10

Metadata and spatial searching as key spatial information infrastructure component: future standardization developments

Marcel Reuvers (Geonovum) and Henri J.G.L. Aalders (TU Delft)

During the last years a lot of attention is drawn for developing Spatial Information Infrastructures (SIIs). Metadata is a major component for searching data by a SII. In this chapter the broader context of metadata is addressed and also open points for further work are mentioned. Because of the background of the authors this article has a standardisation viewpoint[1].

10.1 Why create metadata?

Any organisation providing information should disseminate their data in a way that non-specialised users can discover, evaluate and use it. The basic strategy is searching for words or phrases in the contents resources as a hit-or-miss strategy. When the spatial resource of interest has some word or phrase uniquely associated with it, this can be quite successful but mostly, hundreds or thousands of irrelevant 'hits' may be returned, as anyone can confirm who has spent frustrating hours searching for something whose name is a common word.

An alternative is to use metadata to describe resources in terms of certain well-defined attributes, such as resource title, geographic extent of the resource, resource topic category, or keywords. This allows users to search for keywords, names and phrases in particular contexts or structured search. This means that effective use of spatial metadata is based on three components:

- a set of commonly-understood terms that are used to describe the content of the information resources;
- a standard grammar for connecting those terms into meaningful metadata concepts, and
- a framework that allows the transfer and recombination of those metadata concepts across different applications and subjects.

Together these three elements provide architecture for spatial information description that can work across all associated subject areas (Aalders 2002). Therefore the mission for spatial data inventory applications is to make it easier to find resources using the Internet through the following functions:

- developing spatial metadata standards for information search and retrieval;
- defining frameworks for the interoperability of spatial metadata sets, and
- facilitating the development of community- and/or subject-specific metadata sets that work within these frameworks.

This may be done by providing user application software with tools as buttons, menus, or navigational structures on a website (or, in the ultimate situation) by providing free-text search capabilities.

---

[1] Material is re-used from the work that has been done by the INSPIRE Drafting Team Metadata (Reuvers is chair of INSPIRE Drafting Team Metadata) and Dutch architecture papers.

For example, an organisation’s name might be defined as a responsible party or as the distributor of a spatial dataset, in contrast to having no such information or being one of the many organisations that are described in a document linked to the spatial dataset. If this capability is combined with the use of ‘controlled vocabularies’ (i.e. standardised lists of terms, such as abbreviations for countries or code lists for categories) and standardised formats for values such as time, dates or longitude/latitude, it can greatly improve the efficiency of discovery.
Also, if all spatial resources are assigned metadata such as a resource topic category, it becomes much easier for a user to find resources that match a query for a specific topic.

From the perspective of a governmental organisation, it is important to help users obtain accurate and appropriate information: if users suffer some kind of loss as a result of finding incorrect or inappropriate information, they will make wrong decisions. In order to improve the discovery, evaluation and application of government information, the metadata created to describe resources at different websites and by different organisations must share a common form and meaning, so that users do not have to learn a different set of terms and search strategies for each site they visit. Such ‘interoperability’ is especially important for users who need to combine or compare information from multiple resources, but it is useful for any user attempting to discover information provided by government. This means that metadata standardisation is needed to get better results in discovery and understanding.

10.2 Standards based approach chosen for INSPIRE

Many factors encourage the adoption of standards. Very detailed information on this topic is provided in the GSDI Cookbook (GSDI Cookbook, 2004), but the following information is more particularly applied in the INSPIRE (Infrastructure for Spatial Information in Europe) context:

