Comparative Quality Assessment of Metadata. Two Regional SDI case studies.

P. Díaz[1], J. Masó[2] and J. Guimet[3]
[1]Department of Geography at the Autonomous University of Barcelona, Spain.
[2]Center for Ecological Research and Forestry Applications (CREAF), Spain
[3]Spatial Data Infrastructure of Catalonia, (IDEC), Spain

INSPIRE conference 2010. 23-25 June. Kracov

Summary
Since 2003 many tools for editing and cataloging metadata have been created. Metadata is formalized using the standard ISO 19115 that defines a set of possible elements. Only some of them must be filled in the metadata document to conform to all mandatory and condition requirements. However, metadata generation remains a costly and methodical process that depends on the good will of the producer. SDIs (Spatial Data Infrastructures) centralize these metadata and provide search tools based on metadata content. This paper examines the presence of conformance errors and good practices of metadata records, their nature and proportion for two regional SDI in Spain, very different in nature: Spatial Data Infrastructure of Catalonia (IDEC) which has collected approximately twenty seven thousand metadata records since 2002 and the Infrastructure Spatial Data of Castilla la Mancha (IDE-CLM) which was created in 2006 and has approximately one hundred metadata records. The paper describes the methodology followed to compile and analyze all the mandatory elements required by the ISO 19115 and some other considered relevant. The study focuses on the origin of these errors and suggests good practices to increate the quality and completeness of metadata records.

Keywords: geospatial data catalogue, metadata, SDI, ISO, INSPIRE, quality.

1. INTRODUCTION.

The distribution of the geographic data has experienced a renewal during the last decade. Some examples are the geographic data libraries, the geoportals and the diversification of the geographic data. The diversification in the data sources arise difficulties in finding data products and catalogues and services try to solve this discovery problem. Since 2002, the INSPIRE initiative has been promoting the geoservices, enabling the deployment of a European Spatial Data Infrastructure (ESDI) in which the data catalogues and services an essential technological part (INSPIRE 2007).

The first formal definition of the term SDI was held in the USA. The term SDI was defined as the technologies, policies and people necessaries to promote the exchange of geographic data at all levels of government, private and nonprofit sectors and academic (FGDC 1994).

The main functions of an SDI (Goodchild 2007) are:

- Data publication provided by produces.
- Data: access to heterogeneous resources.
- Data integration: gathering information and prevent duplication.

To accomplish these functions SDIs develop and maintain data catalogues (Nebert 2004). These catalogues register each geographic information dataset by means of a metadata record that describes it. Generally, the geographic information datasets are maintained and distributed by the providers and the catalogue of the SDI only centralizes metadata describing these datasets. Metadata is essential in the selection, transfer and manipulation process of the geospatial data. Depending on its use, metadata can be divided into three levels: discovery, selection and exploitation (Nebert 2004). Some metadata elements are more important for a specific level of use and consequently, the metadata records may be used for one level or another depending on the number of elements described.

This paper aims to detect and analyze errors in the metadata records, to determine the nature of these errors and their ratio of presence and also to make recommendations for avoiding them. The analysis of the quality of geographic datasets is beyond the scope of this study. In the current context of Geographical Information Systems (GIS), several studies evaluated the way metadata is produced (Batcheller 2007, Nebert 2004) but only a few of them focus on the quality of the metadata records (Edvardsen 2009, Tolosana 2006). The quality of the metadata has a direct impact on the success of data catalogue searches (Comprovets 2004).

2. CONTEXT OF THE IDEC AND IDE-CLM.

The concept of SDI was defined in 1994, the year of the National Spatial Data Infrastructure (NDSI) creation (FGDC 1994). The IDEC began to function in November 2002 with collecting metadata records in a catalogue and releasing the first version of its own the metadata generation program (MetaD) a few months after the project started. MetaD was released one year before the adoption of the standard 19115:2003, five years before the adoption of 19139:2007 and six years before the adoption of the INSPIRE directive; it was based on the draft standard ISO 19115 metadata information. In those days nobody knew which level of acceptance this standard would have by the international community. Moreover the only commercial products available for the creation of metadata were based on the FGDC standard CSDGSM. In 2004 MetaD introduced internationalization in collaboration with FGDC, added the capability to import FGDC metadata and adopted the final version of the ISO 19115 standard, and several other improvements and utilities to the final user. In 2007 the schemes that provided the ISO 19139 were finally approved. Since June 2009 the 3.0.5 version is been used and currently they are preparing a new version compatible with the INSPIRE metadata requirements. The wide variety of versions reflects the progressive development of ISO standards as well as the successive improvements and facilities of the application, giving us the idea of a dynamic evolution in the creation of metadata. IDEC was the first data infrastructure in Spain, to create and publish metadata in a standard catalogue (2003), long before other initiatives that emerged.

The high number of metadata record in the IDEC catalogue (Figure 1) is due to the constant influence and pressure that IDEC has been doing on the information providers. Is important to note that this high volume of metadata records come from a wide variety of providers that have created the metadata with different tools, such as different versions of MetaD (IDEC 2009); the GEMM (Pons 2001), also based on the work of Zabala (2001, 2002); or the CatMDEdit (TeIDE 2009). These metadata records describer varied datasets, some of which are non-cartographic documents. The high volume of metadata records is also due the lack of series model, so that each sheet is represented by a metadata record and products with high level of detail are cut in hundreds of sheets.

On the other hand, the IDE-CLM was founded in 2006 by the Junta de Comunidades de Castilla-La Mancha, in cooperation with the Spatial Data Infrastructure of Spain (IDEE) under the INSPIRE directive. Its main objective is the same that inspired the creation of the INSPIRE directive: facilitate the access to geographic information in the region to all citizens via the Internet (IDE-CLM 2010). Their catalogue has a few metadata records. They only store a metadata record for each geographic information series (and not for each sheet as in the case of the IDEC), so the volume of this metadata record should be smaller, but also the number of metadata providers is lower (Table 1).

Both IDEC and IDE-CLM are integrated in INSPIRE. In both cases providers are public agencies, different levels of public administration, university departments, research centers and private organization.

In this paper we describe an quality analysis and assessment done on the collection of metadata records present in both IDEC and IDE-CLM. IDEC metadata catalogue was analyzed in January 2010 and the IDE-CLM was analyzed in September 2009 (the current data catalogue is under review and has not been updated).

Organization	Providers	Total documents
IDEC	138	27386
IDEC excluding l'ICC	137	14616
IDE-CLM	9	98

Table 1. Organizations and number of SDI data analyzed.

2.1. Contextualization in the state level.

The Spanish SDI Working Group (GTIDEE) was created in 2002 as a mechanism to fulfill the needs for the European INSPIRE initiative. Public, and private sector from local, regional and national levels representatives meet periodically. A result of this initiative is the IDEE project. Currently, according to the IDEE, there is a total of 43 SDI in Spain: 10 of them are national, 19 local and 14 regional. All of these have their own metadata catalogue. Table 2 summarizes the regional SDIs, the number of metadata records and the metadata standard provided (IDEE 2010). Both IDEC and IDE-CLM provide access to their catalogues through the CSW standard, for downloading massive metadata records in XML format under the ISO 19139 scheme, which makes possible to known in full detail about the content of metadata records.

