Análisis comparativo de los metadatos distribuidos por la IDEC y la IDECLM como ejemplos de IDE autonómicas.

P. Diaz[1], J. Masó[2] y A. Zabala[1]

[1] Departamento de Geografía de la Universidad Autónoma de Barcelona.
[2] Centro de Investigación Ecológica y Aplicaciones Forestales (CREAF)

Resumen

Desde 2003 han aparecido numerosas herramientas de edición y catalogación de metadatos. Los metadatos se formalizan usando la ISO 19115 que define un conjunto de elementos posibles. Del conjunto, sólo algunos resultan de obligada cumplimentación. Sin embargo, la generación de metadatos continúa siendo un proceso costoso y metódico que depende de la buena voluntad del productor. Para cumplir sus objetivos, las IDE centralizan estos metadatos y ofrecen herramientas de búsqueda de cartográfica basadas en metadatos. Este estudio analiza la presencia de errores y malas prácticas en los documentos de metadatos, su naturaleza y proporción para dos IDE autonómicas de naturaleza muy distinta: la Infraestructura de Datos Espaciales de Cataluña (IDEC) que ha sumado 27007 registros desde 2002 y la Infraestructura de Datos Espaciales de Castilla la Mancha (IDECLM) que empezó en 2006 y cuenta con un centenar de registros. El artículo describe la metodología de recopilación del subconjunto de elementos de metadatos obligatorios y otros considerados relevantes. El estudio indica las fuentes de estos errores y malas prácticas y sugiere la necesidad de incidir en la calidad de los metadatos.

Palabras clave: catálogo de datos, metadatos, IDEC, IDECLM, calidad.

1. INTRODUCCIÓN.

En la última década, la distribución del dato geográfico ha experimentado una renovación. Las librerías de datos espaciales, los geoportales y la diversificación de la producción y uso de los datos geográficos dan buena cuenta de ello. Este cambio se traduce en la aparición de las Infraestructuras de Datos Espaciales (IDE) y sus herramientas comunes como los catálogos de datos y metadatos. La iniciativa INSPIRE aboga desde 2002 por una ampliación de los geoservicios que hagan posible la Infraestructura de Datos Espaciales de Europa donde los catálogos de datos y de servicios tendrán un papel fundamental [INSPIRE 2007].

La primera definición formal del término IDE (Spatial Data Infrastructure, SDI en inglés) se realizó en EEUU. El término IDE se define como las tecnologías, políticas, y las personas necesarias para promover el intercambio de datos geográficos en todos los niveles de gobierno, los sectores privado y no lucrativo, y la comunidad académica [FGDC 1994].

Las funciones básicas de una IDE [Goodchild 2007] son:

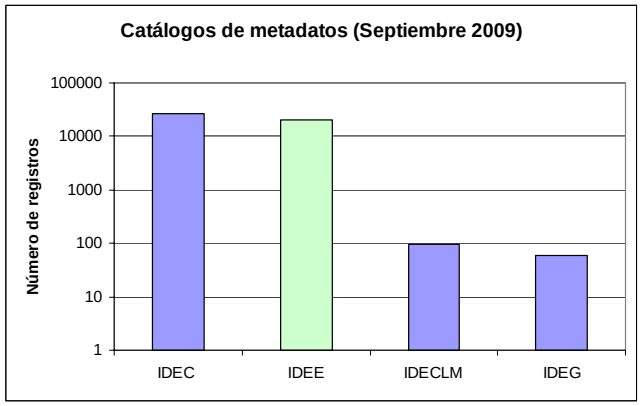
- Publicación de datos facilitados por ciertos proveedores.
- Transmisión de datos: acceso a recursos heterogéneos.
- Integración de datos: Recopilación de información, evitando su duplicación.

Para cumplir estas funciones las IDE elaboran y mantienen catálogos de datos [Nebert 2004]. Estos catálogos registran cada conjunto de información geográfica a partir de una ficha de metadatos que lo describe. Generalmente, los conjuntos de información geográfica son mantenidos y distribuidos por el proveedor mientras que el catálogo de la IDE centraliza los metadatos que describen estos conjuntos de información geográfica. Por ello, los metadatos juegan un papel relevante en el proceso de selección, transferencia y manipulación del dato geográfico.

Este estudio analiza la calidad, en sus diferentes aspectos, de los documentos de metadatos almacenados en los catálogos de metadatos en las IDE autonómicas. Además, analiza la naturaleza de los factores que pueden conducir a errores o imprecisiones en la creación de estos metadatos y se emiten recomendaciones para

evitarlos. El análisis de la calidad de los conjuntos de datos geográficos queda fuera del alcance de este estudio. Dado que los catálogos de metadatos son utilizados básicamente para buscar productos cartográficos, nos motiva la idea que apuntan algunos estudios sobre la calidad de metadatos [Tolosana 2006], referente a la situación frecuente en que los usuarios de catálogos de datos no obtienen resultados satisfactorios en la búsqueda inicial y deben modificar sus criterios de búsqueda para obtener algún resultado.

Actualmente, según datos de la IDEE, se dispone de 29 IDE dentro del Estado Español. De ellas, 12 son autonómicas y todas ellas disponen de catálogos de metadatos propios. Sin embargo, solamente 3 de las IDE autonómicas han hecho público un geo-servicio de un catálogo estándar OGC-CSW: la de Cataluña (IDEC), la de Castilla la Mancha (IDECLM) y la de Galicia (IDEG). La IDEE también permite el acceso a su catálogo a través del estándar CSW (Gráfica 1). Sin embargo, sólo la IDEC y la IDECLM permiten la descarga de los documentos de metadatos completos bajo el esquema ISO 19139, por lo que permiten conocer con detalle el contenido de los registros de metadatos bajo el estándar ISO 19115. Actualmente los catálogos de la IDEE y la IDEG sólo retornan un pequeño conjunto de elementos de metadatos bajo el estándar Dublin Core. Este conjunto era demasiado restringido para permitir el mismo nivel de detalle en todos los casos por lo que finalmente se han descartado para el estudio realizado de calidad.



