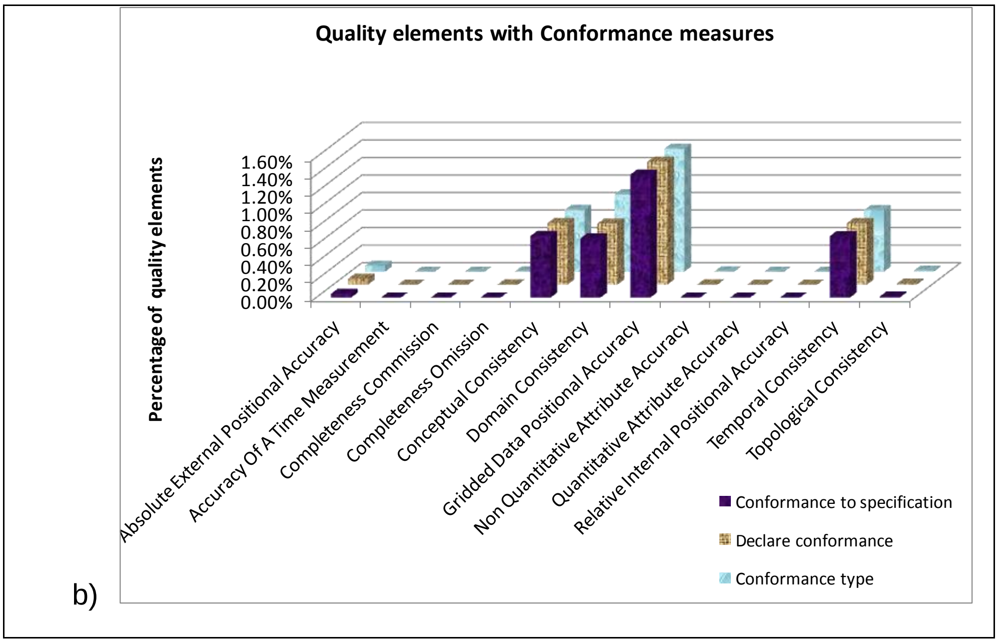
Regarding conformance results (figure 4c), 3361 (91.56%) metadata state the data product is conformant with a certain specification, in contrast with the 128 (3.49%) false statements. Only a few fail to report conformance status and do not declare to which specification they are in conformance with (see conformance to a specification category in figure 5b).

More detailed analysis for each quality indicator is detailed in figure 5. The higher bar represents all the quality indicators declaring the measure type, either quantitative (figure 5a) or qualitative (figure 5b). The middle bar represents the number of indicators containing the indicator type and also a value when quantitative or the conformance status when qualitative. The shorter bar represents the number of indicators containing any indication of the measure technique (*e.g*. name of the measure, description, etc.) and units if quantitative (*e.g.* "root mean square error with 1 pixel of error"); or an indication of the specification the data is in conformance with, if qualitative (*e.g*. "it is true that is conformant with INSPIRE"). The figures 5a and 5b represent that the more complete information we search the lower results number we obtain.

Figure 5: Quality Indicators Filtered by the Completeness of the Provided Information. a) Quantitative Measures b) Conformance Measures.



a)



3.2. Lineage

The ISO 19115 specification model for lineage is shown in figure 6. Lineage offers at least 4 possible options: a list of sources, a list of process steps, a list of process steps that use sources and a list of sources linked with process steps. From our point of view, the third option provides the better way to report a complete record on provenance.

In the following figures the number of lineage elements and the number of metadata records contained on them are shown. The multiplicity (represented by the symbol "(0..*)") of sources and processes in the lineage information should be considered, therefore, the number of lineage elements can be higher or smaller than the number of metadata records with source elements.

The Clearinghouse has 5851 (6.02%) metadata records using the first option, the list of sources. These include 1798 (1.85%) metadata records with temporal elements, which are extent (see table 5). Note that 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. This is the simplest way to highlight that this resource was derived from previous sources, at the same time giving credit to the providers (*i.e.,* provided attribution, and eventually made the sources trustworthy). Moreover, this option enables some form of descriptive quality report based on sources. In this case, if quality indicators are not provided for this dataset, quality indicators from the source records could still provide a clue of the quality.