Currently, the metadata working group of the IDEE has completed the work of several years of adaptation to the INSPIRE initiative with a ISO19115 profile that is called "*Núcleo Español de Metadatos*" (Metadata Spanish Core) (NEM v 1.1) that is only waiting the final approval in the GTIDEE.

As shown in the table 2, most of the Spanish regional SDI are adapted to ISO 19115. This has been carried out in recent months; an example of this is the recent adaptation to the ISO 19115 metadata of the Galicia SDI (IDEG) and the IDEE which had their metadata in Dublin Core standard. The best examples of adaptation to an international standard and of the growth activity of national SDIs are also reflected in the current update of Castilla y León (IdeCyL) metadata records, or the current update of 24 000 metadata records more to the Andalucia SDI (IDEAndalucia) catalogue.

Regional SDI	Type of catalogue	Metadata records	Open access	Metadata standard	catalogue protocol
Andalucía	data	23344	yes	ISO 19115	OGC-CSW
Aragón	data and services	17408	yes	ISO 19115 - FGDC	no
Islas Canarias	services	---	---	---	---
Castilla-La Mancha	data and services	98	temporarily unavailable	ISO 19115	OGC-CSW
Castilla y León	data and services	≈200	being updated.	ISO 19115	---
Cataluña	data and services	27386	yes	ISO 19115	OGC-CSW
Comunidad Foral de Navarra	data and services	275	yes	ISO 19115	OGC-CSW
Comunidad Valenciana	data and services	*116*	yes	ISO 19115	OGC-CSW
Extremadura	data and services	1425	yes	ISO 19115	OGC-CSW
Galicia	data and services	43	yes	ISO 19115	OGC-CSW
Illes Balears	data and services	7388	yes	ISO 19115	OGC-CSW
La Rioja	data and services	550	yes	ISO 19116	OGC-CSW

Table 2. Number of metadata records available in the regional SDI with CSW server.

2. 2. Contextualization in the European arena.

In April 2007 was signed the INSPIRE (Infrastructure for Spatial Information in Europe) directive agreement, with the aim to create a Spatial Data Infrastructure in Europe. INSPIRE includes legal rules, standards and protocols to facilitate the integration of geographic information relevant to the European level. All member states had an initial period of two years to adapt to the directive (INSPIRE 2007).

The INSPIRE directive regarding metadata (INSPIRE 2008) provides a total of 25 mandatory metadata elements, whereas the standard ISO 19115 currently requires 9 (Table 3). This higher requirement of the European directive indicates a greater effort of the SDIs currently regulated under the ISO 19115, to adapt to the European directive, an action that, as mentioned above, IDEC has already begun.

MANDATORY ELEMENTS IN INSPIRE	ELEMENTS IN ISO 19115
Title Abstract Resource language Metadata date Topic category	Mandatory
Responsabile party information Responsible party role	Mandatory at least organisation name, individual name or rol
Date of publication Date of last revision Date of creation	Mandatory at least one of the three
Geographic bounding box Metadata language	Conditional
Keyword value Originating controlled vocabulary Temporal extent Lineage Spatial resolution Specification (rules adopted) Degree (of conformity of the resource) Conditions applying to access and use Limitations on public access Metadata point of contact Resource locator Unique resource identifier Resource type	Optional

Table 3. Degree of obligation of the metadata elements in ISO 19115 and INSPIRE.

Recently, the European Commission Joint Research Centre has published an analysis of a sample of eleven regional European SDIs (Craglia 2009). The regional SDIs included in that study are the SDI of Navarra (IDENA) and the IDEC. Some of the main conclusions are that the eleven SDI use similar technology (OGC-based services and ISO compliant metadata), and most regional SDI need to maintain partnership with other local agencies to develop their activity, due to the presence of multiple actors in the territory.

According to the study, there are three starting points for a the SDI; first, a preexisting regional/national mapping agency (NMA) that creates its own geo-portal, like the Navarra's case; secondly, the agreement of local and national map agencies, and thirdly, the result of governmental policy for the creation and maintenance of geographic information. In this sense, IDEC is one of five regional SDI that has developed a legal framework; a fact that connects with the idea exposed by Craglia about the importance of the regional composition of states like Spain, Belgium, Italy and Germany.

At the time of Craglia's study only IDEC had completed an impact assessment, that shows how an SDI can quickly recover the investment, even without taking into account the cost of data production.

2.3. Contextualization in the international arena.

In 2004, Comprovets had examined the global situation of major national SDIs (Comprovets 2004). The author exposes that satisfactory results in data search catalogue depends largely on the quality of metadata records that will impact the search algorithms capacity. The success of the search is directly related to the accessibility of data, and therefore the usability of the metadata.

Comprovets classifies the people involved in a geo-portal in three groups: providers, administrators and end users. It estimates an average of about 50 data providers of geo-portal. According to our informationthe IDEC had 79 providers, in summer 2008, and138 in April 2010; the IDE-CLM had about ten providers in September 2009 (Table 1).

The same study detects a decrease in the number of providers that is directly related to a reduction of data publication in recent years. It happens in our study as well as reflected in the temporal evolution of the datasets (created, published and revised), in the last 10 years (Figure 1). The creation of data presents a peak in 2003, the revision in 2005 and the publication in 2006, in the three cases it's followed by a decline. We do not find a reduction of the data providers in the catalogue of the IDEC but in the number of published metadata records created and revised.

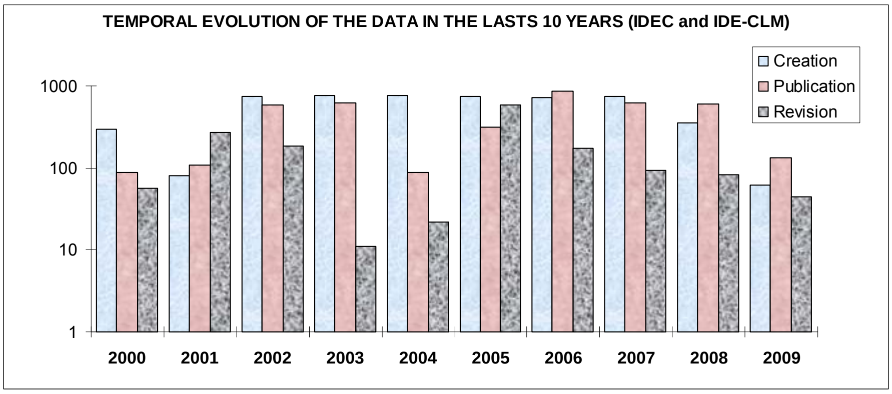


Figure 1. Temporal evolution of the creation, publication and revision of geographical data of the IDEC and IDE-CLM according to datasets dates contained in the metadata records.

According to Comprovets study, the number of final users of the SDI has been stabilizing, and the average of visitors is around 5 000, although the median is around 1 000, due to wide differences between geoportals. Furthermore, it shows the volume of data 13 times greater in U.S. than in Europe, and the average of datasets accessible in the Geoportals is established in 3 500.

