Universitat Autònoma de Barcelona

INSPIRE Conference 2010

Comparative Quality Assessment of Metadata.

Two Regional SDI case studies (IDEC & IDE-CLM).

Paula Díaz[1], Joan Masó[2] and Jordi Guimet[3]
[1]Dep. de Geografia de la UAB.
[2]CREAF
[3]IDEC

22-25 June 2010, Kraków

CONTENT

1. INTRODUCTION.

- 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 the SDI develop and maintain data catalogs [Nebert 2004].
- Aim of this study:
  - Detect and analyze errors in the metadata sets,
  - Determine the nature of these errors,
  - Determine their percentage of presence,
  - Make recommendations for avoiding them.

2. CONTEXT

- The concept of SDI was created in 1994.
- The only commercial products available for the creation of metadata were based on the FGDC standard CSDGSM.

THE IDEC (Catalonia)

- Created in **2002**
  - The Metadata catalog.
  - The program **MetaD** → Creation and edition of metadata sets.

IDE-CLM (Castilla la Mancha)

- Created in **2006**. Under the INSPIRE directive.

Composition of the SDI

- Organisms of different levels:
  - Public administration,
  - University departments,
  - Research centers.
- Geographic Data:
  - Cartography
  - Text Tables
  - Magazines,etc.
- The ICC metadata records have been excluded of the IDEC records (main provider). Due to its large volume metadata records (high quality).

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

2.1. Contextualization in the state level.

- GTIDEE was created in 2002 → INSPIRE → IDEE
- In Spain there are a total of 43 SDI,
  - 10 of them national,
  - 19 local,
  - 14 regional, with metadata catalogs .
- Most SDIs in Spain were created in the context of INSPIRE directive.
- We consider the IDEC was the first data infrastructure in Spain to create and publish metadata in a standard catalog (2003).

2.1. Contextualization in the state level.

- Mostly spanish regional SDI have recently adapted their metadata to ISO 19115.
- The IDEC is the spanish SDI with a major number of metadata records.

Regional SDI	Tipe of catalogue	Metadata datasets	Open acces	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	---
Catalunya	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

2. 2. Contextualization in the European arena.