1. The importance of using a dedicated spatial metadata standard to support the implementation of a Global Spatial Data Infrastructure has been demonstrated by the different initiatives conducted since the early 90's particularly in:
   - North America with the development of the Content Standard for Geospatial Metadata by the US Federal Geographic Data Committee (FGDC) and the Presidential Executive order that all Federal government agencies were required to produce metadata for their spatial data holdings (ISO 2002);
   - Europe with the experimentation of ENV 12657 (CEN 2005) and more recently with the emergence of national and sub-national Spatial Data Infrastructures and their increasing adoption of ISO 19115. (ENV/Temporary European Norm, CEN/Comité Européenne de Normalisation, ISO/International Organisation for Standardisation)
2. More than the lack of metadata, the lack of compatibility between the existing and upcoming metadata solutions is certainly one of the greatest challenges of the INSPIRE Directive (EU 2007, European Union). At this stage and due to the importance of the community, the use of a standard lexicon is a key to success.
3. The standardisation activity in this geo-world has reached level of maturity. A metadata standard dedicated to Geographic Information is available with the publication of ISO 19115. Its applicability to the European context was established with its adoption by CEN in 2005. The reference materials provided by the INSPIRE community for the establishment of these Implementation Rules (IRs) show a general

> endorsement of this international standard by the different European actors of the geographic information domain. Most of the legally mandated organisations and the Spatial Data Information Community (SDIC) have already adopted ISO 19115 or have on-going activities to adopt it as indicated by the results of the INSPIRE survey conducted in Spring 2006 (INSPIRE 2006).

Access to on-line metadata repository is fostered by standard interface specifications, such as the OGC CSW 2.0 (Geospatial Consortium Catalog Services for the Web) which can accommodate the use of different abstract metadata standards and related encoding, such as:

- ISO 19115 and ISO 19119 through their ISO TS 19139 XML Schema implementation (CSW2 AP ISO);
- Dublin Core and its XML Implementation, which are relevant to ensure the relationship with other communities (XML/eXtensible Markup Language).

One should realize that multilingualism becomes a very important aspect since the Web connects a diversity of linguistic and cultural aspects from all over the world. So, the Web will fail to achieve its potential as a global information system, unless resources can be made available to users in their native language, in appropriate character sets and with metadata appropriate to resource management. Here *internationalization* and *localisation* become apparent: though they may be contradictory: while global resource discovery is best served by internationalisation (using common conventions of practice, languages and character sets throughout the world), the needs of any given community may be better served by supporting local conventions. Basic to the promotion of a global metadata architecture is to translate relevant specification and standard documents into a variety of languages. DCMI (Dublin Core Matadata Initiative) maintains a list of translations of its basic documents, as the European workshop on Learning technologies is maintaining translations of the LOM (Learning Object Metadata) specifications (Moellering 2005).

The Dublin Core metadata element set (or the basic interlocking brick) is intended to support cross-subject search and retrieval. It can be thought of as a simplistic or pidgin metadata language that helps the user navigate through disparate subjects, languages, and cultures. Adoption of the Dublin Core by governments, libraries, museums, archives, publishers, environmental science repositories, prints and e-print archives, to name a few, testifies to its success in this role. There are emerging applications in the commercial sector, as well, with health care organisations and financial industries using the Dublin Core as the basis for organising and exchanging information.

10.3 Metadata in a few words

ISO 19115 defines metadata as 'data about data' [1]. This basic definition implies an unlimited scope to what can be seen as metadata. It allows some experts to see information as data or metadata with an unrealistic border between both, and also including data services in metadata. The INSPIRE Directive clarifies the definition of metadata as information describing spatial resources, making it possible to discover, inventory, and use them.

---

[1] INSPIRE has changed this definition in 'Information describing spatial resources, making it possible to discover, inventory and use them'. This definition of metadata originates from the directive. It is compatible with the general definition of metadata provided in ISO 19115 and the OGC abstract specification for metadata: 'data about data'. It clarifies the expected role of metadata within the INSPIRE Infrastructure.