The information extracted from the IDEC geo-portal shows that there were 96 412 visits in 2008, 4 708 of them to the catalogue. The IDEC data catalogue had 27 386 metadata records in April 2010 and the IDE-CLM data catalogue had 98 in September 2009 (Table 1). The actual number of metadata records of the IDEC catalogue is tripled due to metadata is translated in three languages: Catalan, Spanish and English. The set of documents stored in the IDEC is not only composed by cartographic data, but it also includes geographic information available in text format, books, magazines, studies or tables, covering different thematic categories. The main contributors to the IDEC metadata catalogue are the Cartographic Institute of Catalonia (ICC) with 12 770 records, and 3 613 coming from IDEUnivers project. The agency that provides more metadata records to the IDE-CLM is the Regional Development Institute (Instituto de Desarrollo Regional; IDR), with 53 records, followed by the Statistical Institute of Castilla La Mancha, with 21.

Finally, the SDI are containers of geographic information generated by multiple providers under different operation criteria and with different objectives. The IDEC is the Spanish regional SDI with the largest number of geographic data, thematic variety and number of providers.

3. METHODOLOGY.

Thanks to the standardization the interoperability is possible and, in our case, we can interact with metadata catalogues and compare them with the same methodology. Generally speaking interoperability contributes to the unification of efforts, aimed at improving quality, avoiding duplication of data and efforts.

The methodology of this work uses three standards: ISO 19115 (ISO/TC211 Moellering 2003 and 2006) that defines the content of each item in a metadata record, ISO 19139 (ISO / TS 19 139 2007) that defines how to encode this metadata in XML and OGC-CSW (Catalogue Service for Web) (OGC-CSW 2007) that provides a unified interface to publishing and search metadata records. Using the the IDEC (http://delta.icc.es/indicio/csw) and the IDE-CLM (http://161.67.130.140:8080/geonetwork/srv/en/csw) server addresses is possible to enumerate the whole set of record identifiers (with the "GetRecords" operation) and then get metadata records one by one encoded with the ISO 19139 standard (with the "GetRecordByID" operation).

The ISO 19115 standard establishes three element categories:

- Mandatory (that must be populated).
- Conditional (that become mandatory elements in the absence of other elements).
- Optional (that are not truly required, but are useful in many cases)

Once we had all the metadata XML files, we extracted all the ISO mandatory elements (ISO/TC211 2003 and Moellering 2006), and also other optional elements we consider important for the better understanding of the geographic datasets, including the spatial resolution and the keywords, considered mandatory in the INSPIRE directive (INSPIRE 2007) metadata profile. All the information was stored in database tables, containing the mandatory and optional elements in columns (fields) and the XML files, identified by their unique identifier, in rows (records). These database tables are easier to analyze than the original XML files, allowing us to work with a large amount of elements and records, systematizing the search process and error detection. This methodology was applied to the IDEC and IDE-CLM catalogues and the results are shown and compared.

The IDEC database has 14 616 rows and 35 columns. We analyzed all metadata records except those coming form the Cartographic Institute of Catalonia (ICC), because we assumed a high quality and rigorous production that we have checked in a different study (Figure 2). In the case of the IDE-CLM we have analyzed all available metadata records, creating a database of 98 rows and 35 columns (Figure 3).

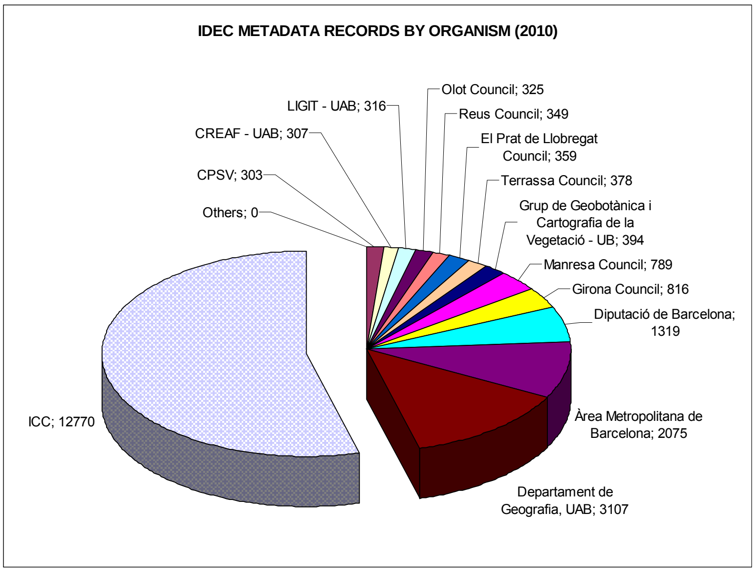


Figure 2. IDEC metadata records per organism with more than 300 records in the catalogue. The ICC records have been excluded.

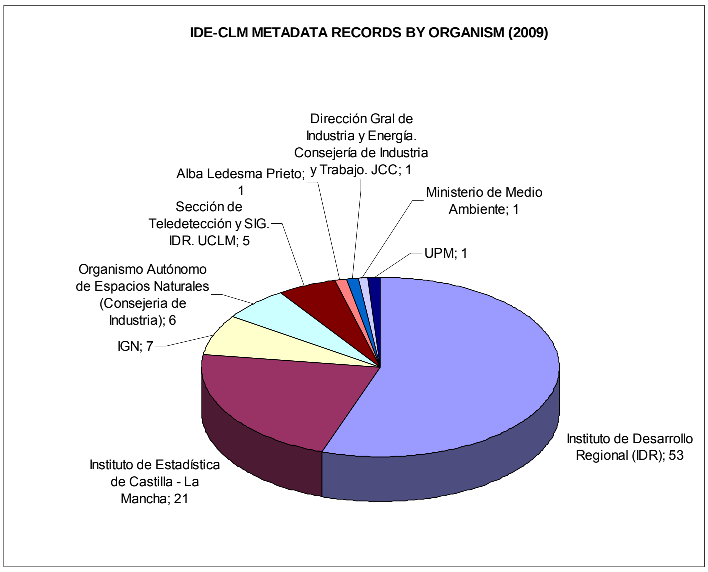


Figure 3. IDE-CLM metadata records per organism.

4. QUALITY ANALYSIS OF THE METADATA RECORDS.

This analysis is divided into three parts. The first part analyzes the lack of compliance with the requirements of the ISO 19115, mainly characterized by mandatory and conditional elements missing in metadata records which can be considered an unintentional mistake or a misinterpretation of the standard requirements. It possible to determine the degree of completeness of metadata records analyzing the errors of omission, in other words, the lack of required information, with a ratio of each type of error.

The second part analyzes other aspects that can be considered lack of good practices; these practices don't strictly contravene the requirements of ISO 19115, but make difficult data discovery or metadata records understanding, and consequently, diminish significantly the possibility of successful searches of SDI data catalogues. This document is complemented by a database containing the singular errors and lack of good practices, as well as some suggested solutions for each metadata record, with the purpose of correcting or mitigating these errors.

The analysis also takes into account which organism committed each error, with the aim of determining the source of errors and lack of good practices. Finally, in the third part suggest possible reasons for these errors.

4.1. Errors in the metadata records in autonomic catalogues.

4.1.1. Errors in the mandatory elements of the ISO 19115 standard.
The mandatory elements are:

- The title,
- The abstract,
- The metadata date,
- At least one of the three dates relating to the creation, publication or revision of a dataset,
- The topic category,
- The contact information of the metadata creator,
- The data language.

The title, the abstract and the data language are also considered in the part 4.2, concerning lack of good practices, in conjunction with other aspects.

