Metadata: where we are now, and where we should be going

Massimo Craglia, Ioannis Kanellopoulos, and Paul Smits
European Commission- DG Joint Research Centre
massimo.craglia@jrc.it, ioannis.kanellopoulos@jrc.it, paul.smits@jrc.it

INTRODUCTION

The purpose of this paper is to review the current state of the art in metadata developments, critically analyse both opportunities and limitations, and suggest new directions for future work in this field. The perspective is primarily European and particularly focused on the geographic domain, although wider issues are discussed where relevant. The main thrust of the paper is that aside from the opportunities provided by technological developments in the ICT field, there are major societal, economic, and legislative drivers pushing for greater transparency of, and access to, information particularly in the public sector. These drivers include legislation on Freedom of Information, Reuse of Public Sector Information (PSI), and more specifically to the environmental and geographic sector, legislation on access to environmental information, and the new directive setting up an infrastructure for spatial information in Europe (INSPIRE).

It is the pressure to open up the stores of PSI, and environmental./geographic information for access and use to others than those who collected in the first place that is increasing the visibility of metadata, i.e. the information necessary to discover what information resources exist, who has them and what are the conditions to access and use them. Many international initiatives are therefore converging to provide standards, tools, and technologies to create and manage metadata. Despite this apparent progress however, a number of barriers remain which include the organizational and financial cost of creating and managing metadata, the immaturity of some of the standards, and specifications, and the uncertainty created by constant technological change that makes the case for metadata investment less clear cut. If these are some of the challenges in respect to metadata for data sets, even greater ones are currently being faced by metadata for web services which need to be overcome to allow for (semi) automatic search, retrieval, choosing, and chaining of services to process information resources necessary to respond to a problem.

With these considerations in mind, the paper is structured as follows: the next section introduces some of the key concepts of metadata including different levels of metadata, granularity, and perspectives. The importance of metadata in the context of the development of spatial data infrastructures is also discussed in this section. Section Three reviews the key international standardization initiatives that provide the framework for creating and managing metadata at the present time. A particular sub-section is devoted to the effort of the INSPIRE Drafting Team on Metadata to develop detailed implementing rules able to provide for a coherent and interoperable implementation of the infrastructure for spatial information in Europe. Section Four discusses ways in which metadata once created can be searched and managed, including harvesting methods and catalogues. The instability of current standards and specifications for cross-catalogue searching are discussed here based on a project recently funded by the JRC (Senkler et al. 2006). Section Five discusses the issues raised in the previous sections and concludes with a call for a shift from the current data producer- centric view of SDIs in general and metadata in particular, to one that is more user focused. From an SDI development point of view, this needs a much greater emphasis on service-based processing, and hence on service metadata, while for a specific metadata perspective, the move towards a user focus needs new mechanisms to harvest user feedback, engage users participation in the development and maintenance of information resources following the example of social networks and communities currently being developed in the Internet world.

.

METADATA DEFINITIONS AND PERSPECTIVES

Metadata is structured information describing an information resource, such as a written document or report, a table of statistical data, a topographic dataset, an image, or anything that one may want to find and utilize, including a service able to process data in some way rather than data itself. Metadata may have different levels of "granularity" i.e. refer to a collection of (digital) objects or datasets, individual objects within the collection, or parts of an object. So for example you may have metadata for a topographic dataset (e.g. the 1:50,000 scale map of Great Britain), then metadata for individual tiles as they are updated at different times, and finally metadata for individual features or records in the database. These different levels of granularity are not mutually exclusive, and users from different communities and areas of interest may pay particular attention to one or more of such levels. There are also different levels of metadata, including:

- Metadata for Discovery: the minimum amount of information that needs to be provided to convey to the inquirer the nature and content of the available resources; this is information that basically answers questions about what resources exist, where, and held by whom.
- Metadata for Evaluation: adds detail to the level above to allow potential users of the resource to assess whether it would be suitable for their purposes. This may include information on the characteristics of the resource, but also on the financial and legal conditions for access and re-use.
- Metadata for Use: includes further details enabling access, transfer, interpretation, and use of a resource in an application.

How these different levels of metadata are interpreted may well vary depending on the perspectives: data producers may see metadata as a mechanism for advertising their products, and educate potential users on the characteristics of the data. In this sense, quality is often described through objective parameters of the production process. On the other hand, users maybe less interested in such technical details but be more interested in minimizing the costs of getting the data (time, money, procedures), as well as details of what could be done with the data, applications possible, or details of previous users or experts in the field. This social dimension of metadata is borne out of focus groups with a range of different user groups in an EC-funded project in the late 1990s[1], but is largely ignored by the dominant view of metadata that is data-producer driven and enshrined in emerging standards (see Section 3).