Gráfica 1. Número de registros de metadatos disponibles en las IDE autonómicas con servidor CSW y en la IDEE.

2. CONTEXTUALIZACIÓN DE LA IDEC Y LA IDECLM.

El concepto de las IDE nace en 1994, año de la creación de la National Data Spatial Infrastructure (NDSI) [FGDC 1994]. No será hasta el 2002 que la IDEC inicia su trabajo y catalogación de metadatos con la primera versión del catálogo de metadatos y el programa MetaD; lo que significa un año antes de la aprobación del estándar 19115:2003 y 5 años antes de la aprobación del 19139:2007. El elevado volumen de metadatos de esta IDE (Gráfica 1) se debe a su incansable labor de motivación de los diversos organismos proveedores de información. Es importante destacar que este elevado volumen de datos disponibles actualmente proviene de una gran diversidad de organismos proveedores de estos datos, que han elaborado metadatos con diferentes herramientas, diferentes versiones de MetaD [IDEC 2009] u otras herramientas de ayuda a la elaboración de metadatos (Tabla 1). Estos metadatos documentan información de naturaleza muy variada, siendo algunos sobre documentos no cartográficos. El elevado volumen también se debe al hecho de que la unidad de almacenamiento es la hoja, por lo que algunos productos de alto nivel de detalle pueden estar cortados en decenas o centenares de hojas, cada una de las cuales genera un registro de metadatos.

Por otra parte, la IDECLM fue creada en 2006. Su catálogo dispone de un número reducido de metadatos. La unidad de almacenamiento es el conjunto de información geográfica (y no la hoja como en el caso de la IDEC) por lo que el volumen de metadatos presente debe ser necesariamente menor; sin embargo también se observa que el número de proveedores de metadatos es considerablemente más reducido (Tabla 1). Los metadatos analizados del catálogo de la IDEC son de junio de 2009, los analizados del catálogo de la IDECLM son los disponibles en septiembre de 2009.

Ambas IDE se enmarcan en la directiva INSPIRE y reciben el soporte de diversas Administraciones públicas.

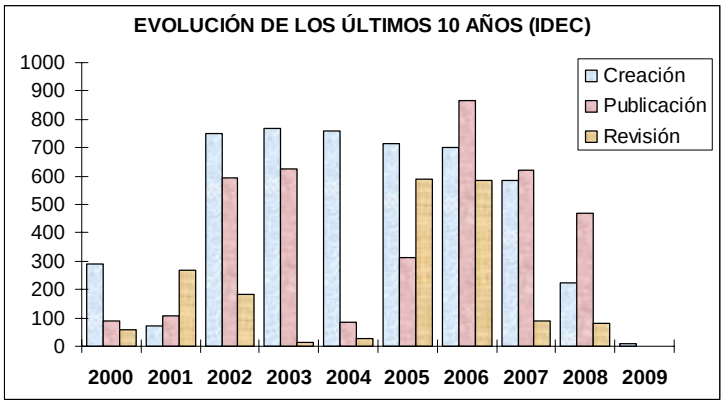
	Organismos registrados	***Documentos totales***
IDEC (06/2009)	112	27007
IDEC excluyendo el ICC (06/2009)	111	14231
IDECLM (09/2009)	9	98

Tabla 1. Organismos y número de datos de las IDE analizadas.

Comprovets analiza la situación de las principales IDE nacionales en 2004 [Comprovets 2004]. Se hace especial hincapié en la importancia de los metadatos en las búsquedas exitosas y del éxito de las búsquedas para la accesibilidad a los mismos.

El estudio de Comprovets reduce a tres los grupos de personas implicados en un geoportal: los proveedores, los administradores y los usuarios finales. Se establece un promedio próximo a 50 proveedores de datos espaciales por geoportal. Según nuestra información la IDEC contaba en verano de 2008 con 79 organismos y en junio de 2009 con 112, actualmente la IDECLM dispone de aproximadamente de una decena de proveedores (Tabla 1).

En el mismo estudio se perfila una evolución decreciente de proveedores. Esa tendencia a la baja se relaciona directamente con una reducción de datos publicados en los últimos años. De igual modo sucede en la IDEC, tal y como refleja la tendencia a la baja de la gráfica de evolución de los datos publicados, creados y revisados en los últimos 10 años (Gráfica 2). La creación de datos presenta su máximo en 2003, la revisión en 2005 y la publicación en 2006, los tres casos seguidos de una disminución. El conjunto de proveedores de la IDEC lo forman organismos públicos, privados, en menor medida, e instituciones de diferente índole. Principalmente se compone de organismos de diferentes niveles de la Administración pública, departamentos universitarios y centros de investigación. De esto se desprende que los productores de datos geográficos no representan únicamente instituciones estrictamente especializadas en la generación de datos geográficos.



Gráfica 2. Evolución temporal de la creación, publicación y revisión de metadatos en la IDEC.