4.1.1.a) IDEC catalogue.
The metadata date is mandatory, and also one of the three dates relating to the dataset creation, publication or revision. There are 353 metadata records (2.42%) without metadata date; as well as none of the three dataset dates were found in 1 779 metadata records (12.17%). The dataset revision date is previous to the creation date in three records (0.02%), as well as creation date of the dataset is after the creation of metadata in 491 cases (3.36%) both been obviously incoherent.

There are 19 topic categories established by the ISO 19115 (Figure 4) that help us to characterize the type of data. Due to the multiple cardinality of this element all topic categories in metadata records have been considered (Figure 4). More than a half of the metadata records are "ImageryBaseMapsEarthCover": 8 385 (57.37%). "Environment" (13.95%), "structures" (11.56%), "location" (9.79), "boundaries" (9.15) and "PlanningCadastre" (8.93%) are next more present topic categories, but with a ratio of presence quite below the first one. The absence of the topic category is a mistake, which happens in 499 cases (3.41%).

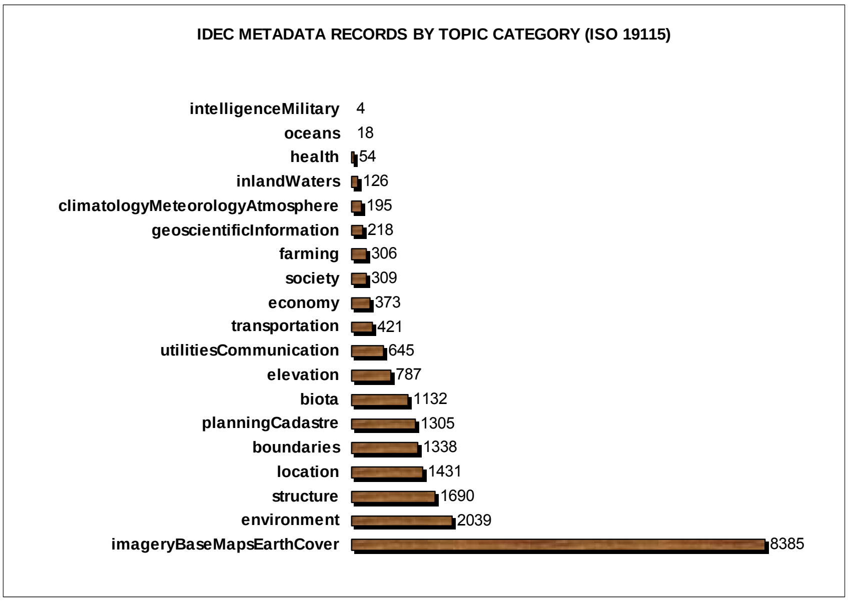


Figure 4. IDEC metadata records by topic category.

The topic categories are defined in the ISO 19139 as a non extendable list (enumeration) of 19 codes written in English. 1 418 topic categories in metadata records are not in the enumeration (9.70%), generally equivalent descriptions in the metadata language are used, such as "Bases de mapes" or "Imatges de la cobertura terrestre" instead of "ImageryBaseMapsEarthCover".

In ISO 19115 standard the metadata creator contact has several elements but we consider here: individual name, organization name and title. Each metadata document should have at least one of them. In our metadata collection, we found that 39 metadata documents (0.27%) does not have any of the three elements.

4.1.1.b) IDE-CLM catalogue.
Metadata publication date is described in all the metadata records; but there are 36 metadata records (36.7%) without any of the three datasets date types.

Topic categories are described in all metadata records except in 3 (3.06%). "PlanningCadastre" topic category is the most present, with 53 records (54.08%). Six topic categories are not present in any of the metadata records in the IDE-CLM (Figure 5).

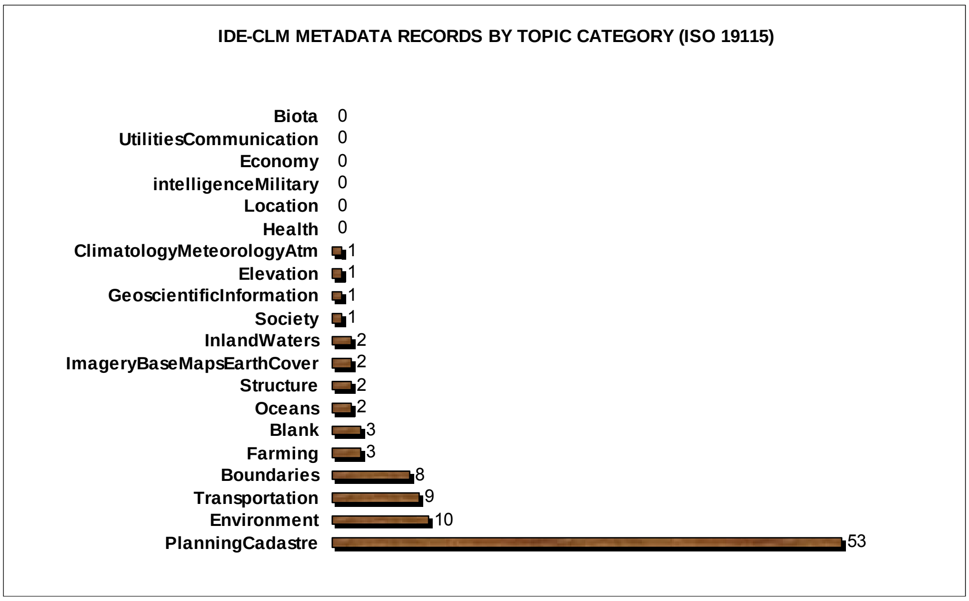


Figure 5. IDE-CLM metadata records by topic category.

Two metadata records lack do not follow the conditional rule about the metadata creator contact minimum information (2.04%).

4.1.2. Errors in the conditional elements of the ISO 19115 standard.
The conditional elements are:

- The extension, that should be introduced in geographical coordinates (latitude/longitude), and if it's not indicated, description of the geographical extension should provided.
- The language of the metadata, if it is not described in the document encoding.

4.1.2.a) IDEC catalogue.
Geographical extent should be defined by four geographical coordinates or by a description, there are 33 metadata records that lack both (0.23%). The ISO 19115 standard requires that extension must be expressed in angles (latitude/longitude). However, we have 27 metadata records (0.18%) with numbers clearly out of range, that seems to be in the data reference system units: UTM31N ED50 / m. One metadata record has the minimum coordinate greater than the maximum one (0.01%).

ISO 19115 standard does not provide any way to indicate the reference system of the geographical extent. Nevertheless, there is some consensus in using longitude/latitude in the WGS84 datum, which is internationally valid. In our opinion this information could be include in an alphanumeric description, for example in the description of the extension (Ex_GepgraphicDescription).

33 of the Catalan metadata records have lack of metadata language element (0.23%). we have detected 51 metadata records written in Spanish even if the element says Catalan. It could be due to a confusion between the language of the metadata and the language of the data.