Figure 6: Lineage UML Schema under ISO 19115, including Source (LI_Source) and Process step (MD_ProcessStep)

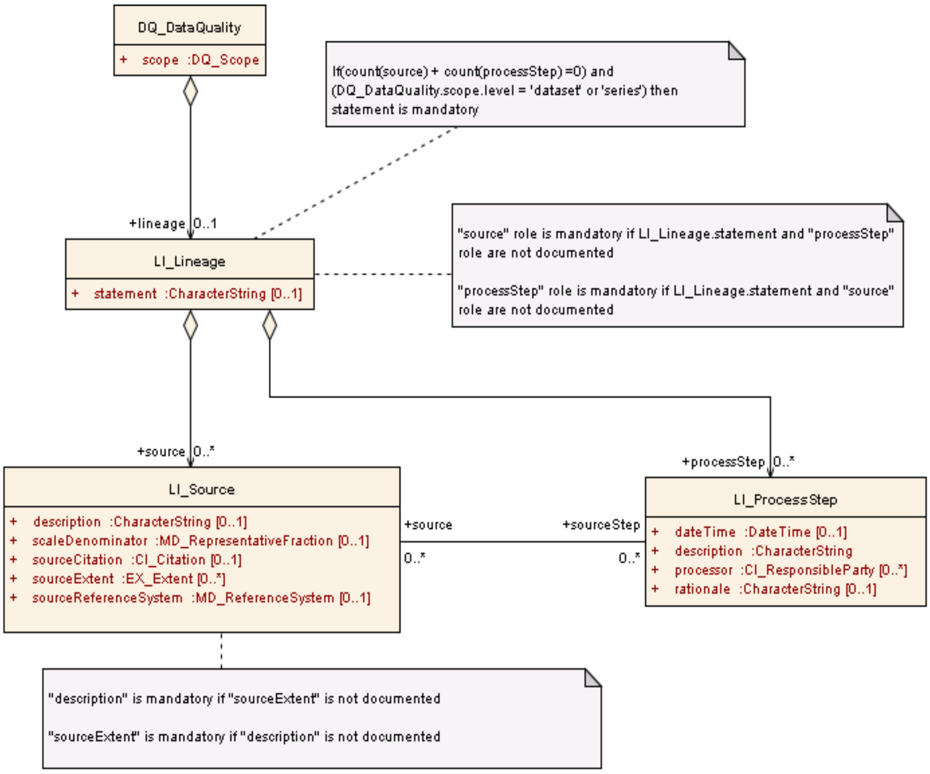

DQ_DataQuality
+ scope :DQ_Scope
If(count(source) + count(processStep) =0) and (DQ_DataQuality.scope.level = 'dataset' or 'series') then statement is mandatory
+lineage
0..1
LI_Lineage
+ statement :CharacterString [0..1]
"source" role is mandatory if LI_Lineage.statement and "processStep" role are not documented
"processStep" role is mandatory if LI_Lineage.statement and "source" role are not documented
+source
0..*
+processStep
0..*
LI_Source
+ description :CharacterString [0..1]
+ scaleDenominator :MD_RepresentativeFraction [0..1]
+ sourceCitation :CI_Citation [0..1]
+ sourceExtent :EX_Extent [0..*]
+ sourceReferenceSystem :MD_ReferenceSystem [0..1]
+source
0..*
+sourceStep
0..*
LI_ProcessStep
+ dateTime :DateTime [0..1]
+ description :CharacterString
+ processor :CI_ResponsibleParty [0..*]
+ rationale :CharacterString [0..1]
"description" is mandatory if "sourceExtent" is not documented
"sourceExtent" is mandatory if "description" is not documented

Table 5: Direct List of Source Elements 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 (extent)	**0**
Metadata with source elements		**5851**
Of which	metadata with temporal element (extent)	1798

Referring to the second option, the list of process steps, 9261 metadata records (9.53%) describe the processes, containing either processes or processes linked to sources. Among these, 8035 (8.26%) metadata records have a list of processes without mentioning sources. There are 292 (0.30%) providing the process date (see table 6). Note that 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. The direct list of processes provides information on the exact processes execution and the order of these executions. Having this option without any data source information, it is difficult to infer the quality of this resource.