The concept of metadata is not new, as librarians for example have been documenting and cataloguing information resources (both physical like books, and digital) for a very long time. Nevertheless, its importance has been steadily growing with the emergence of Internet-based services providing access to government information (e-government), and other dedicated initiatives focused on geographically-referenced infrastructures (Spatial Data Infrastructures or SDI). In Europe, the push towards the development of an Information Society (Craglia and Masser, 2003; Blakemore and Craglia, 2006) and legislation promoting access to environmental information (CEC, 2003a) and the re-use of Public Sector Information (CEC 2003b) have also fostered indirectly the importance of metadata by requiring Member States to develop "registers or lists of the environmental information held by public authorities or information points, with clear indications of where such information can be found" (CEC 2003a, art 3 para 5c), and "practical arrangements [..] that facilitate the search for documents available for reuse, such as assets lists, accessible preferably online, of main documents, and portal sites that are linked to decentralized assets lists" (CEC, 2003b, Art. 9). A step-change in the importance of metadata is provided by the recently approved INSPIRE Directive establishing an infrastructure for spatial information in Europe (http://www.ec-gis.org/inspire) which specifically

[1] Methods for Access to Data and Metadata in Europe:
http://www.shef.ac.uk/~scgisa/MADAMENew/Content.htm

requires Member States to “ensure that metadata are created for the spatial data sets and services corresponding to the themes listed in Annexes I, II and III, and that those metadata are kept up to date” (Art 5), and specifies some of the information that must be included in the metadata, as well as the fields that must be searchable through Discovery Services to be provided by the Member States (Art 11). This Directive is very significant because moves the discussion from metadata as a good data management practice (that may or may not be endorsed) to the level of obligation to create metadata conforming to a minimum level of service for a wide range of data themes specified in the Directive. It is no longer a question of “if” and “when” but only of “how” to do it. In this respect, detailed technical rules for metadata are being developed by an international team of experts with the support of the Commission. Once approved, these technical implementing rules will be mandatory to ensure a coherent implementation of the Directive. These rules and the international standards within which they are framed are discussed in the next Section.

PLENTY OF STANDARDS

If there is a lack of metadata, it is certainly not due to a lack of standards. As the number, complexity, and diversity of geographic datasets have grown, methods for providing an understanding of all aspects of this data grew in importance as well. In the last decades, various initiatives were launched to standardize the way in which information about geographic datasets was presented. Ten years ago a pre-standard on GI metadata was published by the European Committee on Standardization (CEN) Technical Committee 287. This expertise and that from other stakeholders was funneled into the International Standardization Organization’s (ISO), Technical Committee 211 project, resulting in ISO 19115:2003 Geographic Information - Metadata. When implemented by a data producer, ISO 19115 will:

1) Provide data producers with appropriate information to characterize their geographic data properly.
2) Facilitate the organization and management of metadata for geographic data.
3) Enable users to apply geographic data in the most efficient way by knowing its basic characteristics.
4) Facilitate data discovery, retrieval and reuse. Users will be better able to locate, access, evaluate, purchase and utilize geographic data.
5) Enable users to determine whether geographic data in a holding will be of use to them.

So the standard focuses on the content and structure, and not on the encoding. In fact, in June 2001 the Open Geospatial Consortium (OGC) embraced ISO 19115 as an abstract specification (i.e., it has semantic value rather than syntactic). CEN/TC 287 Geographic Information , and therefore all the European national standards bodies, have also 19115 in their catalogue (EN ISO 19115:2005). As a consequence, a number of countries have also translations of the standard in their national language which greatly helps its uptake.

ISO/TC 211 follows a model based approach to the standards it develops, and there is a great inter-dependency between the 19100 series of standards it produces. ISO 19115 is based on a set of foundation standards from the ISO 19100 series, e.g. ISO/TS 19103, ISO 19107 and ISO 19108. ISO 19119 Geographic Information - Services extends ISO 19115 for metadata for spatial services. The applicable XML Schema Implementation of ISO 19115 is defined in ISO/TS 19139.

Outside the geographical domain, the Dublin Core (DC) metadata elements set (aka ISO 15836) gained wide acceptance in the communities that deal with the more general information sources such as e-Government. Its wide acceptance is perhaps also related with the fact that DC is simple.

The INSPIRE Metadata Drafting Team has been tasked to develop a set of metadata elements to be used in connection with the INSPIRE Directive. The Directive itself says that Implementing Rules be developed in accordance to relevant European and international standards and best practices. In addition, a survey was carried out by the JRC among the INSPIRE stakeholders (Nowak and Craglia, 2006), which provided further evidence that the standards-based approach adopted by the Drafting Team was the right way to go. The proposed INSPIRE metadata elements are tailored to the discovery of geospatial resources, taking into account the requirements of the Directive. In particular, the Directive (EU, 2007) requires for metadata to include:

(a) the conformity of spatial data sets with the implementing rules on harmonization;

(b) conditions applying to access to, and use of, spatial data sets and services

(c) the quality and validity of spatial data sets;

(d) the public authorities responsible for the establishment, management, maintenance and distribution of spatial data sets and services;

(e) limitations on public access and the reasons for such limitations.