4.1.2.b) IDE-CLM catalogue .
The geographical extent is not indicated in 29 metadata records (29.21%) and 64 (65.30%) metadata records describe the extent coordinates with values clearly out of range that seems to be UTM coordinates matching

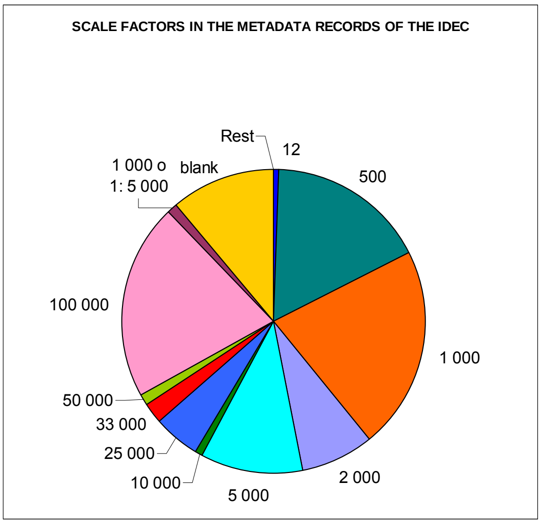
the area. According to the standard ISO 19115 it should be in angles (longitude/latitude). A single document presents inconsistencies between the minimum and maximum coordinates (0.01%).

The metadata language is not specified in 21 metadata records (21.42%), which also does not specify the data language. The language defined in three metadata records is English (3.06%), which as proved an error due to metadata are written in Spanish.

4.1.3. Errors in optional elements of the ISO 19115 standard.
Optional elements that have been covered are:

- scale,
- coordinate reference system (CRS),
- geographic data models,
- topology.

4.1.3) IDEC catalogue.
There are 64 different scale factors, 6 of them are highly improbable in a map, or very small ("1", "12", "13", etc) or mixed (1000 or 1:5000), or unlikely ("302"), (Figure 6). These scale factors are inconsistent in 341 metadata records (2.33%). Although it is not a mandatory element in the ISO 19115 standard, we consider it an important element for geographic data processing and search.



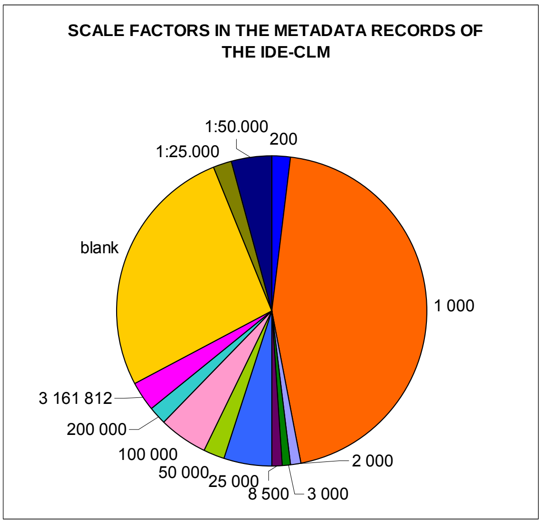


Figure 6. Scale factors commonly used in the IDEC metadata catalogue ("Rest" includes those below 100 entries).
Figure 7. Scale factors used in the IDE-CLM metadata records.

There are 6 different CRS in the IDEC metadata. We found that 20 coordinates relative to the reference system EPSG: 32 633 - WGS84/UTM zone 33N are not consistent with the area indicated in the extent. IDEC metadata records without CRS are 731 (5.00%).

4.1.3.b) IDE-CLM catalogue .
There are 72 metadata records containing a scale factor (73.46%). Unusual scales denominators and fraction formats not according to ISO 19115 ("3161812", "1:25000" "1:50000") are presents in 9 records (0.09%) (Figure 7).

Finally, all the metadata records have metadata describing the CRS.

4.2. Lack of Good Practices.
We have exposed the conformance with ISO 19115 standard errors that we had detected in the metadata records classifying them in three levels: mandatory elements, conditional elements and optional elements. This section describes practices that are not considered errors under the standard ISO 19115 but can reduce interoperability with other systems or the ability to search in the data catalogues.

The elements that we consider important for understanding geographic data are:

- title,
- abstract,
- geographic data models,
- topology,
- keywords,
- dataset dates,
- dataset language.

4.2.a) IDEC catalogue.
The title and abstract are two essential elements in a metadata document, allowing users to understand the data and to easily identify it on a search. Some practices, that do not help to understand the geographic data were studied, complements the previous aspects about the lack of these elements. There are 34 metadata records that do not contain title (0.23%), and 4 070 metadata records (27.85%) that could improve the content of it, to increase the users understanding (Table 4).

TITLES	Total	Percentage
Slight descriptive alphanumeric sequences	4070	27.85%
Titles too long (more than 110 characters)	153	1.05%
Metadata without title	34	0.23%
Amount	***4257***	***29.13%***

Table 4. Aspects to improve the understanding of titles (IDEC).

Although the standard does not mention length limitations, we established the criterion of lack of conciseness in a threshold of 110 characters, considering that a title should be concise and direct. Under this criterion there are 153 titles (1.05%) too long, so we recommend moving some parts of it to the abstract. When the title is too short, we recommend adding specific information about the product type, date and scale to complete it.

The abstract must be completed but the IDEC catalogue has populated them in 99.76%. Additionally, 0.5% of these cases has the abstract identical to the title, which does not provide additional information and is not in line with the INSPIRE different definitions of these two elements. According to the INSPIRE metadata title is “a characteristic, and often unique, name by which the resource is known”, and the abstract “is a brief narrative summary of the content of the resource”.

Vector is the most abundant geographic data model listed in the metadata records distributed by IDEC (in 8 950, 61.23%). Grid model has been listed in 4 375 metadata records (29.93%), text table in 50 metadata records (0.34%) and TIN in 5 (0.03%) (Figure 8).

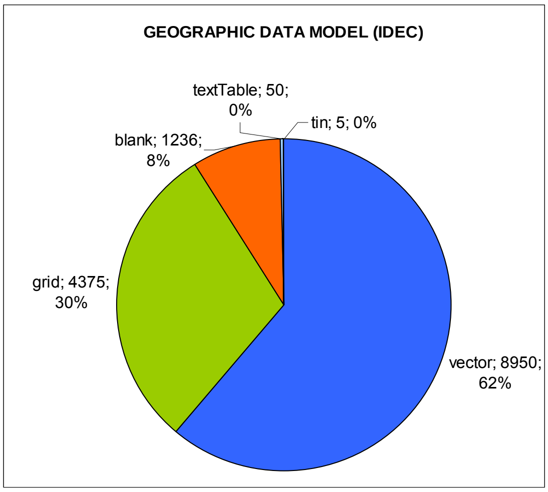


Figure 8. Geographic data model (IDEC).

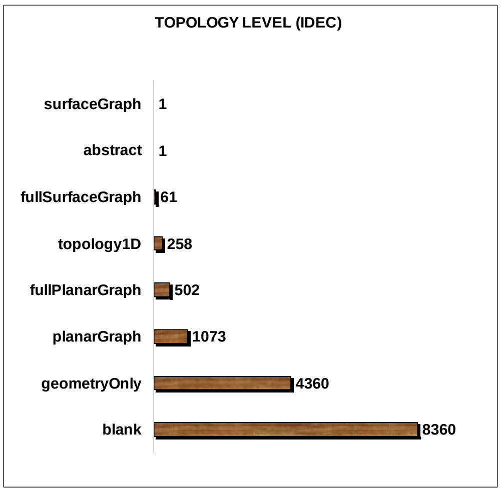
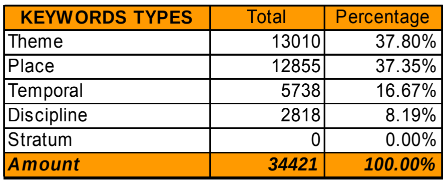


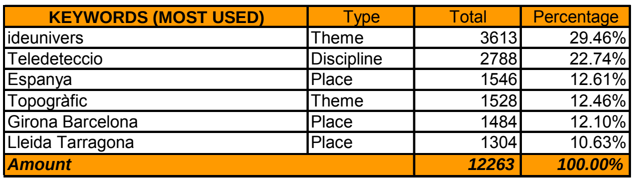
Figure 9. Topology level in the metadata records (IDEC).