Table 6: Direct List of Process Step Elements in the Metadata extracted from the Clearinghouse

Process Elements	Description	**12914**
	Process responsible	**1800**
	Date and time	**437**
	Rationale	**15**
Metadata with process elements		**9261**
Of which	processes without mentioning sources	8035
	MD with Date and time element	292

Referring to the third option, only 1226 metadata records (1.26%) have been identified in the Clearinghouse (table 7). This option provides a list of processes execution, the order of these processes and how and when the data sources were used. With this information it is possible to infer which sources represent a higher influence over quality and the associated final result (Moré and Pons, 2011). No record containing the fourth option was found.

Table 7: Summary of Results of Metadata following the Complete Provenance Path including Linking Process Steps and Sources

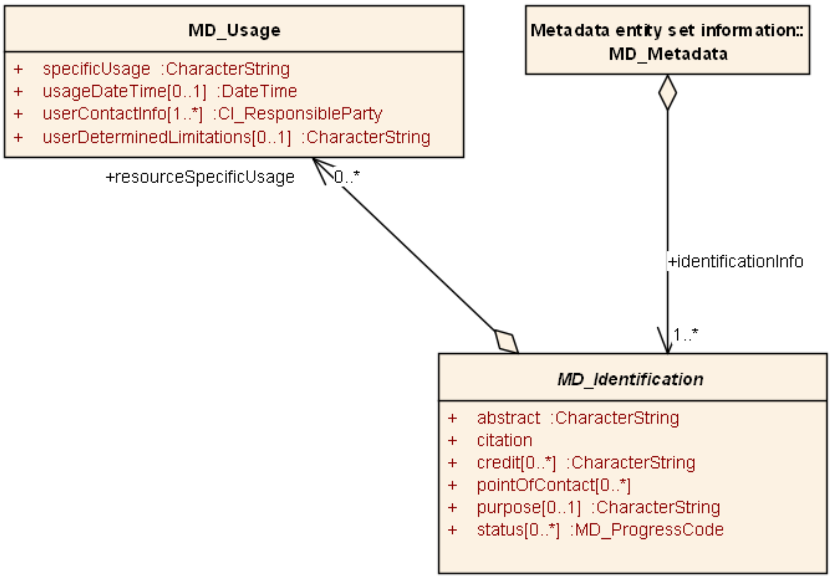
Process Elements with Sources	Citation	**2094**
	Description	**0**
	Scale denominator	**0**
	Source extent	**0**
Metadata with process elements		**1226**

3.3. Usage

There is one specific entity in ISO 19115 intended for enabling producers provide a brief description of ways in which the resource is currently being used, or has been used, which is the usage information (MD_Usage; see figure 7). Producers can get this information from uses establishing some reporting channels.

Even though in the analysis performed 1133 records were identified containing usage information (1.17%), only the mandatory specific usage and user contact information elements were described. Additionally, it could contain a brief description of procedures of using data, the list of resources, the date and time of the use, and limitations of use.

Figure 7: Metadata Usage UML Model in ISO 19115


MD_Usage
+ specificUsage :CharacterString
+ usageDateTime[0..1] :DateTime
+ userContactInfo[1..*] :CI_ResponsibleParty
+ userDeterminedLimitations[0..1] :CharacterString
Metadata entity set information::
MD_Metadata
+resourceSpecificUsage
0..*
+identificationInfo
1..*
MD_Identification
+ abstract :CharacterString
+ citation
+ credit[0..*] :CharacterString
+ pointOfContact[0..*]
+ purpose[0..1] :CharacterString
+ status[0..*] :MD_ProgressCode

4. DISCUSSION

The overall conclusion of the analysis is that the number of metadata records without quality information of any kind is relatively small (about 10%); but the completeness of the quality information is not achieved. When providing quality indicators, either conformance or quantitative, producers don't provide the necessary elements to quantify the real quality. An increase of the awareness and training on how these quality indicators can be obtained, recorded and propagated is needed, complemented by more research in automatic ways of capturing metadata.