El estudio de Comprovets afirma que el número reciente de usuarios finales de las IDE se ha estabilizado y no ha aumentado. El promedio se establece en 5000 visitantes, aunque la media se encuentra en torno a los 1000 debido a las amplias diferencias entre geoportales. Además, se pone de manifiesto el volumen de datos 13 veces mayor en EEUU que en Europa y se establece en 3500 el promedio de datos accesibles en los geoportales. La información extraída del geoportal de la IDEC revela que esta Web contó en 2008 con 96412 visitas, 4708 de ella al catálogo. El catálogo de datos de la IDEC en junio de 2009 albergaba 27007 documentos de metadatos, el de la IDECLM contaba con 98 en septiembre del mismo año (Tabla 1). El número real de metadatos del catálogo de la IDEC se ve triplicado debido a la traducción de éstos a tres idiomas: catalán, castellano e inglés.

El conjunto de documentos almacenados en la IDEC no sólo constituyen datos estrictamente cartográficos, sino que se dispone de información geográfica en formato de texto, libros, revistas, estudios o tablas, perteneciente a diferentes categorías temáticas. Las aportaciones más importantes de metadatos al catálogo de la IDEC son los documentos generados por el Instituto Cartográfico de Cataluña (ICC) y aquellos del proyecto IDEUnivers. El organismo que más documentos de metadatos aporta a la IDECLM es el Instituto de Desarrollo Regional: IDL, seguido del Instituto de Estadística de Castilla La Mancha.

En definitiva las IDE actúan como contenedores de información geográfica, generada por múltiples proveedores bajo diferentes criterios de operabilidad y con diversos objetivos. La IDEC constituye la base de datos geográficos regional más importante del estado español en volumen de datos, variedad temática y número de proveedores.

3. METODOLOGIA.

La importancia de la estandardización radica en la homogeneización del *trabajo*, lo cual facilita la interoperabilidad y la comparación. Todo ello contribuye a la unificación de esfuerzos encaminados a la mejora de la calidad, evitando la duplicación de documentos y esfuerzos.

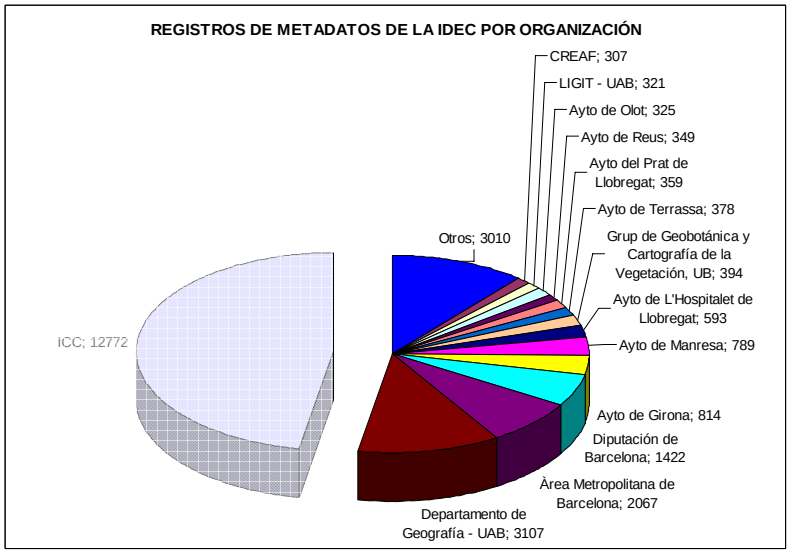
La metodología de este trabajo se basa en 3 estándares: la ISO 19115 [ISO/TC211 2003 y Moellering 2006] define el contenido de cada elemento de metadatos, la ISO19139 [ISO/TS 19139 2007] define como codificar estos metadatos en XML y el OGC-CSW (Catalog Service for Web) [OGC-CSW 2007] admite la posibilidad de publicar y buscar conjuntos de metadatos. A partir de las direcciones de los servidores del IDEC (http://delta.icc.es/indicio/csw) y de la IDECLM (http://161.67.130.140:8080/geonetwork/srv/en/csw) es posible enumerar los identificadores de registro con la operación GetRecords y posteriormente obtener todos los documentos de metadatos en formato ISO19139 con la operación GetRecordByID.

El estándar ISO 19115 establece tres categorías de elementos:

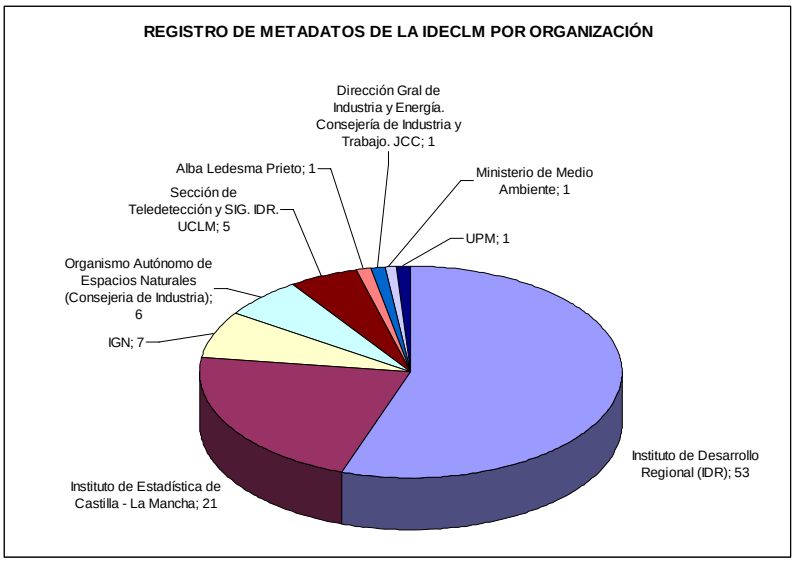
- Obligatorios, elementos que deben ser cumplimentados.
- Condicionales, elementos obligatorios en caso de falta de otros elementos.
- Opcionales, no son de obligada cumplimentación.