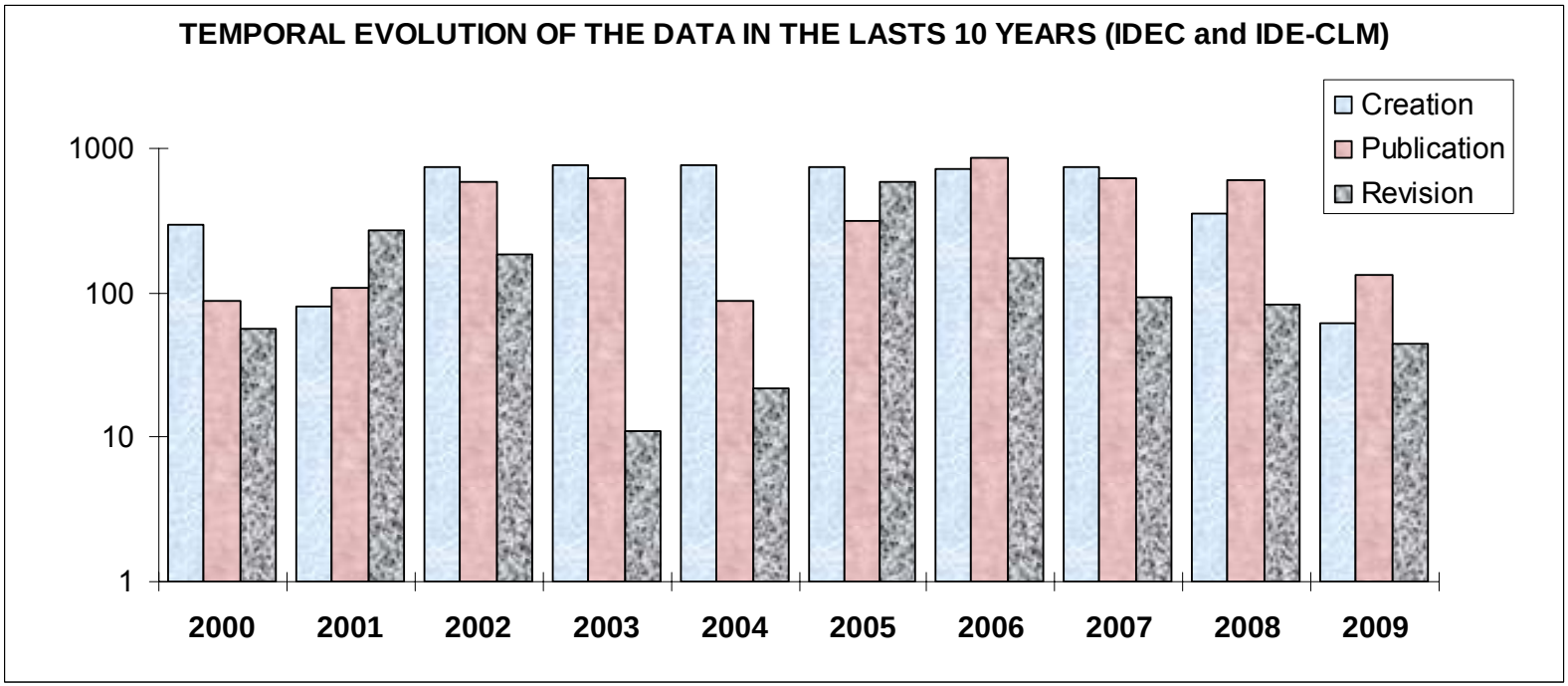
- INSPIRE stablish **25 mandatory elements.**
- ISO 19115 stablish **9**.
- If SDI currently have difficulty validating metadata under a less rigorous standard, the integration with the European directive can be very difficult.

MANDATORY ELEMENTS BY INSPIRE	ELEMENTS BY ISO 19115
Title	Mandatory
Abstract	
Resource language	
Metadata date	
Topic category	
Responsabile party information	Mandatory at least organisation name, individual name or rol
Responsible party role	
Date of publication	Mandatory at least one of the three
Date of last revision	
Date of creation	
Geographic bounding box	Conditionals
Metadata language	
Keyword value	Optional
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	

2.3. Contextualization in the international arena.

- Comprovets has examined the global situation in 2004.
- The author claims to find satisfactory results in a search to a data catalog depends largely on the quality of metadata sets.
- The Average of data providers of geo-portal is **50**
- The Average of visitors is **5000** (median: 1000)
- There is a decrease of providers in the last decade that represents:
  - A reduction of data publication in last decade.

2.3. Contextualization in the international arena.



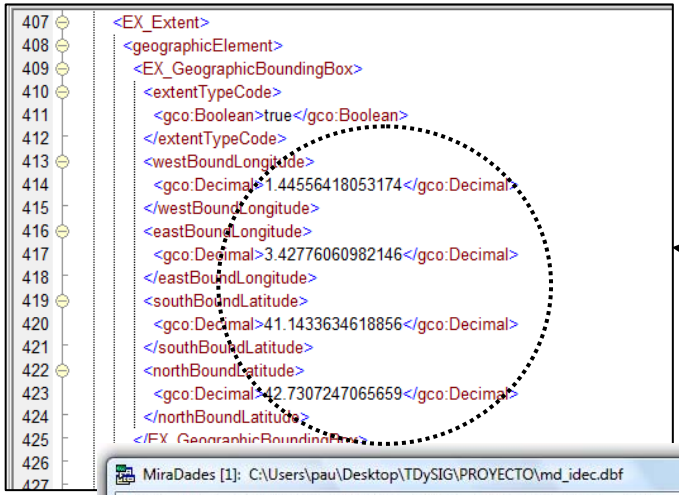
- In the IDEC and IDE-CLM catalogues, the created, published and revised data present a decline in the lasts years.