Regarding more concrete findings related to the ISO scope, note that “application” is not in the code list. It could be interpreted as a synonym of “software” and if so, the results obtained for application should be considered valid and added to the software category in the results. This is an example of the semantic issue that need to be addressed in further interoperability studies. In addition, 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. On the other hand, an important immediate result is that all quality indicators, and almost all indicators, are described in the GEOSS Clearinghouse, representing the five different classes. Particularly, the high number of records with completeness, consistency and temporal accuracy quality indicators contrast with the previously published study about the Spanish regional SDIs, in which they do not represent more than 5% overall (Díaz et al, 2010).

In relation to the quality indicators expressed as conformance results, we might wonder why a metadata provider would explicitly state not being compliant with a particular quality assessment methodology; in any case, it is an interesting finding proving quality awareness by the provider and their efforts to carefully follow the standards. The number of empty conformance values is lower compared to the quantitative values. Nevertheless, a more detailed examination unveils that quality measure reported are far from complete. Indeed, almost half of the measurements provide only quantitative numbers not indicating neither their units nor the methodology used to determine them. This reveals the need for more clear tools, tutorials and a registry of measure types helping users to completely describe the quality indicator. A step forward in this direction is to include UncertML (http://www.uncertml.org/) in the quantitative values and extend the same solution for categorical variables. One of the most important sources of Earth observation data is remote sensing imagery. A surprising finding of this study is that ISO 19115-2 earth observations extensions for the quality part are almost unused. In fact, quality indicator that provide a per pixel quality index (coverage results) as a quality measure are commonly obtained by remote sensing production products; but coverage quality distribution appears only 5 times and the use of lineage extensions were not found. ISO 19115-2 adds an extension of process step called LE_Processing that contains a “runTimeParameters” attribute allowing the description of the exact list of parameters used in the execution of the process. This extension enables recording the exact information needed to repeat the execution of the process, so that if the uncertainties on the sources are known, and the process supports quality propagation, the quality of the resource can be reassessed. In addition, a citation of the used algorithm (LE_Algorithm) can also be included. No record using this extension was found in the Clearinghouse. A possible reason for this is that the standard is relatively new and not well known; besides, most of the metadata tools do not provide support for it (Zabala and Pons, 2002). Even, the

ISO 19115 is seen as too complex and arduous to be rigorously followed manually (Batcheller, 2008); so, ISO metadata extensions make the situation even worse. This confirms again the need for new tools; preferable automatic tools that help producers transfer the production process information into metadata records, guidance and good practices.

Referring to the lineage, it is important to highlight that, although abundant information about lineage has been found, automated update of metadata remains a largely elusive goal, particularly in the area of data quality, because of the difficulties associated with processing metadata, (Goodchild, 2007). Rich lineage information is found in about 10% of metadata records. The fact that some of the documents present even more than a hundred sources and process steps entries unveils its importance for producers and the fact that the right tools to describe processes and sources exists and are used in some domains. Nevertheless, the low percentage of metadata record containing lineage information reveals that more work is needed to generalize this practice.

Regarding usage information, curiously, all records were provided by the same institution, which, in the end, produces a non representative scenario. The main reason for this is that ISO 19115 standard lacks the right emphasis in user feedback, which could be an important addition to the current producer oriented metadata description (Goodchild, 2007). The current metadata ‘usage’ entity has demonstrated to be insufficient as a way to convey this information. Providing an agile way for users to contribute to add extra information, *e.g.* providing a mechanism to add rich comments to datasets properly liked to the corresponding producer metadata records, could be an important component to add to the GCI, as well as an important contribution to GEOSS. Additionally, these is a need for a seamless harmonizing of both producer oriented ISO metadata and more informal user feedback inputs.