Para facilitar el estudio, se ha generado una base de datos a partir de un subconjunto de los elementos de metadatos disponibles en los documentos XML de la ISO 19139. Se han extraído la totalidad de los elementos obligatorios, y aquellos que, aun siendo opcionales, han sido considerados de importancia relevante para la comprensión del dato geográfico. El resultado final es una base de datos en la que los elementos obligatorios y opcionales corresponden a campos de las tablas, mientras que los registros corresponden a los documentos de metadatos descargados, designados con su identificador único. La utilización de una base de datos como método para el tratamiento de la información, ha permitido trabajar simultáneamente con gran cantidad de elementos y registros, sistematizando el proceso de búsqueda y de detección de errores. Debido a la cardinalidad múltiple de algunos elementos, en ocasiones concretas hemos recopilado sólo la primera entrada, considerándola la más significativa.

En el caso de la IDEC se generó una base de datos de 14231 registros y 32 campos. Se analizaron todos los documentos a excepción de los pertenecientes al Instituto Cartográfico de Cataluña (ICC) al presuponerse su calidad y presentar un conjunto muy homogéneo y de gran número al estar cortados por hojas. En consecuencia, los metadatos analizados pertenecen a datos muy variados de instituciones no estrictamente especializadas en la generación de cartografía (Gráfica 3). En el caso de la IDECLM se han analizado la totalidad de documentos disponibles, generando una base de datos de 98 registros y 32 campos (Gráfica 4).



Gráfica 3. Registros del catálogo de la IDEC por organización con más de 300 registros en el catálogo. Los registros del ICC han sido excluidos del estudio.



Gráfica 4. Registros del catálogo de la IDECLM por organización.

4. ANÁLISIS DE LA CALIDAD EN LOS DOCUMENTOS DE METADATOS.

A continuación se analizará la calidad de los documentos de metadatos, basándonos en dos criterios. Primero se estudiarán los errores que presentan los documentos al no cumplir alguna regla impuesta por el estándar ISO 19115. En segundo lugar se efectuarán recomendaciones, que no estando impuestas por el estándar, consideramos que son necesarias para la comprensión del dato, y la capacidad de los catálogos de responder mejor a las búsquedas de los usuarios.

4.1. Errores presentes en los documentos de metadatos de los catálogos autonómicos.

4.1.1. Los elementos obligatorios en el estándar ISO 19115 son:

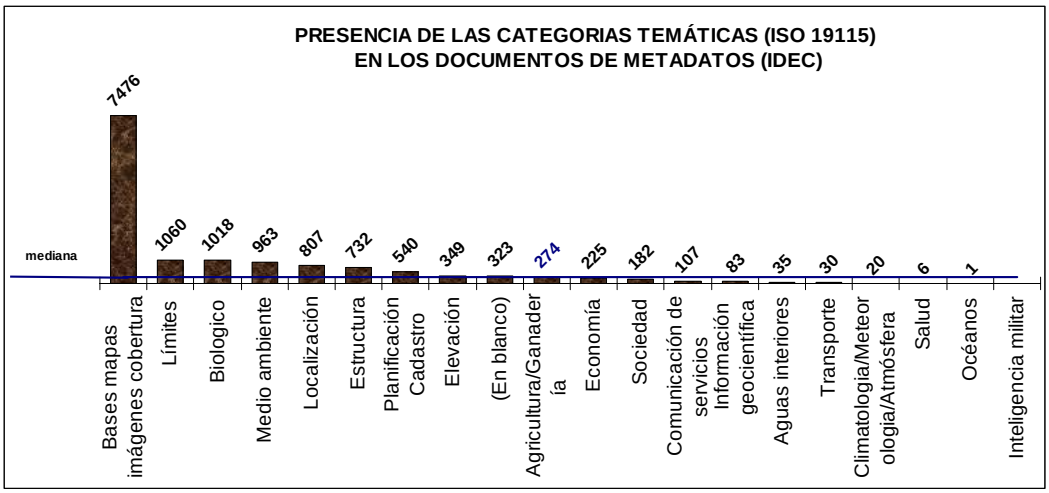
- el título,
- el resumen,
- la fecha de edición del metadatos,
- al menos una de las tres fechas referentes a los datos: creación, publicación o revisión,
- la categoría temática,
- información de contacto del creador del documento de metadatos,
- el idioma de los datos.

El título, el resumen y el idioma de los datos se tratarán en el prado 4.2., referente a malas prácticas en la documentación de metadatos, ya que se analizarán otros aspectos además de su obligatoriedad.

4.1.1.a) Catálogo de la IDEC.
La fecha de edición de los metadatos es obligatoria, así como una de las tres fechas referentes a los datos: creación, publicación o revisión. Son 320 documentos (2.25%) los que carecen de la fecha de los metadatos, además de no estar descrita ninguna de las tres fechas referentes al dato en 1746 documentos (12.3%). La fecha de creación del dato es sospechosamente posterior a la de creación del metadato en 219 casos (1.54%).

Son 19 las categorías temáticas que establece la ISO 19115 (Gráfica 4).

Las categorías temáticas ayudan a caracterizar el tipo de dato presente en la IDEC, el gráfico ilustra más de la mitad de los documentos como "Bases de mapas o imágenes de la cobertura terrestre". Las categorías temáticas se definen en la ISO 19139 como un *Codelist* con 19 códigos escritos en lengua inglesa. En 1419 documentos (11%) se definen las categorías temáticas con palabras que no pertenecen al *Codelist* (generalmente se usan descripciones equivalentes en el idioma de creación de los metadatos). Dado que la categoría temática es obligatoria, consideramos un error la ausencia de ésta en 323 casos (2.27%).