The metadata for those resources comprise:

- Identification information, i.e. information to uniquely identify the resource such as:
    - Title, abstract, reference dates, version, purpose, responsible parties, …
    - Geographic extent,
    - Browse graphics (overview, thumbnail, …),
    - Possible usage;
- Legal and security constraints;
- Content Description, i.e. information identifying the feature catalogue(s) used and/or information about the coverage content;
- Reference system information, i.e. identification of the spatial and temporal system(s) used in the resource data;
- Spatial Representation, i.e. information concerning the mechanisms used to represent spatially the resource data;
- Quality and validity information, i.e. a general assessment of the quality of the resource data including:
    - quality measures related to the geometric, temporal and semantic accuracy, the completeness or the logical consistency of the data;
    - lineage information including the description of the sources and processes applied to the sources;
    - validity information related to the range of space and time pertinent to the data; to whether the data has been checked to a measurement or performance standard or to what extent the data is fit for purpose.
- Portrayal information, i.e. information identifying the portrayal catalogue used;

- Distribution information, i.e. information about the distributor of, and options for obtaining the resource;
- Maintenance information, i.e. information about the scope and frequency of updating of the resource data.

The INSPIRE Metadata Use Case (see Figure 10.1) creates a general context federating the existing and upcoming metadata based solutions around three base user activities[1] (see Figure 10.2):

1. The *discovery* of resources. The user expects to identify a set of resources satisfying a basic set of search criteria. The user interacts with a Search Engine connected to a set of metadata repositories which document available resources. The Search Engine transfers the search criteria to the metadata repository and expects a minimum set of metadata related to the matching resources. It consolidates the answers and provides them to the user through an adapted interface.
2. The *evaluation* of available resources. The user has now identified a candidate resource, potentially as a result of the discovery activity and wants to determine whether this resource satisfies his/her requirements or not. For this purpose it may use a Metadata Browser to examine more detailed metadata about the resource.
3. The *use* of adequate resources. The user has chosen a resource and some access and use rights have been granted to him/her. The resource is accessible and can be used through a

[1] Some INSPIRE terms differ from the ones in the GSDI cookbook, e.g. ‘evaluation’, ‘use’ (INSPIRE) versus ‘exploration’, ‘exploitation’ (GSDI cookbook).

series of dedicated tools. Metadata will support the user in fully understanding the data and using it properly resulting in more reliable analysis and more confidence in the results.

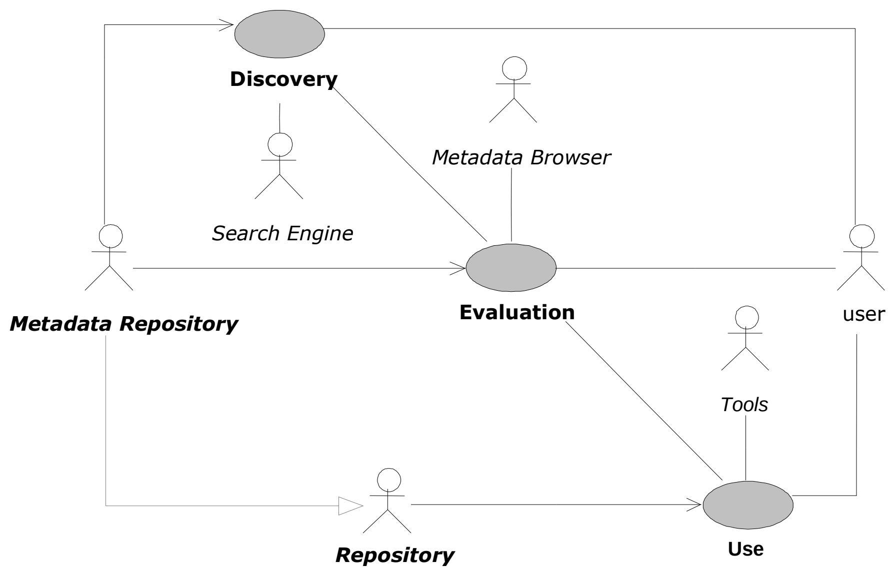

Figure 10.1 The INSPIRE Metadata Use Case

Different kinds of users may be involved in the different activities: some experts may evaluate the resources while operators may use them. There are cases where the user may be a software program performing automating searches. Depending on the users, different types of Search Engines, Metadata Browsers and Tools may be necessary. From this perspective, this use case does not constraint the software market and it even creates a general context fostering the emergence of new markets and the satisfaction of new user requirements.