5. CONCLUSION

The main objective of this work was to determine the current status of the GCI metadata records in terms of quality metadata. Results are part of the requirements study conducted by the GeoViQua EC FP7 project, and will help to make informed decisions on how the project can contribute to the improvement of the quality metadata and the development of tools to visualize information with its quality uncertainties. These results will enable to infer what quality indicators’ queries will have better effect on filtering results in the GEO Portal search engine. Better tools for metadata editing and producing, connected to the data generation processes are needed. In GeoviQua, attention will be put on creating a user feedback component for GEOSS integrated in the GCI; and in providing formats and tools for encoding pixel based indicators, such as the NetCDF-U format conventions and NetCDF editors.

Despite the current difficulties, the adoption of geospatial standards provide obvious advantages: allowing communication, comparison and minimizing data integration efforts and enabling an interdisciplinary system of systems that can benefit several communities of practice. The new GCI connects tens of new catalogues through the EuroGEOSS broker that also convey quality information. In the future, new tools will be integrated in GeoViQua Broker and in the GEO Portal, such as a method for sorting the search results, allowing specialised search, inserting quality indicators and measure value thresholds as filters in a search, making the result of the search more understandable for users, and providing a method to easily intercompare dataset metadata parameters. It also will enhance visualization components in a way that a well presented quality information will accompany map views and will increase the trust on the data products integrated in GEOSS. The GeoViQua project will contribute to GEOSS in providing such these tools.

ACKNOWLEDGEMENTS

This paper was written thanks to the support of the European Commission through the FP7- 265178- GeoViQua (ENV.2010.4.1.2-2) and a grant to the Consolidated Research Groups given by the Catalan Government (2009 SGR 1511). Xavier Pons is recipient of an ICREA Acadèmia Excellence in Research grant (2011-2015).

REFERENCES

Bai, Y., Di, L., and Y. Wei (2009). A taxonomy of geospatial services for global service discovery and interoperability. *Computers & Geosciences*, 35(4): 783-790.

Batcheller, J. K. (2008). Automating geospatial metadata generation—An integrated data management and documentation approach. *Computers & Geosciences,* 34 (4): 387-398

Bruce, T. R., and D. I. Hillmann (2004). The continuum of metadata quality: Defining, expressing, exploiting. in D. Hillmann and E Westbrooks (eds.), *Metadata in Practice* , pp. 238-256. Chicago: ALA Editions. ISSN: 0-8389-0882-9

Christian E.J. (2008). GEOSS Architecture Principles and the GEOSS Clearinghouse. *IEEE Systems Journal*, 2(3): 333-337.

Craglia, M., Goodchild, M.F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S., Parsons, E. (2008). Next Generation Digital Earth: A position paper from the Vespucci initiative for the advancement of geographic information science. International Journal of Spatial Data Infrastructures Research, 3:146-167.

Cressie, N. and J. Kornak (2003). Spatial Statistics in the Presence of Location Error with an Application to Remote Sensing of the Environment. *Statistical Science*, 18(4): 436-456.

Devillers, R., Bédard, Y. and R. Jeansoulin (2005) Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS. *Photogrametric Engineering and Remote Sensing*, 71(2): 205-215.

Díaz, P., Masó, J. and J. Guimet (2010), Comparative Quality Assessment of Metadata. Two Regional SDI case studies, *Proceedings of INSPIRE Conference, June 22-25 2010, Kraków, Poland*.

Donaubauer, A., Kutzner, T. and F. Straub (2008). Towards a Quality Aware Web Processing Service, *Accuracy 2010 symposium, July 20-23, Leycester, UK.*

Fellous, J. L. and J. Béquignon (2010). *A report prepared by the European Space Agency in the framework of the GEO Science and Technology*, GEO and Science.

GEO (2005). The Global Earth Observation System of Systems 10-Year Implementation Plan, Group on Earth Observations, (February). at: www.earthobservations.org/docs/10-Year Implementation Plan.pdf, [accessed 14 January 2012].

GEO (2009). AR-09-01a GEOSS Core Architecture Task Report, (September), at:http://www.earthobservations.org/documents/committees/adc/200909_11thADC/AR-09-01a_20090915.pdf [accessed 24 March 2012].

