www.terradue.com

On Metadata about geographic information

Jo Walsh

The following essay provides an overview of concerns around providing metadata about geographic information to users of a spatial data distribution service.

Identifying best practise, it is intended to provide supporting context for the use of a minimal abstract model and to offer an overview of current and future concerns in geographic metadata production and use.

Short Title	White Paper Template
Prepared by	Jo Walsh
Approved by	Pedro Gonçalves
Reference	T2-Research-07-001-OnMetadata
Issue	1
Revision	0
Date of issue	2007-09-07
Status	Final
Document type	White Paper
Distribution	Public
Category	Research
Keywords	metadata, geographic information
Comments	The first of three documents about Geospatial Services and Applications

Table of Contents

Introduction

Metadata is placed centrally in Spatial Data Infrastructure initiatives. It provides the groundwork from which to share data between applications. After all,the technological part of an SDI or a research data distribution infrastructure has as its goals:

- Making data easier to find
- Offering clear statements about the right to access reuse and redistribute data once it is found
- Improving the accuracy and currency of geographic data over time by facilitating sharing and reuse of it.

Both producers and users of geospatial data and metadata, lament that metadata lacks proximity to the systems they work with. Too often, the task of creating metadata becomes part of a later documentation phase - quite likely not carried out by the original creator of the data, and thus lacking their knowledge contribution. Just as documentation better reflects reality when it is constantly written in the process of a software project, so metadata better reflects reality when it is created and updated as constantly as the data it describes, is handled.

The needs of the producers, and potential consumers of geographic data provider different driving forces in creating a climate where more of it can be more useful. It should be easier for data producers to maintain metadata themselves, with minimal extra task load - by means of semi-automatic, assisted collection of metadata, and better integration into client software. Recommended fields or properties are kept to the minimum that is necessary and useful for data documentation and reuse.

One must find agreement between the goals of any SDI as a network, and the needs of data-producing agencies and those building spatial data search services. Not to impose too heavy a burden on the data producer, the generation and maintenance of geospatial metadata should be:

- Done close to the data production process itself, extending into the same tools.
- Storable in parallel with the data, possibly in the same data storage.
- Easy to synchronise across repositories and to reconcile possibly disparate parallel versions of "the same" description.

A Minimal Model

All of these concerns influence the creation of a minimal model for metadata about geographic information for practical use. The Draft Implementing Rules for the European SDI Directive take this approach; DCLite4G is an expression of this model as a Dublin Core Application profile. [DClite4G, INSPIRE-MD]. This is a common subset of the major metadata standards. What is recommended is characterised by a concern for machine-readability, semi-automatability, and what is of most use for the sharing of data and the building of search and discovery services.

Where data ends and metadata begins, can be a philosophical discussion with many viewpoints. The geospatial standards community has a fairly fixed idea of what metadata *is*, exemplified by the ISO standard document 19115:2003. This standard specification has gained a lot of traction, and many national- and regional- level SDI initiatives are mandating some subset of ISO19115 as a "core profile", not including all the properties ostensibly mandatory in 19115 proper.

For this reason it is for the best to maintain an approach that can be compatible with ISO19115. However, it is not enough simply to state "just use ISO19115". To accurately describe online resources one needs to be more specific. Many of the classification taxonomies that it mandates lack immediate use either for data publishers or for search services and end users of an SDI. One needs the ability to extend a standard with domain specifics, the ability of software to read and reuse it without human intervention - this seems necessary if geospatial metadata is going to become available in sufficient fluency and volume to build search and discovery services as useful as can be.

Machine-readability, or more importantly machine-reusability, of data description information is key. The FGDC standard for metadata in North America has a lot cultural commitment to it, and quite widespread implementation. Though it recommends an XML format, much FGDC metadata is published in a fragile tab-separated format for which a complex custom parser is needed, which it is not possible to extend with namespaces, or do data typing if that is desired. ISO19115/ISO19139 (XML representation of ISO19115 metadata) does not provide a significant difference or technical advantage over the XML version of the FGDC profile, for which suggested DTDs are published.

Simple, machine-readable models become especially important in the context of an active network of data sharing. It is here that some of the current standards provisions are missing connections. Thus the provision of an extension namespace and recommended usage constraints for Dublin Core - an *Application Profile* for geospatial metadata - is appealing. Dublin Core is well-understood and widely propagated. Many metadata standards documents and specifications provide a model mapping to Dublin Core.

Principles for metadata systems

While wishing to avoid unnecessary complexity, one is warned against keeping things *too* simple. The balance between expressiveness and usability appears quite differently to different people. Exasperatingly, one does not want to set the bar too low. These are the core suggestions for improving metadata reward in return for investment of time and effort. They are suggestions also broadly echoed in the *Commons of Geographic Data Whitepaper on Metadata* [CGD-Metadata].

Machine readability

Maximise machine readability of what is collected, and expressed, in metadata:

- Identifiers for things
- Reuse of Well-known domain models with URLs
- Future use in data distribution

ISO19115 mandates in a lot of places the use of external "code lists" (to indicate high-level category classification; restrictions on usage; even organisation roles of responsible parties). The looking up of properties in code lists requires a lot of out-of-band effort on the part of software designed to help create, or to retrieve and understand, geospatial metadata. In cases where code lists appear, a more effective tactic would be to use URI schemes - unique identifiers which as a side-effect tend to be much more human-readable as well. No "agent" can know that the semantics of a classification are expressed in a code list which may, possibly, be looked up via a service to resolve a human-meaningful value.