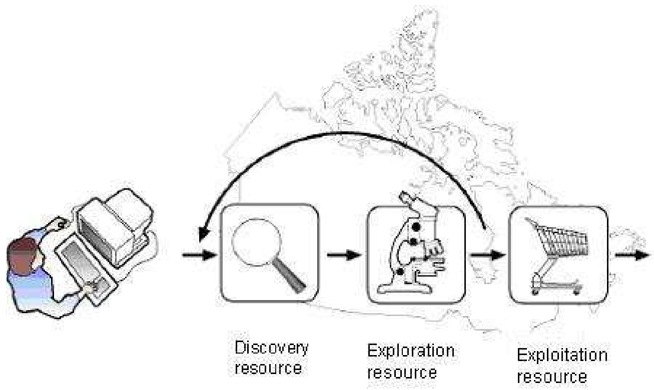

Figure 10.2 The three steps: discovery, exploration and exploitation of resources

10.4 The Metadata implementation

The metadata production within organisations has recently become professional. There are a lot of initiatives where metadata is built-up. The way this is done has to be critically reviewed. An example is that the metadata production is done after the data production process. Consequences are that the metadata is difficult to catch and that means that quality and completeness is not guaranteed. The integration of the metadata production within the data production[1] causes problems explained by the following reasons (Wayne 2005), but is the way to go:

1. metadata standards are too extensive and difficult to implement
2. metadata production requires time and other resources
3. there are few tangible benefits and incentives to produce metadata.

The achievement is to solve these three problems. Below are mentioned organisational benefits (Wayne 2005)

Organisational benefits of metadata	
Data Archive	Data are the most expensive components of a GIS. Metadata is a means of preserving the value of data investments. This is of particular significance to local and regional governments experiencing rapid staff changes.
Data Assessment	GIS data development has shifted form data producers to data consumers. From a consumer perspective, metadata is the truth in labelling required to assess available data products. From the producer's perspective, metadata is a means of declaring data limitations and serves as a form of liability insurance.
Data Management	Metadata enables organizations to retrieve in-house data resources by specific criteria for global edits and annual updates.
Data Discovery	Metadata is the primary means of locating available spatial data resources via the Internet. Metadata is a primary public information resource as it is a non-technical means of presenting technical information.
Data Transfer	Metadata is increasingly used by software systems as a means of properly ingesting data and by analysts as a means of properly displaying data.
Data Distribution	By building metadata in compliance with national standards, one can

[1] At the end the distinction between metadata and data will disappear in the production process. Advanced organisations are thinking in processes and information (models) and not in metadata and data.

	participate in the Global Spatial Data Clearinghouse. Participation promotes your agency and frees staff from answering data inquiries.

10.5 The importance of temporal extent

Besides the spatial extent the temporal extent is worthy mentioning because of the importance for temporal queries. Temporal aspect is also important in several domains, e.g.: earth science (see chapter 5), cadastral registration (see chapter 9) and (Oosterom 2006, Oosterom 2002)

Some general user queries on temporal extent might be:

- A travel company wishes to search for the January snow climatology for the Alps, for inclusion in a brochure. The full temporal coverage might be the summary statistics of snowfall and snow cover for the period 1971 to 2000 taken over all Januaries. Here the discovery metadata is specific for area and for temporal extent. The classification of temporal extent is and ordinal reference system of named months.
- A hydrologist is looking for winter rainfall climatology in a catchment area. Here the ordinal classification of winter may be more important than a specific geographical position, as the hydrologist will expect to generalize to the catchment from whichever specific locations are found.
- A geologist may wish to find surface beds of Tertiary minerals across Western Europe. Again the ordinal temporal classification is initially more important than location.
- A marine investigator requires weather information for a specific period for an accident at sea, the location being only approximately known.
- An archaeologist wishes to compare sites across regions, which are known to be active in

the same time period. The discovery is to find locations for a known period.

- Anyone who wants a weather forecast for 3 days hence. For a weather forecast, the sequence of weather is usually more important and may be more accurately forecast than the expected weather at a specific location and time.