Subject to available information, topology level should be defined. More than a half of the metadata records distributed by the IDEC do not specify any topology level, 8 360 (57.20%); 1 896 metadata records indicate some topology level (12.97%), while 4 360 metadata records (29.83 %) specify that non-present topology (Figure 9).

Due to the abundance and variety of keywords, we collected only the first keyword of each type. The ISO 19115 standard identifies five types of them: discipline, place, stratum, temporal and theme.



KEYWORDS TYPES	Total	Percentage
Theme	13010	37.80%
Place	12855	37.35%
Temporal	5738	16.67%
Discipline	2818	8.19%
Stratum	0	0.00%
Amount	***34421***	***100.00%***

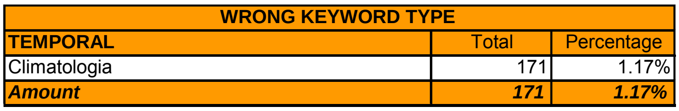


KEYWORDS (MOST USED)	Type	Total	Percentage
ideunivers	Theme	3613	29.46%
Teledeteccio	Discipline	2788	22.74%
Espanya	Place	1546	12.61%
Topogràfic	Theme	1528	12.46%
Girona Barcelona	Place	1484	12.10%
Lleida Tarragona	Place	1304	10.63%
Amount		***12263***	***100.00%***

Table 5. Metadata records by keyword type (IDEC). **Table 6.** Most used keywords (IDEC).

The results indicate that keywords are very frequently used for the 4 types: place, theme, discipline and temporal (Tables 5). Table 6 shows the six most used keywords; some of them (such as "Ideunivers") represent IDEC specific terms used as internal method of data classification. The importance of keywords to find a dataset in searches of the IDEC should be emphasized, because it currently are one of the search criteria in the IDEC catalogue.

We detected an error in "temporal" keywords that is defined in the ISO 19115 standard as "(it) identifies a time period related to the dataset". There are 171 metadata records categorized as temporary keywords that are not a period of time (Table 7).



WRONG KEYWORD TYPE		
TEMPORAL	Total	Percentage
Climatologia	171	1.17%
Amount	***171***	***1.17%***

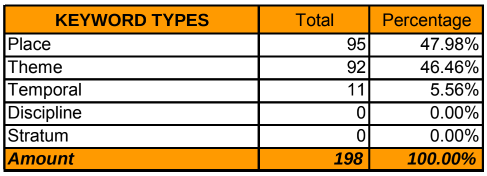
Table 7. Number and percentage of metadata records keyword wrong type.

The IDEC uses data creation "1/1/1900" as a no-data flag. (It was found in 1 385 metadata records (9.48%)). We do not recommend this practice for two reasons: first, it's a non standardized practice that could cause problems of interpretation and validation if the metadata is exported to other catalogues; second, as a required metadata element to persuade the user to indicate an approximate date (only year, for example) or even to include a range of uncertainty in the date are better practices.

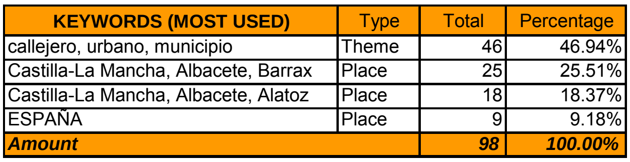
The language of the data is mandatory and is not indicated in 357 metadata records, a 2.44% of cases. This situation could be justified in cases where data do not contain textual information (isolines, altimetry, etc.).. However, this case could tolerate future problems as other catalogues may include this element as a requirement for acceptance of a metadata document.

4.2.b) IDE-CLM catalogue .
Title and abstract are present in all metadata records. There is no case with numerical descriptions or title and abstract redundancy. There is not any title longer than 110 characters.

Keywords are present in three categories: place, theme and temporal. Keywords containing more than one word were detected, which can lead to confusion (Tables 8 and 9).



KEYWORD TYPES	Total	Percentage
Place	95	47.98%
Theme	92	46.46%
Temporal	11	5.56%
Discipline	0	0.00%
Stratum	0	0.00%
Amount	***198***	***100.00%***



KEYWORDS (MOST USED)	Type	Total	Percentage
callejero, urbano, municipio	Theme	46	46.94%
Castilla-La Mancha, Albacete, Barrax	Place	25	25.51%
Castilla-La Mancha, Albacete, Alatoz	Place	18	18.37%
ESPAÑA	Place	9	9.18%
Amount		***98***	***100.00%***

Table 8. Metadata records by keyword type (IDE-CLM). **Table 9**. Most used keywords (IDE-CLM).

Finally, the data language is blank in 25 metadata records (25.3%), and topology never indicated. Figure 10 represents the geographic data model of the metadata records published in the IDE-CLM.

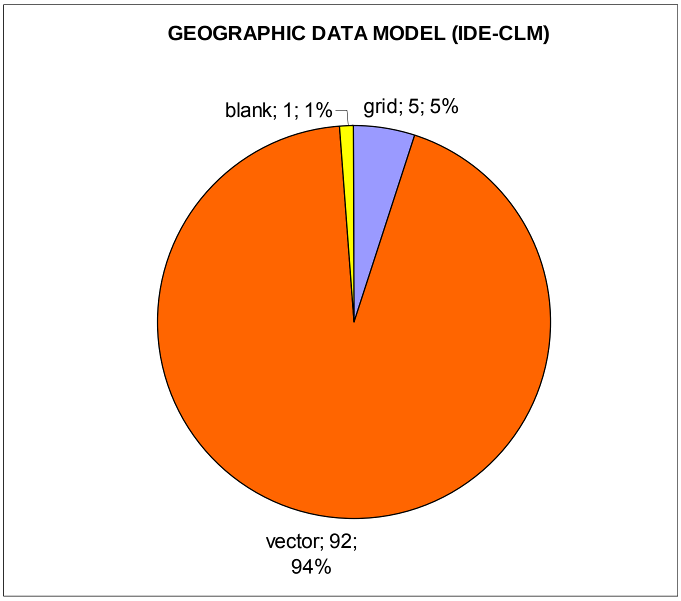


Figure 10. Geographic data model (IDE-CLM).