3. METHODOLOGY.

- The **standardization** facilitates the:
  - Interoperability.
  - Comparison.
- This study is based on three standards:
  - ISO 19115 → to establish the elements as a basis for study.
  - ISO 19139 → to understand the XML documents.
  - OGC-CSW → to download XML metadata documents.

  *http://delta.icc.cat/wefex/client?id=urn:uuid:**UUID\b87a8b73-394c-9ad6ccc**&idioma=**ca**&do=mostraResum&skipStyling*
  - At time of this study only IDEC and IDE-CLM had OGC catalogs.

Creation of a metadata BataBase.



- Metadata Sets in XML.
- Data base containing all the metadata sets

MiraDades [1]: C:\Users\pau\Desktop\TDySIG\PROYECTO\md_idec.dbf

Fitxer Edició Visualització Registres Camps Informació Eines Ajuda

Registres: |< << < > >> >| Camps: |< << < > >> >|

	UUID	XMIN	XMAX	YMIN	YMAX	ESCALA
14223	146ef0be-a0e6-11dd-a834-8dd3749edb99	1.0858963630697	1.1532026128995	42.713759601520	42.774818078269	33000
14224	15053d52-a0e6-11dd-a834-8dd3749edb99	1.0055675830264	1.0848255161746	42.727209557736	42.779439294603	33000
14225	159b62d6-a0e6-11dd-a834-8dd3749edb99	1.0055675830264	1.0848255161746	42.727209557736	42.779439294603	33000
14226	161770aa-a0e6-11dd-a834-8dd3749edb99	1.3624869317687	1.4148820739171	42.604456314491	42.641826615822	33000
14227	16d14ace-a0e6-11dd-a834-8dd3749edb99	1.3624869317687	1.4148820739171	42.604456314491	42.641826615822	33000
14228	174fa292-a0e6-11dd-a834-8dd3749edb99	1.3215321756892	1.3890137841563	42.602560838967	42.643041357241	33000
14229	17c93f66-a0e6-11dd-a834-8dd3749edb99	1.3215321756892	1.3890137841563	42.602560838967	42.643041357241	33000
14230	1858600a-a0e6-11dd-a834-8dd3749edb99	1.2621349128855	1.3456838563784	42.589709910313	42.655470639001	33000
14231	18ee858e-a0e6-11dd-a834-8dd3749edb99	1.2621349128855	1.3456838563784	42.589709910313	42.655470639001	33000

- Extraction of all the mandatory elements and also other optional elements.

Creation of a metadata BataBase.

- All the elements extracted are mandatory by the INSPIRE directive regarding metadata.
- DATABASE:
  - Columns: Mandatory and optional elements.
  - Rows: XML files identified by their UUID.
- The IDEC:
  - 35 columns and 14 616 rows .
- The IDE-CLM:
  - 35 columns and 98 rows.