Temporal aspects vary quite significantly with respect to the application domain. The semantics of the temporal extent will be in development for the next years. The domain standardisation will be leading here. From our experience INSPIRE will be an important catalyser to achieve this.

10.6 The changed role of metadata in Service Oriented Architecture (SOA)

Any organisation wanting a SOA application to install has to migrate its application architecture. Figure 10.3 describes a general application architecture for an individual organisation.

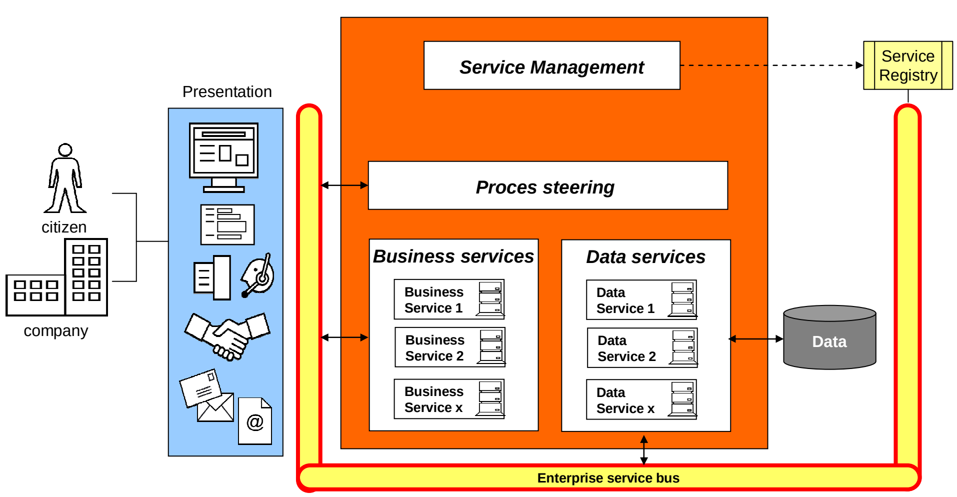

Figure 10.3 The Dutch Reference model service layers

An application architecture brings the functionality of applications in logical layers with standardised interfaces. This creates more flexibility and better possibilities for maintenance because changes in a logical layer do not have consequences for the other layers, as long as the interfaces between layers do not change. Within the Dutch Reference model for Architecture the next layers can be distinguished:

- Data;
- Data services;
- Business services;
- Presentation;
- Process steering;
- Service management.

The layer service bus is not worked out in more detail because its task is mainly logistic (way of delivery, envelope). In the Netherlands one applies a basic bus for interfaces, security and a richer bus with possibilities for schema translation, orchestration, subscriptions and so on.

Data

The data is all digital information that is maintained in databases, file systems, document management systems, geo-information systems, etc.

Data services

The data services give access to data. With these services data can be created, changed, deleted or viewed. Data services are mostly offered by a provider without knowing what the user wants, where a user can be a human or an application. A user has to change his environment in the way the data is served. In the geo-world examples of these services are web mapping services (WMS) of web feature services (WFS).

The metadata described in the previous sections is metadata over data services, making use of ISO 19115, ISO 19119, ISO TS 19139 and OGC capabilities. This is the 'classic' way of describing and using metadata.

Business services

Business services supplies services to users (humans or applications). A characteristic for business service is that the user knows beforehand which performance will be delivered but not how this is done. A business service fulfils that delivered information is focussed on the process in which it is used.

These business services may use data from the data services. A business service can make use of WFS Filter encoding but more regular and needed is the application of SOAP (Simple Object Access Protocol) and WSDL (Web Services Description Language) as a service description.

The metadata for these services is not focussed on the data but on fulfilling of the business needed. That makes the description of these services different from the 'classic' way.

Presentation

The presentation layer takes care for the interaction with the user. Information for applications services is presented in a way it's fits the used channel.