4.3. Reasons for the presence of errors in the metadata.

This paper does not only study the quality and errors of metadata, but also explores the diverse reasons of these errors and makes recommendations to improve this quality. Some errors or deficiencies are due to a lack of information (such as the missing creation date of the dataset or the lineage); the difficulty applying concepts to some types of information (the scale information in tabular data with x,y positions). In general providers can have different methods for creating and publishing metadata records in the SDI: online forms, automated data collection (harvesting), XML document submission or direct transmission from a desktop GIS (Goodchild 2007). GeMM (Pons 2001) and CatMDEdit (CatMDEdit, 2006) tools allow the automatic extraction of metadata from de data, and MetaD enables direct metadata document publication in the IDEC catalogue. These three tools are the most used in Spain. However, these methods are not exempt of errors. Although validation functions control the mandatory metadata, do not prevent introducing erroneous or incomplete information. However, in the IDEC case, some frequent errors in mandatory elements have been associated with some user behaviors that escapes MetaD application controls.

Several studies evidence the fact that automatic extraction of implicit metadata avoids the generation of errors saving user time. A study of Mansó (2009) shows a high ratio of error in the manual compilation of implicit metadata; being the number of errors inversely proportional to metadata knowledge of the operator.

Finally we found that some errors currently on IDEC metadata catalogue can not be generated with the program MetaD, CatMDEdit or GeMM, so it seems that providers have generated metadata records with other metadata programs that do not have the validation functions following ISO 19115 standard requirements.

5. RECOMMENDATIONS.

Throughout this analysis we have reviled some metadata record fail in ISO requirements tests (paragraphs 4). The most immediate recommendation is to correct them and to try to determine their reasons, in order to avoid future repetitions. SDIs should not publish documents that lack some mandatory element.

We acknowledge that SDIs are not metadata producers, so therefore they are not responsible for the quality of the metadata published in the catalogue. However, when SDI detects errors like this paper reviles, we recommended to communicate with producers, responsible for the creation of metadata records, helping in the process of solve the present and future errors.

We recommend reaching a consensus among providers and SDIs about controlled vocabularies containing e.g. the names of the organism providers or other element content that they currently differ but a unification effort can be done.

We also recommend establishing common rules for generic creation of metadata titles to make them easy to read and to avoid repetition with the abstract, as described by INSPIRE (discussed in paragraphs 4.1, 4.3 and 5.4). For titles of series sheets a template is recommended like the one already applied in accordance with the ICC, in the MMZ format distribution (MiraMon):

{Titles series} {indexLeaf } {alfanumeric Leaf Name}

In the IDEC case, two different metadata schemes has been detected, which differs in the encoding of a particular element. A single schema should be accepted to avoid confusion and to improve interoperability. Practices that could mislead the end user should be avoided like the use of "1-1-1900" as a NODATA value, (discussed in paragraph 4.2).

In paragraph 4.4, referring the lack of some good practices, the desirability of using a thesaurus was indicated for the keywords selection, due to the standardization of descriptive criteria facilitates accessibility and transmissibility of geographic data, particularly important for the IDEC because keywords are one of the main strategies in catalogue searches. The indication of a thesaurus is mandatory in INSPIRE, while the ISO 19115 standard gives the option; since IDEC will be part of INSPIRE it is highly recommended to use keywords that comes from a thesaurus, also indicating a link the thesaurus itself. In fact, the latest versions of GEMM and MetaD incorporate thesaurus of keywords (12 161 entries in the case of GeMM, 8 358 of them coming from the Catalogue of the Library of Catalonia, CBUC, that has hierarchical structure; 3 803 entries, without hierarchy, in the case of MetaD). Using a link to a dictionary is also recommended by the mechanism "gmx: Anchor xlink: href":

```
<Keyword>
xlink:href="http://idec.icc.cat/dictionaries/KeywordDictionary.xml#Catalunya"/> <gmx:Anchor
</ Keyword>
```

Figure 1 shows the three dates relating to the data from 2005 to 2009 (creation, publication and revision treated in paragraphs 4.1, 4.4 and 5.1). The publication date is the most used one although, probably, the user prefers to know the creation date of the dataset. In addition, the temporal extent is necessary for many geographic information processes. We recommend providing at least the creation date.

We observed the IDEC catalogue has been renovated and has already included some suggestions that we recommended in a previous metadata quality assessment (Díaz 2009). Going beyond, we also recommend including these recommendations in the IDEC metadata profile, as well as disseminate the detected problems especially among organisms that are making more errors than others in particular issues previously explained.

6. CONCLUSIONS.

Carry out a systematic revision of the metadata records collected in SDI metadata catalogues is possible in order to detect errors, weaknesses, lack of good practices. Furthermore determine the responsible parties for a specific problem and analyze the time trend of some problems can be achieved.

The number of metadata records collected in the SDIs corresponds directly to the number of providers. Information providers produce metadata that is collected in SDI catalogues not being directly responsible for the quality of metadata records they receive. However, periodic metadata quality checks can be made to detect errors or lack of good practices.

Errors in metadata records are due to three general reasons: provider's lack of information, the standard capacity to describe the information required and the tools available to generate the metadata records. These aspects directly affect the quality of metadata and therefore the search results on data catalogues.

As discussed, the INSPIRE is more demanding respect to the completeness of the metadata. ISO 19115 currently requires completing nine elements, while the INSPIRE requires 25 (Table 3). If the SDIs currently have difficulty validating metadata under minimum requirements of the ISO19115 standard, the integration with the European directive will be very difficult.

Metadata analysis of metadata catalogued in the two regional SDI studied manifests the presence of different kinds of errors in these metadata records (Table 10). Approximately 4.33% of the metadata records of the IDEC lack of mandatory element of the ISO 19115 standard, in contrast to 13.47% of the IDE-CLM metadata records. The average error for all metadata records is around 3.84% in the IDEC case and 11.73% in the IDE-CLM case.

Both metadata catalogues show frequent errors or lacks in best practices in the dataset dates and the dataset languages, so the studied SDI must insist on corrections to the producer third parties. In addition, the IDE-CLM need to find out the origin of that high error level in the bounding box coordinates of the datasets published.

COMMON ERRORS FOUND IN THE SDI	IDEC	IDE-CLM
Metadata date in blank	2.42%	0.00%
Data dates in blank (the three)	12.17%	36.73%
Creation date later than metadata date	3.36%	0.00%
Creation date "1900-01-01"	9.48%	0.00%
Topic category not in the enumeration	9.70%	0.00%
Topic category in blank	3.41%	3.06%
Contact information in blank	0.27%	2.04%
Geographic extent not in angles (lat/long)	0.18%	60.20%
Minimum coordinate greater than the maximum	0.01%	1.02%
Data language in blank	2.44%	25.51%
Incorrect metadata language	0.35%	3.06%
Inconsistent scale factors	2.33%	9.18%
Average error	***3.84%***	***11.73%***

Table 10. Abstract of errors found in the IDEC and the IDE-CLM.

Despite the enormous difference in the volume of data, the number of third party providers and the number of years in operation, it has been observed that the error committed by IDEC and IDE-CLM are quite similar, confirming the general need to solve these errors and increase metadata quality.

Much of the metadata records show errors that cannot be involuntarily made from common metadata tools (because all of them have filtering mechanisms and controls that reports and forces the user to correct them), so it's obvious that other tools or methodologies are being used that don't have an strict validation. It's necessary to have a set of consistent validation rules for all the metadata creators that could be applied both to the editing tools as well as metadata catalogues.