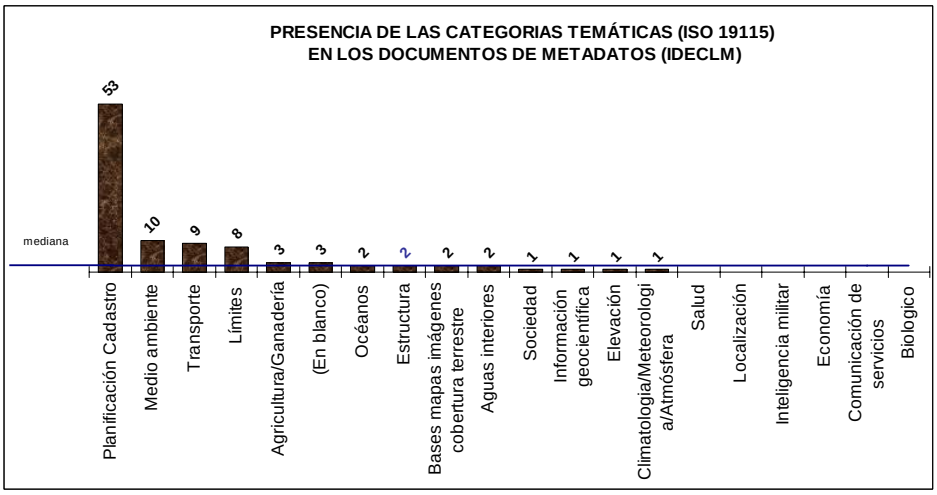


Gráfica 4. Volumen de datos de la IDEC que pertenecen a las 19 categorías temáticas .

Son tres las referencias principales del estándar ISO 19115 a la información de contacto del creador de metadatos: nombre individual, nombre de la organización y el cargo. En nuestra recopilación de metadatos, comprobamos que 6 docuementos (6.12%) no cuentan con ninguna de estas tres informaciones al respecto, siendo obligatoria al menos una de ellas.

4.1.1.b) Catálogo de la IDECLM.
La fecha de edición del documento de metadatos se encuentra descrita en la totalidad de los documentos de metadatos, mientras son 36 los documentos (36.7%) que no presentan ninguna de las tres fechas referentes al dato.

Las categorías temáticas se encuentran descritas en todos los documentos excepto en 3 (3.06%). La más representada es la categoría de "Planificación del catastro" con 53 registros (54%) y 6 categorías temáticas no presentan ningún dato geográfico en la IDECLM (Gráfica 5).



Gráfica 5. Volumen de datos del IDECLM que pertenecen a las 19 categorías temáticas

En cuanto a la información de contacto obligatoria, son 2 los documentos que carecen de la misma (2.04%).

4.1.2. Los elementos condicionales del estándar ISO 19115 a tratar son:

- la extensión, debe introducirse en coordenadas geográficas (latitud/longitud), sino debería realizarse una descripción de la extensión geográfica.
- el idioma de los metadatos, si no se ha descrito en la codificación del documento.

4.1.2.a) Catálogo de la IDEC.
La extensión geográfica se define en la totalidad de los documentos mediante las cuatro coordenadas geográficas. El estándar ISO 19115 establece que debe estar expresada en ángulos, bien radianes o grados (latitud/longitud). No obstante contamos con 765 documentos (5.37%) de metadatos que presentan las coordenadas en números muy elevados, los cuales consideramos coordenadas expresadas en metros, en relación al sistema de referencia que indican: el UTM31N ED50 y el UTM30N ED50. Se han catalogado también, 468 documentos (3.29%) de metadatos en los que la mínima coordenada es mayor a la máxima, (Tabla 2).

	total
XMIN>XMAX	412
YMIN>YMAX	56
Total general	***468***

Tabla 2. Número de incoherencias entre mínima y máxima coordenada (IDEC).

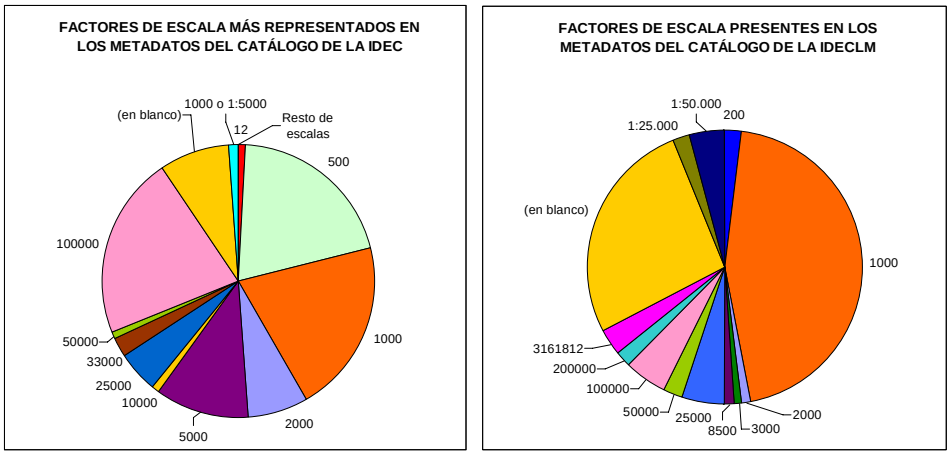
En cuanto al idioma de los metadatos, un 100% de los documentos lo detallan. Sin embargo comprobamos que se han redactado en castellano 68 documentos (0.48%), los cuales indican como idioma del dato el castellano. Esto podría deberse a una confusión entre el elemento idioma de los metadatos e idioma de los datos.