Process steering

These services make use of other services. By each used service agreements are needed about the way of interaction with the user. This has consequences for the implemented process within organisations. Process steering takes care that the needed business services are chained in the right order. This is about orchestration, choreograph, business process management and workflow management. WS-BPEL (Web Services - Business Process Execution Language) is an often applied schema language in these processes.

The metadata has to support these processes. Questions arise like how is the metadata described for a new service as an outcome of a WS-BPEL process and what the role is of the metadata of the individual business services that are part of it?

Service management

An organisation has to take care that services will fulfil in time the requirements of the users. That means creating of new services and removing or changing of existing services . For several reasons it is important to describe and publish the services in a registry. In this way the provider will has to register the services only once instead of many times. Examples of these service registries are UDDI (Universal Description, Discovery and Integration protocol), ebXML (Electronic Business using eXtensible Markup Language), RIM (Registry Information Model) and CSW (Catalogue Service). These registries fulfil also other requirements: the metadata in these different kind of registries is also different on aspects based on the purpose of the service itself. A good investigation of the differences between these registries is done by ISO TC/211 in the ad hoc group on the study of ebXML RIM, 2007-02-16, http://www.isotc211.org/protdoc/211n2165/.

In the geo-world we can distinguish three types of registries:

1. Discovery services

   Dataset metadata: basic information about a dataset, e.g. about the identification, the extent, the quality, the feature types, spatial reference, and distribution.

2. Publishing services

   - Service descriptions: basic information about a service, e.g. a description of the operations and their parameters as well as information about the geographic information available from a service offering.
   - Service types: list of service types (service taxonomy).
   - Data specifications: detailed description of one or more datasets that will enable it to be created, supplied to and used by another party.

3. Register services

   The number of registers that need to be maintained in the infrastructure can be significant. As a result, a clear and sustainable operational model forms a key part of the setup of the infrastructure. A starting list of potential registers, i.e. kinds of spatial information items, includes:

   - Feature catalogues: catalogues containing definitions and descriptions of the feature types, their attributes and associated components occurring in one or more data sets, together with any operation that may be applied as part of a data specification.
   - Application schemas: conceptual schema for data required by one or more applications. Part of a data specification and specified in a formal conceptual schema language (typically UML (Unified Modelling Language).
   - Code lists: dictionary describing the attribute value domains for selected property types in

a feature catalogue / application schema. In these cases the value domain is not fixed in the feature catalogue / application schema, but that is managed separately; i.e. this establishes a controlled vocabulary.

- Thesauri: similar to code lists with additional information how terms of the vocabulary relate to each other (hierarchies, etc.). It is unclear whether this is needed in the first step.
- Coordinate reference systems: dictionary of coordinate reference systems, datums, coordinate systems and coordinate operations which are used in datasets.
- Units of measurements: dictionary of units of measurement, which are used in datasets.
- Spatial object identifier namespaces: a mechanism is required to guarantee uniqueness of feature identifiers across various content providers. One approach is to use existing 'local' identifiers of the provider, but define namespaces to distinguish between different providers (and between different offerings of a provider). These namespaces need to be managed.
- Portrayal rules: rules that are applied to a feature to determine the portrayal of a feature in a map.
- Symbols: depictions to be used in portrayal rules to describe the styling of features in a map.

As we can see many services will exists; the context is much broader then only a registry for metadata in the 'classic' way. Besides the discussion which registries are needed on the software, the interesting part also will be which metadata will be needed to fulfill the user requirements. It is expected that more and more metadata will be interpreted by machines instead of a human interpretation.

10.7 The Geospatial Web

The Geospatial web is growing fast: most significant platforms are Google Earth, Microsoft Virtual Earth and NASA (National Aeronautics and Space Administration) World Wind. Behind this there is new technology growing like geo-tagging, semantic web, open source, open data and much more mostly driven under the general Web 2.0 developments.

The Geospatial web, How geo-browsers, Social software and the Web 2.0 are shaping the Network Society, is a excellent book in describing new developments regarding the web.