On the other hand, description in the optional elements that we consider quite relevant, has even a lower quality. In order to facilitate the task of creating metadata records, some tools for automatic metadata extraction should be considered (Manso, 2004 and 2009) facilitating and speeding a homogenized creation of metadata records. Therefore the metadata producer could focus on describing additional elements, improving the understanding of the data, which is very important for items that are currently optional in the ISO standard but mandatory in INSPIRE (Table 3).

There is a need for implementing more quality control procedures, beyond the presence or absence of data, to review the consistency of document elements and even make a quality report (i.e. quality controls could give a score to a metadata record, taking into account different criteria, such as the abundance of the elements described in a hierarchical rank). These quality controls could even suggest some improvements in the quality of metadata records, such as developing concise titles, non-redundant abstracts or consolidation of data dictionaries for the description of keywords. We consider that forthcoming metadata profiles could include a set of testable criteria (as a test recipe) which could be used by metadata tool developers and catalogue managers to provide a set of homogeneous and practical criteria in the heterogeneous SDI environment.

Quality on metadata records is a compromise between agility for third party providers who create metadata and the needs of the end-user who wants as much detailed information as possible. In this sense, IDEC has done an excellent job of motivation, collection and integration of the various actors involved. This has resulted in a critical mass of information that makes the current infrastructure useful. Since last year, IDEC has been refocusing from volume production to quality improvement. This paper shows that this quality refocusing process is needed and many errors can be detected with data analysis procedures and therefore corrected or prevented.

Acknowledgments.

The authors would like to express their gratitude to Victor Pascual Ayats and Jordi Guimet (IDEC) for the information provided during this study, as well as those regional SDIs that had contributed, and those people who had shared their expertise and suggestions.

We also acknowledge the finance that ICC has provided to expand this study.

REFERENCES.

COMPVOETS, J., BREGT, A., RAJABIFARD, A., WILLIAMSON, I. (2004). "Assesing the worldwide development of nacional spatial data clearinghouses". International Journal of Geographical Information Science, 18:7, 665-689.

CRAGLIA, M., CAMPAGNA, M. (2009). "Advanced Regional Spatial Data Infrastructures in Europe". European Commission, Joint Research Centre Institute for Environment and Sustainability. Italy. ISBN: 978-92-79-11281-2.

DIAZ, P., MASÓ, J., ZABALA, A. (2009). "Análisis comparativo de los metadatos distribuidos por la IDEC y la IDECLM como ejemplos de SDI autonómicas". Communication to JIDEE 2009. Murcia. ISBN: 978-84-87138-56-0. 12 p.

DIAZ, P. (2009) "Análisis comparativo de los metadatos distribuidos por la IDEC". Master project in Remote Sensing and GIS, 10th edition Autonomous University of Barcelona and CREAF. Bellaterra. 15 p.

FGDC (1994). Federal Geographic Data Committee. Executive Order 12906: Coordinating geographic data acquisition and access: The National Spatial Data Infrastructure. April 11th, 1994.

GOODCHILD, M.F., FU, P., RICH, P. (2007). "Sharing geograpic information: An assesment of the Geospatial One-Stop". Annals of the Assosiation of American Geographers, Vol. 97:2, 250-266.

IDEC (2009). "Manual de l'usuari de l'aplicació MetaD per a la creació, edició i exportació de metadades (ISO 19115 - ISO 19139). Versió: 3.0.5". Available in Internet: http://www.geoportal-idec.cat/geoportal/cas/meta-d/Manual_MetaD_3.0.5.pdf, last access May 21st 2010.

IDEE (2010). "Infraestructura de Datos Espaciales de España". Web: http://www.idee.es, last access June 07th 2010.

IDE-CLM (2010). "Infraestructura de Datos Espaciales de Castilla la Mancha". Web: http://SDI.jccm.es/, last access May 11st 2010.

INSPIRE (2007). "Infrastructure for Spatial Information in the European Community". Official Journal of the European Union. Directive 2007/2/CE of the European Parliament and of the Council of 14 March 2007.

INSPIRE (2008). "Infrastructure for Spatial Information in the European Community as regards metadata". Official Journal of the European Union. Directive 2007/2/EC of the European Parliament and of the Council of 3 December 2008.

Cartographic Institute of Catalonia (2001) "Especificacions per al format Mapa de MiraMon comprimit (MMZ) de la Base topogràfica de Catalunya 1:5 000 (BT-5M) v2.0". Barcelona. Available in Internet: http://www.icc.cat/cat/content/download/9992/32603/file/bt5mv20mm0_01ca.pdf, last access March 23rd 2010.

Cartographic Institute of Catalonia (2008) "Especificacions per al format Mapa de MiraMon comprimit (MMZ) de la Base topogràfica de Catalunya 1:25 000 (BT-25M) v1.0". Barcelona. Available in Internet: http://www.icc.cat/cat/content/download/9995/32618/file/bt25mv10mm0_01ca.pdf, last access 23rd March 2010.

MANSO, M. A., NOGUERAS-ISO, J., BERNABÉ, M.A., ZARAGOZA-SORIA, F.J. 2004. "Automatic metadata extraction from geographic information". 7th AGILE Conference on Geographic Information Science. Spatial Data Infrastructure III, 379-385.

MANSO, M. A., BERNABÉ, M. A. (2009). "Metadatos implícitos en la información geogràfica: caracterización del coste temporal y de los tipos y tasas de errores en la compilación manual". Geofocus (Articles), nª9, p. 317-336. ISSN: 1578-5157.

MOELLERING, H., AALDERS, H.J.G.L., CRANE, A. (2006). "World spatial metadata standards". Elsevier, The Netherlands.

NEBERT DD. (2004). "The SDI Cookbook. Global Statial Data Infrastructure technical working group". Available in Internet: http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf , last access June 07th 2010.

NEBERT DD, WHITESIDE, A., VRETANOS, P. (2007). "OGC CSW OpenGIS Catalogue Service Implementation Specification". OGC 07-006r1.

TeIDE (2009)."Manual de usuario de la aplicación CatMDEdit para la creación y edición de metadatos geográficos. Versió 4.5". Available in Internet: http://catmdedit.sourceforge.net, last access May 21st 2010.

TOLOSANA, R., NOGUERAS, J., ZARAZAGA, F.J. (2006). "El impacto de la calidad de los metadatos en los servicios de búsqueda de una SDI". Communication to JIDEE 2006.

PONS, X. (2001) "MiraMon. Geographic Information Systems and Remote Sensing software. v. 4". Center for Ecological Research and Forestry Applications, CREAF. Bellaterra. 286 p. ISBN: 84-931323-4-9

ZABALA, A. (2001) "Disseny d'una aplicació de Gestió de Metadades: el Gestor de Metadades del MiraMon (GeMM)". In: "Treballs del Màster en Teledetecció i Sistemes d'Informació Geogràfica, 4ª edició". 8 p. Institute for Space Studies of Catalonia (IEEC). Barcelona. ISBN: 90-77017-71-2

ZABALA, A., PONS, X. (2002) "Image Metadata: compiled proposal and implementation" In Benes, T. (ed.) "Geoinformation for all". Millpress, Rotterdam. ISBN: 90-77017-71-2 (p. 674-652)

ZABALA, A., PONS, X., MASÓ, J. (2006) "Metadatos para Capas y Series Cartográficas. Modelo de Herencia de Metadatos". Communication to JIDEE 2006. 10 p.