In addition, the Directive requires Member States to deploy Discovery Services to search and display the content of the metadata. Searching should be available on

(a) keywords;

(b) classification of spatial data and services;

(c) the quality and validity of spatial data sets;

(d) degree of conformity with the implementing rules on harmonization;

(e) geographical location;

(f) conditions applying to the access to and use of spatial data sets and services;

(g) the public authorities responsible for the establishment, management, maintenance and distribution of spatial data sets and services.

Most of these criteria are uncontroversial but some discussion is needed in the review process to identify the appropriate way to describe the quality and validity of spatial data sets which reflect not just the perspective of data producers but also the assessment by users in relation to the fitness for purpose of the resource.

SEARCHING FOR METADATA

So we have standards for metadata and a legal framework in the making which will push for the creation of metadata for a large number of resources. However, metadata are only useful if they can be searched for and discovered. This usually happens through metadata catalogue services.

Catalogues of geographic resources are one of the core components of a Spatial Data Infrastructure. Geographic data catalogues are discovery and access systems that use metadata as the target for query on geographic information. In addition to catalogues that contain metadata about geospatial data, there are also catalogues that describe geographic services. For the purposes of this paper we will refer to the geographic data and services catalogues as catalogues of geographic resources.

Catalogues have three essential purposes:

- To assist in the organization and management of diverse geospatial resources for discovery and access,

- To discover resource information from diverse sources and gather it into a single, searchable location, and
- To provide a means of locating, retrieving and storing the resources indexed by the catalogue.

A recent study funded by JRC (Senkler et al 2006) examined the current state of the art in catalogue services and the relevant International standards. The overall objective was to set up and test an environment where a centralized catalogue service is integrated into the INSPIRE geoportal, which is required by the Directive, to provide harmonized and interoperable access to federated catalogues throughout Europe.

The distributed catalogues provided different service interfaces and different information models for their metadata; the goal was to realize the best degree of interoperability in the catalogue service network and to report shortcomings and advantages of up to date software, implementations and specifications.

The main requirements in the study were catalogue services compliant to the OGC CSW 2.0 specification and metadata compliant with the ISO 19115:2003 standard.

Many catalogues from different European member states contributed their catalogue implementations and knowledge to the study. Specific interoperability tests were developed to test how the considered catalogue services supported the underlying specification.

The interoperability tests demonstrated that specifications considered in the study are not robust and are not adequately supported by the implementations: none of the distributed catalogue services could be queried by the centralized catalogue service broker without the development of special adaptors. An adaptor is a filter that is plugged between the broker and the target catalogue service to translate the request to the federated catalogue service and the response back to the broker in a way that a standardized communication could be established. A number of adaptors were therefore developed to enable access to the distributed catalogues and deal with the implementation specific inconsistencies of the considered catalogue services. In general the following issues were identified:

- Many aspects of the OGC CSW 2.0 specification are ambiguously defined. This leads to different interpretations of the specification and results to non interoperable catalogue service implementations.
- Concerning the underlying information models (ebRIM and ISO AP), it is erroneous to translate one model to the other (semantically and syntactically) since no standardized mapping rules exist.
- The concept of federated search should be better documented and integrated into the specifications: i.e. what it means, how it works, etc.

The reasons for this are, in most of the cases, too many degrees of freedom in the underlying OGC CSW 2.0 specification. This leads to different interpretations of the specification and, in the end, in non interoperable catalogue service implementations. The results of this test have been fed back to OGC because it is clearly unsustainable to have to develop specific adaptors for each of the catalogues being searched.

Alternative methods to access metadata from different locations include harvesting. Harvesting refers to accessing metadata resources through a web server and adding them to a local database. This takes place at predefined time intervals and no specific catalogue service is required. Harvesting is efficient, as the success of Google demonstrates, and also allows for richer statistics to be returned to the user including ranking of frequencies, which are more difficult to do in the case of distributed catalogue searches. The obvious disadvantage of harvesting is that multiple copies of the metadata are stored at different locations but at this stage of development of SDIs, it may well be that both centralized and distributed searches need to be supported.