The metadata way of Web 2.0 thinking comes from another point of view as the current metadata thinking. The Metadata way of Web 2.0 thinking concentrates on for example:

- Conceptual search by ontology's (making use of OWL (Web Ontology Language)), Chapter 5, Conceptual Search: Incorporating Geospatial Data into Semantic Queries and Chapter 23, SWING – A Semantic Framework for Geospatial Services)
- Making use of communities for tagging geo-data and building metadata
- Creating geo-tagged pictures by address (geographic extent), Chapter 15, Sharing, Discovering and Browsing Geo-tagged Pictures on the World Wide Web

… and much more. This means that the processes of building metadata can be enriched by making use of the Web 2.0 developments. Otherwise it is important to align the current metadata developments with the Web 2.0 developments; especially when the geo-community wants to discover not only geographic data but also new data (as pictures and others sensors) that have a location too.

The Geospatial Web will have impact on the current metadata in building the metadata and conceptual search. New types of metadata for different kind of data will arise. Therefore it is important to align continuously the current metadata developments with Web 2.0 developments.

Too arrange this, sustainable standards are needed and the work of ISO TC/211 and OGC with the semantic web world of W3C needs programming and more detailed cooperation. For example, the hierarchical structure of ISO 19115 makes this alignment more difficult.

10.8 General conclusion

Metadata will support the user in fully understanding the data. It appears that metadata for data and data services are well arranged in the field of standardisation. At the other hand we cannot sit back because new challenges come up and the traditional background of data (maps) focussing will not help us. These new challenges are pointed out in the sections: The importance of temporal extent, the changed role of metadata in Service Oriented Architecture (SOA) and the Geospatial Web. Of course there are and will be other important developments but these will come from different directions and from different communities. This means that the next step in metadata development will require more and more cooperation with other domains. Good steering of this process: *do not worry us on the outcome!*

Literature

Aalders H.J.G.L. and G. J. Hunter, 2002, *Spatial Data Standards* UNESCO - EOLSS Encyclopedia Knowledge for sustainable development - An Insight into the Encyclopedia of Life Support Systems, Paragraph 6.72.3.4. August 2002

CEN 2005, *EN ISO 19115 Geographic information – Metadata*, EN European Norm

ENV 12657: 1998, Geographic Information - Data description - Metadata

INSPIRE 2006 Drafting Team Metadata, Draft Implementing Rule Metadata as provided to the

SDICs and LMOs , see http://www.ec-gis.org/inspire/reports/ImplementingRules/draftINSPIREMetadataIRv2_20070202.pdf

EU 2007, *Directive of the European Parliament and of the Council establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)* January 2007

INSPIRE 2006, see *http://www.ec-gis.org/inspire/reports/INSPIRE_Metadata_Survey_2006_final.pdf*

ISO 2002, *ISO 19115: 2002 Geographic information - Metadata*

ISO 2003, *ISO 19119 Geograohic Information - Services*

ISO 2004, *ISO TS 19139 Geographic Information - Metadata - XML schema implementation*

Moellering 2005, H. Moellering, H.J.G.L. Aalders and A. Crane, *World Spatial Metadata Standards.* Published on behalf of the International Cartographic Association by Elsevier, Amsterdam, The Netherlands. 2005. ISBN 0-08-043949-7.

van Oosterom, P.J.M. and C.H.J. Lemmen, 2006, *'Spatial data management on a very large cadastral database'*. In: *Computers, Environment and Urban Systems,* Volume 25, 4/5, pp. 509-528. V30 (1), 2006

van Oosterom, P.J.M., B. Maessen and C.W. Quak, 2006, Generic query tool for spatio-temporal data. In: *International Journal of Geographical Information Science*, Volume 16, 8, pp. 713-748 December 2002

Scharl, A., Tochtermann, K. The Geospatial Web, How Geobrowsers, 2004 Social Software and the Web 2.0 are Shaping the Network Society, *Springer*, ISBN 978-1-84628-826-5 GSDI 2004

Onsrud, H. *GSDI Cookbook v2.0* (PDF) January 2004.

Wayne 2005, *see http://www.fgdc.gov/metadata/documents/InstitutionalizeMeta_Nov2005.doc*