GEO (2012). Task IN-05 GEOSS Design and Interoperability, GEO Work Plan Symposium (April), at: ftp://ftp.earthobservations.org/201205_Work_Plan_Symposium/Presentations/3_Work Plan Review/Infrastructure/GEO_WPS1205_Presentation_IN-05.pdf [accessed 14 May 2012].

Goodchild, M.F. (2007). Beyond Metadata; towards user-centric description of Data Quality. *Keynote presentation at the 5th International Symposium on Spatial Data Quality, June 13-15, Eenschede, Netherlands*.

Goodchild, M.F., Fu, P. and P. Rich (2007). Sharing geograpic information: An assesment of the Geospatial One-Stop. *Annals of the Assosiation of American Geographers*, 97(2): 250-266.

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 (2003). ISO 19115:2003, Geographic information-Metadata. ICS: 35.240.70

ISO-TC211 (2009). ISO 19115-2:2009, Metadata pre-standard extension for imagery and gridded data. ICS: 35.240.70

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

ISO-TC211 (2007). ISO 19139:2007, Geographic information-Metadata-XML schema implementation. ICS: 35.240.70

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

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I. and A. Manitsaris (2008). A Conceptual Framework for Metadata Quality Assessment, *Proceedings of International Conference on Dublin Core and Metadata Application, "Metadata for Semantic and Social Applications", September 22-26 2008, Berlin, Germany.*

Masó, J., Serral, I. and X. Pons (2011). GEOVIQUA: a FP7 scientific project to promote spatial data quality usability: metadata, search and visualization, *Proceedings 7th International Symposium on Spatial Data Quality, October 12-14 2011, Coimbra, Portugal*.

Moen, W. E., Stewart, E. L. and C. L. McClure (1998). Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS), *Proceeding ADL '98 Proceedings of the Advances in Digital Libraries Conference. April 22-24 1998, Santa Barbara, California, USA,* pp. 246. IEEE Computer Society.

Moré, J. and X. Pons (2011). Preliminary considerations about the assessment and visualisation of the quality on geometric corrections of satellite imagery depending on the number of ground control points, *Proceedings 7th International Symposium on Spatial Data Quality, October 12-14 2011, Coimbra, Portugal*.

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

Nebert D., Whiteside, A. and P. P. Vretanos (2007). OGC Catalogue Service Implementation Specification, Version 2.0.2, OGC 07-006r1, at: http://portal.opengeospatial.org/files/?artifact_id=20555 [Accessed 20 March 2010].

Ollier, G., Béroud, F. and K. Fontaine (2009). *Catalyzing Research and Development (R&D) Resources for GEOSS*, Committee in support of the GEO Task ST-09-01.

Pons, X. (2002). *MiraMon. Geographic Information System and Remote Sensing software*. Centre de Recerca Ecològica i Aplicacions Forestals, CREAF. Bellaterra. ISBN: 84-931323-5-7.

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

Stvilia, B., and I. Gasser (2008). Value-based metadata quality assessment, *Library & Information Science Research* 30 (1): 67–74.

Teillet P.M. and G. Chander (2010). Terrestrial reference standard sites for postlaunch sensor calibration. *Canadian Journal of Remote Sensing*, 36(5): 437-450.

Tolosana, R., Álvarez, J. A., Lacasta, J., Nogueras, J., Muro, P. R. and F. J. Zarazaga (2006). On The Problem Of Identifying The Quality Of Geographic Metadata. *Lecture Notes in Computer Science (LNCS)*. 4172: 232-243, ISSN 0302-9743.

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

van Zyl, T. L., Simonis I. and G. McFerren (2009). The Sensor Web: systems of sensor systems, *International Journal of Digital Earth*, 2(1): 16-30.

Xiao, N., Calder, C. A. and Marc P. Armstrong (2007). Assessing the effect of attribute uncertainty on the robustness of choropleth map classification, *International Journal of Geographical Information Science*, 21(2): 121-144.

Zabala, A. and X. Pons (2002). “Image Metadata: compiled proposal and implementation”, in Benes, T. (ed.) *Geoinformation for all*. Millpress, Rotterdam pp. 674–652. ISBN: 90-77017-71-2