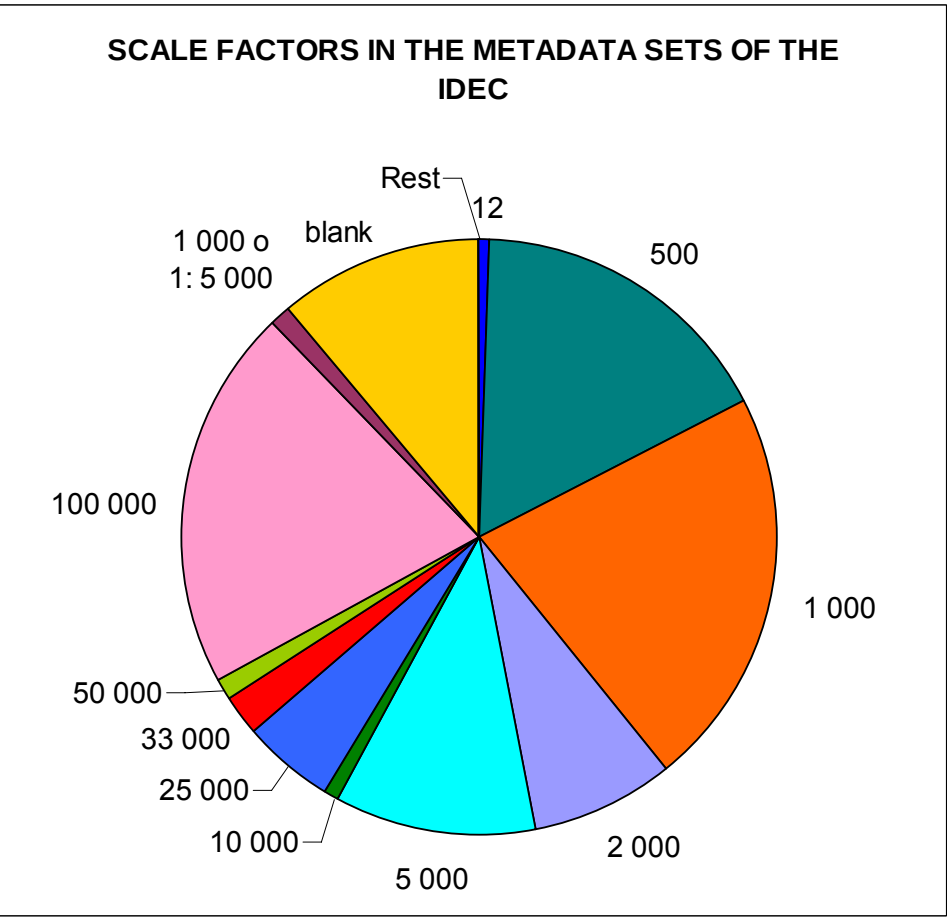
4. QUALITY ANALYSIS OF THE METADATA SETS.

- **4.1. Errors in the metadata sets of autonomic catalogs.**
  - Lack of compliance with the ISO 19115 requirements.
  - Some examples of **mandatory** elements:

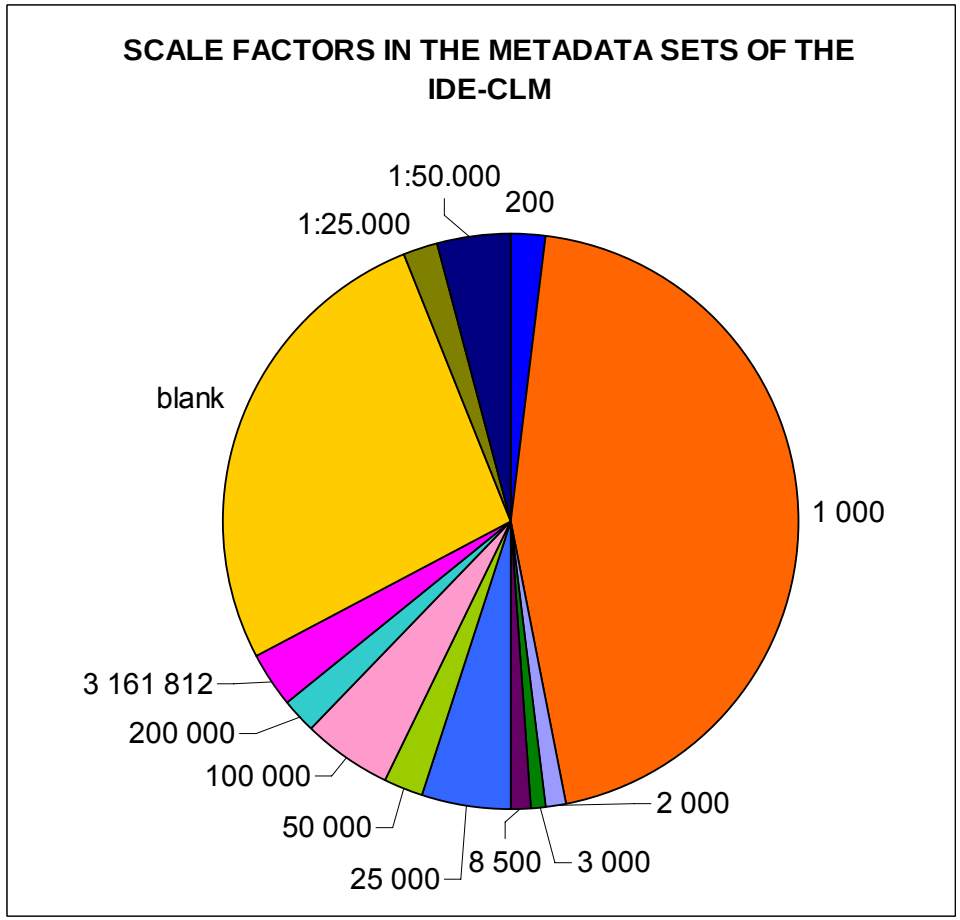
	IDEC	IDE-CLM
Lack of metadata date:	353 (2.42%)	----
Lack of datasets dates:	1 779 (12.17%)	36 (36.7%)
Lack of extent:	33 (0.23%)	29 (29.21%)
Lack of creator contact:	39 (0.27%)	2 (2.04%)