4.1.2.b) Catálogo de la IDECLM.
La extensión geográfica no se indica en 29 documentos (29.21%). 64 (64.4%) documentos describen la extensión en coordenadas con valores muy altos y semejantes a las coordenadas UTM de la zona, debiendo estar, según ISO 19115, en ángulos (longitud/latitud). Un solo documento presenta incoherencias entre la mínima y la máxima coordenada.

El idioma de los metadatos se encuentra sin especificar en 21 documentos (21.42%), los cuales tampoco especifican el idioma del dato. El idioma definido en 3 documentos es el inglés (3.06%), lo que se constata como error al comprobar que los metadatos se han redactado en castellano.

4.1.3. Los elementos opcionales que han sido tratados son:
- la escala,
- el sistema de referencia horizontal (SRH),
- los formatos de representación del dato geográfico,
- la topología.

4.1.3.a) Catálogo de la IDEC.
Los factores de escala reunidos son 63, de los cuales 7 son incoherentes para un mapa, o bien muy pequeños ("1", "12", "13", etc) o bien mixtos ("1000 o 1:5000"), (Gráfica 6). Estos factores de escala incoherentes se encuentran en 390 documentos (2.74%). A pesar de no ser obligatorio, se considera un elemento importante para el tratamiento del dato geográfico y su búsqueda.



Gráficas 6. Factores de escala más utilizados en los metadatos del catálogo de la IDEC ("Resto de escalas" engloba aquellas inferiores a 100 registros). **Gráfica 7**. Factores de escala usados en los metadatos de la IDECLM.

Los SRH de los datos de la IDEC son 5 diferentes. Se ha comprobado que las 20 coordenadas relativas al sistema de referencia EPSG:32633 - WGS84 / UTM zona 33N no son coherentes con la zona indicada.

4.1.3.b) Catálogo de la IDECLM.
Son 72 los documentos que contienen la especificación del factor de escala (73.46%). Destaca la utilización de escalas poco comunes y en formato de fracción no contemplada por la ISO 19115 ("3161812", "1:25.000", "1:50.000"), aunque representan sólo 9 registros, (Gráfica 7).

Finalmente todos los documentos de metadatos presentan la descripción del SRH.

4.2. Malas prácticas.

Esta sección describe prácticas que, aun sin considerarse errores bajo el estándar ISO 19115, pueden reducir la interoperabilidad con otros sistemas o la capacidad de búsqueda de los catálogos de datos. Se trata el contenido de los elementos que consideramos importantes para la comprensión del dato geográfico. Los elementos tratados son:

- el título,
- el resumen,
- el formato,
- la topología,
- las palabras clave,
- SHR,
- fechas sobre los datos,
- idioma de los datos.

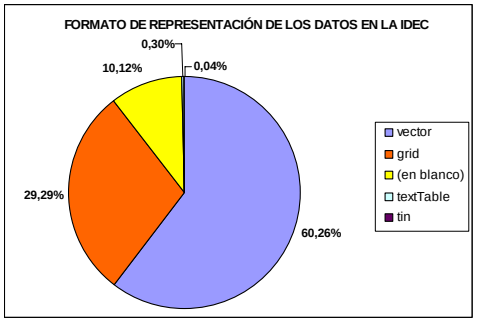
4.2.a) Catálogo de la IDEC.
El título y el resumen son dos de los elementos esenciales en un documento de metadatos, que permiten entender el dato e identificarlo fácilmente durante una búsqueda. La totalidad de los documentos contienen título, pero a 3998 documentos (28%) podrían aplicarse mejoras que aumentaran la comprensión del mismo por parte de los usuarios (Tabla 3).

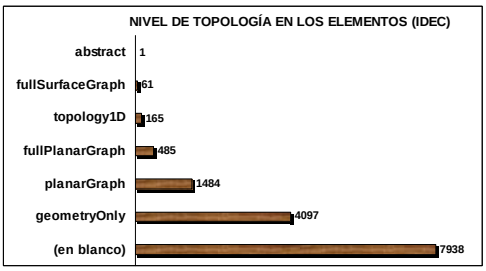
TÍTULOS	totales	
Secuencias alfanuméricas no descriptivas	3869	27%
Falta de concisión	129	1%
Total general	***3998***	

Tabla 3. Aspectos que albergan mejoras de la comprensión en los títulos (IDEC).

Aunque el estándar no establece limitaciones de longitud, hemos establecido el criterio de falta de concisión en un umbral de 110 caracteres por considerar que un título debe ser conciso y directo. Obtenemos así 129 títulos (1%) demasiado largos con partes que recomendamos sean trasladadas al resumen. Asimismo, son abundantes los títulos (3869, un 27%) que presentan sólo secuencias alfanuméricas no suficientemente descriptivas (generalmente códigos de hoja) que disminuyen la potencialidad de búsquedas en el catálogo. En este último caso recomendamos añadir información concreta sobre el tipo de producto, la fecha y la escala para completar el título. El resumen se debe cumplimentar en el 100% de los documentos, en el catálogo de la IDEC, este porcentaje llega al 99.9%, pero en un 2% de estos casos el resumen contiene el título al inicio lo que no aporta información adicional y desconcierta al usuario final.

El formato más abundante catalogado en los metadatos distribuidos por la IDEC es el vectorial, indicado en 8576 documentos (60.26%). El formato *grid* ha sido catalogado en 4168 documentos (29.29%), las tablas de texto en 42 (0.3%) y el TIN en 5 (0.04%) (Gráfica 8). Para los 60.26% conjuntos de información vectorial es posible y conveniente definir el nivel de topología. No obstante el nivel de topología no se documenta en 7938 documentos (55.77%), en 2196 documentos (15.43%) se indica algún grado de topología, mientras que 4097 documentos (28.78%) indican que no la presentan (Gráfica 9).