DISCUSSION AND CONCLUSIONS

The INSPIRE Directive represents a significant step change in the development of SDIs in Europe, and worldwide. At the European level, it mandates the creation and maintenance of metadata and related discovery services which are the first visible and value-adding element of any SDI. Its significance worldwide is not just as a component of a nascent global SDI, but also because it will be the first large scale implementation of federated SDIs across 27 countries and 23 languages. This challenge already exposes some of the limitations of current standards implementations and specifications which need to be overcome to achieve interoperability. In prospect, Europe becomes then a leading laboratory to research and test multilingual, multicultural, and disciplinary SDIs, where the complexity of semantics reference systems already poses several challenges (see for example Kuhn, 2003 and 2005). More research (and implementation) challenges are to follow as we move from the current generation of SDIs, largely focused on data search and retrieval for local processing by expert users, towards a much more open generation of SDIs aimed also at the non-specialist user and the general public, in which the emphasis will not be on data per se but on information i.e. the outcome of a process or workflow that interprets the question posed, finds the relevant data and processing services, chains the services, and return an answer understandable by the user, and hence tailored to his/her profile. Such Service Driven Infrastructure will require much more machine-to-machine interaction and therefore the embedding in machine-readable form of much of the knowledge currently used by humans to search data, understand their structure and semantic meanings, find services that are appropriate, make choices when needed between alternative offerings of data and services, and at the current state of the art, even chain services manually. Hopefully, in this next generation SDI, it will not even be necessary to manually create metadata but it will be automatically extracted during the production and usage process of the data (see for example Bulterman, 2004; Gould, 2005; Howison and Goodrom 2004). Similarly, there is a need to move much beyond the current descriptions of services which are largely for human consumption with machine-readable ones that encode a much richer description of what the service can do, what data it can process, how trustworthy its outcomes are, and how to resolve potentially conflicting Digital Rights Management rules pertaining to the different data sets (and potentially also services) used in the chain. If we consider that the current descriptions of “quality” of a service refer only on its response time and reliability meaning how often the service is down, then one can see how far we have to go in this field. Until such pressing research issues are solved, we are still operating in the index-card paradigm of librarians pre-dating the digital information revolution, and in the artisan’s world of service chaining. With INSPIRE we are giving an extra boost, but maybe it is really time to think of another way of documenting resources to include active feedback from users as well as more automated means of clustering user preferences and searching, mining association rules, and deploying the results for the benefits of users as well as producers that are now standard practice among on-line retailers (e.g. Amazon) but have yet to make it to the geospatial world (see for example Pike and Gahegan, 2004).

References

Blakemore, M. and Craglia, M. 2006. Access to Public Sector Information in Europe: Policy, Rights, and Obligations. The Information Society 22(1): 13-24

Bulterman D. (2004) Is it Time for a Moratorium on Metadata? IEEE Multimedia, October-December, 10-17.

Commission of the European Communities. 2003a. Directive 2003/4/EC of the European Parliament and the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC. Brussels: CEC.

Commission of the European Communities. 2003b. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Brussels: CEC.

Craglia M. and Masser I. 2003. Access to Geographic Information: a European perspective, URISA, Vol 15(1) http://www.urisa.org/Journal/protect/APANo1/craglia.pdf

European Union. 2007. Directive of the European Parliament and of the Council establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) Joint text approved by the Conciliation Committee provided for in Article 251(4) of the EC Treaty. Available at: http://register.consilium.europa.eu/pdf/en/06/st03/st03685.en06.pdf

Gould, M., 2005, Meta-Findability: Part 1. GEO:connexion, Magazine available at http://www.geoconnexion.com/uploads/meta_intv5i7.pdf [accessed 17/01/07]

Howison J. and A. Goodrom 2004. Why can't I manage academic papers like MP3s? The evolution and intent of Metadata standards, available on http://www.freelancepropaganda.com/archives/MP3vPDF.pdf [accessed 17/01/07]

Kuhn W. (2005) Geospatial Semantics: Why, of What, and How? Journal on Data Semantics, Special Issue on Semantic-based Geographical Information Systems, Lecture Notes in Computer Science, 3534: 1-24.

Kuhn W. (2003) Semantic Reference Systems International Journal of Geographic Information Science, Guest Editorial. 17(5): 405-409.

Nowak J. and Craglia M. 2006. INSPIRE Metadata Survey Results. Luxembourg: Office for Official Publications of the European Communities EUR 22488 EN available at http://www.ec-gis.org/inspire/reports/INSPIRE_Metadata_Survey_2006_final.pdf

Pike WA and M. Gahegan (2004) Visualizing concept relationships in a distributed knowledge sharing environment. GIScience, Adelphi, MD, October 2004.

Senkler, Kristian, Voges, Uwe, Einspanier, Udo, Kanellopoulos, Ioannis, Michel Millot, Gianluca Luraschi, Cathal Thorne, Lars Bernard and Paul Smits, Software for Distributed Metadata Catalogue Services to Support the EU Portal, DG Joint Research Centre, Institute for Environment and Sustainability, Technical Report EUR 22337 EN.