Some examples of optional elements:

- Scale factors: (Inconsistent for a map)



341 (2.33%)



9 (0.09%)

4.2. Practices for improvement.

- This practices don't contravene strictly the ISO 19115 requirements, but the **accessibility** & **understanding**.
- IDEC: Titles

```
<title>
 <gco:CharacterString>314-80</gco:CharacterString>
</title>
```

- IDE-CLM: Keywords

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%***

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

- Three main reasons:
  - Lack of accurate information by the metadata creator (such as the date of creation of the dataset),
  - The difficulty of determining the information required (the scale information in tabular information with x,y positions),
  - Ignorance of certain factors (e.g. processes).
- Methods for creating metadata are not exempt of the generation of errors.
- There is a high percentage of error in the manual compilation of metadata elements (Mansó 2009).

5. RECOMMENDATIONS.

- The most immediate recommendation is correct the lacks of mandatory elements.
- The SDI can inform the metadata providers to facilitate them solve errors.
- We recommend establishing common rules for generic creation of metadata titles.
- Use a thesaurus for the selection of keywords.

6. CONCLUSIONS.

- It's possible to carry out a systematic review of the metadata sets of SDI in order to:
  - Detect errors, weaknesses, or lacks of good practices,
  - Determine the organisms responsible of a specific problem.
- Periodic quality checks of the metadata can be made to detect errors or lacks.
- INSPIRE is more demanding than ISO 19115 respect to the completeness of the metadata.
- This analysis of metadata manifests the presence of different kinds of errors in the metadata sets.
- Much metadata sets have errors that can't be involuntarily made from common metadata tools.
- There is a lower quality of description in the optional elements.

- There is a need to implement more quality control procedures.
- The average error for all metadata sets is around 3.84% in the IDEC and 11.73% in the IDE-CLM.
- This analysis applied to regional SDI, shows that quality is a compromise between agility for providers who create metadata and the needs of the end-user who wants as much detailed information as possible.

COMMON ERRORS FOUND IN THE SDI	IDEC	IDE-CLM
Metadata date in blank	2.42%	0%
Data dates in blank (the three)	12.17%	37%
Creation date later than metadata date	3.36%	0%
Creation date "1900-01-01"	9.48%	0%
Topic category not in codelist	9.70%	0%
Topic category in blank	3.41%	3%
Contact information in blank	0.27%	2%
Geographic extent not in angles (lat/long)	0.18%	60%
Minimum coordinate greater than the maximum	0.01%	1%
Data language in blank	2.44%	26%
Incorrect metadata language	0.35%	3%
Inconsistent scale factors	2.33%	9%
Average error	***3.84%***	***11.73%***

- *Thanks for the attention.*

- *paula.diaz@uab.cat*
- *joan.maso@uab.cat*
- *jordi.guimet@icc.cat>*