Gráfica 8. Formato de los datos (IDEC). **Gráfica 9**. Topología indicada en los documentos de metadatos (IDEC).

Debido a la abundancia y variedad de palabras clave hemos procedido a la selección de la primera perteneciente a cada tipo. El estándar ISO 19115 distingue cinco tipos: disciplina, lugar, estrato, temporal y tema.

PALABRAS CLAVE POR TIPOS (IDEC)	total
Tema	12494
Lugar	12333
Temporal	5643
Disciplina	0
Estrato	0
Total general	***17976***

	PALABRAS CLAVE MÁS UTILIZADAS	Tipo	total
1	ideunivers	Tema	3898
2	Espanya	Lugar	1546
3	Topogràfic	Tema	1532
4	Girona Barcelona	Lugar	1483
5	Lleida Tarragona	Lugar	1304
6	2003	Temporal	1254
7	Referència cadastral	Tema	796
8	Manresa	Lugar	790
9	2002	Temporal	640
10	Cartografia digital	Tema	617
	Total general		***13860***

Tabla 4. Volumen de las palabras clave por tipos (IDEC). **Tabla 5**. Palabras clave más usadas (IDEC).

Los resultados indican un número cuantioso de palabras clave, englobadas en tres tipos (Tabla 4). En la tabla 5 se muestran las diez palabras clave más utilizadas, algunas de ellas (como por ejemplo "Ideunivers") representan términos específicos que la IDEC utiliza como método interno de clasificación del dato.

En cuanto al SRH, información importante a la hora de representar el dato geográfico, se cuenta con 761 documentos (5.35%) carentes de esta información.

La IDEC utiliza la fecha de creación del dato "1900-01-01" como marca de desconocimiento de la misma. (observada en 1219 registros (8.56%)). No recomendamos esta práctica por 2 motivos: Al ser un criterio no estandardizado la exportación de estos metadatos a otros catálogos podría generar problemas de interpretación o validación. Al ser un elemento de metadatos obligatorio es mejor persuadir al usuario a indicar un fecha aproximada (solamente el año, por ejemplo) o incluso extender el estándar para inducir un intervalo de incertidumbre en la fecha.

El idioma del dato es obligatorio y no está indicado en 3719 documentos, un 26.1% de los casos. Esta situación podría justificarse en los casos en que los datos no contengan ninguna información textual (curvas de nivel, cotas altimétricas, etc.). Sin embargo, tolerar este caso podría dar problemas en un futuro dado que otros futuros catálogos podrían incluir la obligatoriedad del idioma como un requisito para la aceptación de un registro de metadatos.

4.2.b) Catálogo de la IDECLM.

El título y el resumen están presentes en la totalidad de los documentos de metadatos. En ninguno de los casos se encuentran descripciones numéricas ni repeticiones del título en el resumen, finalmente es uno el título que supera el largo de 110 caracteres.

Las palabras clave**,** igual que en los documentos de metadatos que distribuye la IDEC, se encuentran representadas en los tres tipos: lugar, tema y temporal. Destaca la presencia de palabras clave que albergan diferentes términos, lo cual puede generar confusión (Tablas 6 y 7).

PALABRAS CLAVE POR TIPOS (IDECLM)	total
Lugar	95
Tema	92
Temporal	11
Disciplina	0
Estrato	0
Total general	***198***

PALABRAS CLAVE MÁS UTILIZADAS	Tipo	total
callejero, urbano, municipio	tema	46
Castilla-La Mancha, Albacete, Barrax	lugar	25
Castilla-La Mancha, Albacete, Alatoz	lugar	18
ESPAÑA	lugar	9
Total general		***98***

Tabla 6. Volumen de las palabras clave por tipos (IDEC). **Tabla 7**. Palabras clave más usadas (IDEC).

Finalmente el idioma del dato está en blanco en 25 documentos (25.3%).

4.3. Razones de la presencia de errores en los metadatos.

Este trabajo también explora la diversa naturaleza de los orígenes de estos errores. Algunos errores o malas prácticas se deben al desconocimiento de la información exacta (como en el caso de la fecha de creación del dato); a la dificultad de determinar la información que se requiere bajo el esquema ISO 19115, (la escala, en informaciones tabulares de información asociada a coordenadas sobre el terreno) o la simple ignorancia de determinados factores (histórico de procesos).

Los proveedores disponen de diferentes métodos para crear y publicar documentos de metadatos en las IDE: formularios en línea, recopilación automática (*Harvesting*), transmisión directa del XML o transmisión directa desde un escritorio SIG [Goodchild 2007]. Los programas GeMM (de MiraMon) y CatMDEdit (de la Universidad de Zaragoza) permiten la extracción automática de metadatos, mientras que el MetaD posibilita la publicación directa de documentos de metadatos en el catálogo de la IDEC. Estos programas son los más utilizados en España. Sin embargo estos métodos no se encuentran exentos de la generación de errores. A pesar de que disponen de funciones de validación que controlan la obligatoriedad de los metadatos, se ha comprobado que no evitan la generación de errores por comisión o por falta de información.

Finalmente se ha comprobado que algunos de los errores presentes actualmente en los documentos de metadatos del catálogo de la IDEC no se pueden generar con el programa MetaD, CatMDEdit o GeMM, por tanto se reafirma la posibilidad de que se estén generando documentos de metadatos con otros programas, que carecen de una función de validación adaptada a las normas ISO 19115.

5. CONCLUSIONES.

El número de metadatos disponibles en una IDE corresponde directamente al número de proveedores con que cuenta la misma. Los proveedores de la información elaboran también sus metadatos por lo que los catálogos de las IDE gestionan esta información sin ser los responsables directos de la calidad de los documentos de metadatos que reciben. Sin embargo, sobre los metadatos de estos catálogos pueden realizarse controles de calidad que descubren errores o malas prácticas.

Los errores que albergan los documentos de metadatos se deben a tres aspectos fundamentales: la información de que disponga el proveedor, la capacidad del estándar para adaptarse a las necesidades del tipo de información a describir y las herramientas de que se dispone para generar los documentos de metadatos. Estos errores repercuten directamente en la calidad del metadato y por tanto en los resultados de búsqueda sobre catálogos de datos.

El análisis de los metadatos distribuidos por las 2 IDE autonómicas estudiadas manifiesta la presencia de errores de diferente naturaleza en estos documentos (Tabla 8). Aproximadamente un 5% de los documentos de la IDEC no presentan algún elemento obligatorio del estándar ISO 19115, en contraste a un 16.07% de los documentos de la IDECLM. El porcentaje medio de error para el conjunto de documentos se sitúa en torno al 7% en la IDEC y al 13% en la IDECLM.

Ambos catálogos de metadatos evidencian frecuentes errores o ausencias en la documentación de las fechas y los idiomas de los datos sobre los que las IDE estudiadas deberían incidir. Además la IDEMLC necesita averiguar el origen del abundante error en las coordenadas del ámbito de los conjuntos de información catalogados.

ERRORES DE LOS METADATOS	IDEC (%)	IDECLM (%)
Fecha de edición del metadato en blanco	2%	0%
Fechas del dato (las tres) en blanco	12%	37%
Fecha de creación del dato posterior a edición	10%	0%
Categorías temáticas en idioma incorrecto	11%	0%
Categorías temáticas en blanco	2%	3%
Información de contacto en blanco	0,1%	2%
Coordenadas no en ángulos	5%	60%
Mínima coordenada mayor a la máxima	3%	1%
Idioma de los datos en blanco	26%	26%
Idioma del metadato incorrecto	0,5%	3%
Factores de escala incoherentes	3%	9%
Promedio de error	***7%***	***13%***

Tabla 8. Resumen de los errores registrados en la IDEC y la IDECLM.

A pesar de la enorme diferencia en el volumen de datos, en la cantidad de productores y en el número de años que llevan funcionando se ha observado que los porcentajes de errores entre los metadatos de la IDEC y la IDECLM son similares por lo que podemos extrapolar que la necesidad de incidir sobre la calidad de los mismos es general.

Abundantes documentos de metadatos presentan errores que no pueden realizarse involuntariamente desde las herramientas de creación de documentos de metadatos más comunes (que poseen mecanismos de filtro y control), por lo que se evidencia que se están utilizando multitud de otras técnicas de generación. Es necesario disponer de un conjunto de reglas de validación homogéneo para todos los metadatos que sea aplicado tanto a las herramientas de edición como a los catálogos de metadatos.

Por otra parte, se observa una menor calidad de descripción en los elementos no obligatorios, aún siendo información relevante en datos geográficos. Con la finalidad de facilitar la tarea de crear documentos de metadatos, debe considerarse la utilización de programas de extracción automática de metadatos [Manso, 2004] como una herramienta que facilita, homogeniza y agiliza la creación de documentos de metadatos. De este modo el autor podría centrarse en describir más detalladamente los elementos adicionales, mejorando la comprensión del dato.

Este análisis aplicado a las IDE autonómicas, revela que la calidad es fruto de la necesidad de las IDE de buscar un compromiso entre la agilidad para los proveedores que crean metadatos y el usuario final que desea la máxima información y precisión posible. En este sentido la IDEC ha realizado una excelente labor de motivación, recolección e integración de los diversos actores implicados, lo que le ha permitido, alcanzar una masa crítica de información que posibilita reemplazar su objetivo inicial de cantidad y producción a un objetivo de mejora de la calidad, un proceso que ha iniciado en el presente año. Este estudio demuestra que este giro es necesario también en las IDE de más reciente creación como la IDECLM y que muchos de los errores pueden ser detectados con procedimientos de análisis de datos y eventualmente corregidos.

AGRADECIMIENTOS.

Los autores agradecen sinceramente a Víctor Pascual Ayats y a Jordi Guimet (IDEC) la información ofrecida a lo largo del desarrollo del estudio; y en general a quienes nos han ofrecido sus conocimientos y propuestas.

BIBLIOGRAFÍA.

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

FGDC, 1994. Federal Geographic Data Committee. Executive Order 12906: Coordinating geographic data acquisition and access: The National Spatial Data Infrastructure. April 11, 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, Infraestructura de Datos Espaciales de Cataluña. Manual del usuario de la aplicación MetaD para la creación, edición y exportación de metadatos (ISO 19115 – ISO 19139). Versión: 3.0.5. http://idec.icc.es/geoportal/cas/docs/.

INSPIRE 2007. Infraestructura de Información Espacial de la Unión Europea. Diario Oficial de la Unión Europea. Directiva 2007/2/CE del Parlamento Europeo y del Consejo de 14 de marzo de 2007 por la que se establece una infraestructura de información espacial en la Comunidad Europea (Inspire).

ISO/TC211, 2003; Geographic Information: Metadata, Internacional Standard 19115.

ISO/TS 19139, 2007; Geographic information: Metadata, XML schema implementation.

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.

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. La internet: http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf

OGC CSW, 2007. OpenGIS Catalogue Service Implementation Specification

TOLOSANA, R., NOGUERAS, J., ZARAZAGA, F.J. 2006. El impacto de la calidad de los metadatos en los servicios de búsqueda de una IDE. Jornadas técnicas de la Infraestructura de Datos Espaciales de España.