Where metadata is serialised in XML form for transport across a network, one should look to provide as much hint in the structure as possible, as to how the data can be used in a consistent way. For example, rather than putting arbitrary coordinates or a string of Well Known Text encoding a geometry into a <dc:spatial> property, one should include a <georss:Box> to indicate extents inside it.

When describing people who are "resource responsible parties", publishing or maintaining a data set; one should attempt to emphasise properties that are consistent can be used partially as machine identifiers - such as a parson's email address, which identifies them reasonably uniquely. Rather than focusing on identifying people through properties which are highly variable - names spelled or misspelled or encoded in different ways - one should seek to connect data sets to people through an abstraction which is reusable.

Extensible Models

Offer extensible models providing a clean way for domain specialists to enhance a constrained "common core". An aim is to allow user communities to provide their own, more granular subset domain models without breaking compatibility or mutual intelligibility with a main body of information. The ability to supply "plugin" vocabularies obviates any need, in providing guidelines, to limit what can be expressed.

The social/technological process surrounding the Implementing Rules for the INSPIRE Directive establishing a framework for Spatial Data Infrastructure in Europe, is taking an approach like this. The creation of recommended vocabularies is delegated to Spatial Data Theme Communities; no explicit description of their contents inside the core legal framework.

A useful precedent in a related domain is that of RSS, the XML format for syndicating feeds of metadata about objects on the web, media and structured data increasingly. An RSS or Atom metadata feed can be extended with many different specialist vocabularies through the use of namespaces. Parsers or agents with no understanding of special terms simply skip over them; adventurous discovery agents can attempt to look up the namespace and find a machine-readable expression of its terms. The semantics of terms can be extended at namespaces (e.g. URLs) using the W3C standards RDF (Resource Description Framework) Schema or OWL (Web Ontology Language), or through XSD or other XML schema description, to learn more about how they should be processed and stored.

Multi-modal user interfaces

Offer multi-modal user interfaces to expose specialists and non-specialists to different levels of detail for metadata properties. The Commons of Geographic Data group, in their assessment of metadata best practise, propose a "Ten Minute

Rule" - that documentation of a data set should take no more than ten minutes for its original creator. One is tempted to consider even ten minutes too long. Many potential users will be non-experts in geographic metadata standards though deep experts in their specific spatial domains, able to provide more detailed abstracts, more knowledge about the history and quality of data sets. Partial metadata, and metadata at different levels of detail, should be accepted. Radically simplified interfaces for quick creation should work in parallel with finely grained interfaces that provide what metadata management specialists have come to expect. [CGD-Metadata]

Low Cost

Use lowest cost components, where cost is financial but is also time and energy commitment cost. This is a classic benefit of open source software use - there is less impact on "psychic budget" and more flexibility in the future direction of the system. Starting small, with a software component such as GeoNetwork to manage and publish metadata emerging from many different, some proprietary backends; extend this principle outwards into distribution and storage systems, distributing system load across many small and cheap components. A common metaphor used in the open source GIS community is that of a "lego model" of interchangeable parts, for ease of support of new developments.

Profiles and Protocols

One should not take the risk of ignoring good practise in other domains - in metadata for media, or the description and collection of spatial data which is not considered "traditional". The massive search engine companies start to index more spatial data from different sources to complement their mapping and viewing services. Is the prospect of a "Google SDI" a threat or a complement to more managed national or NGO SDI efforts? [PR-Google]

Metadata is a place to build bridges between worldviews, to maximise the use of data for different purposes. A spatial search engine won't pick and choose between different data styles, profiles, protocols, but will try to include and understand as many as possible. Thus an approach to metadata should be the same - "be liberal in what you send, liberal in what you receive.". A metadata harvesting engine working for a discovery service can collapse different standard formats expressing metadata into one core internal model.

Thus a common subset profile for geographic metadata concentrates on the needs of the "excluded middle" (not those specialists immersed in the standards process, or the web 2.0 pushpin applications sharing descriptions of points, but all the analyst practitioners in between). One does not want to disappoint specialists or in any way prevent them from continuing to do as they have agreed to do. If issues arise (such as complaints from communities without great commitment to ISO19115, about the verbosity and redundancy of the ISO19139 recommended XML format for metadata), those providing software and services should work with specialists to find solutions to their troubles; drawing help where possible from other domains that share similar problems.

An attempt to stay "protocol neutral" aids in spreading risk around technological investment - not making too deep a commitment to any protocol or profile which is still unproven, retaining the ability to support many. Though there may be a good basic level of agreement on a core profile for geographic metadata, there is disparity in the approach to "catalog services" and different interfaces for transferring it about. None of the currently available protocols is really gaining momentum across the GI community. Nor do any of them seem particularly well prepared for the the movement of geographic data publishing, search and storage onto a more distributed basis.

Synchronisation - a protocol failure?

Thus far the Catalog Services for the Web Version 2 standard has failed to propagate. It is over the bar of complexity for casual implementation or use, and its inter-node content harvesting capacities are not well tuned to exchange of data between services. With potential support for many profiles, CSW2 leaves so many options open that its existence is not helping to resolve the common problems of mapping between models, connecting agents to online resources or providing fine-grained updates on changes. The Electronic Business Registry Information Model, ebRIM, comes as part of a protocol family designed to have extensible query interfaces. Within its terms it is straightforward to extend to do simple spatial queries - e.g. on a bounding box representing spatial extents. ebRIM is part of a very generic framework, again beyond casual use in geographic information systems. Even more than with the use of OWL, ebRIM incurs an extra layer of abstraction away from the data, needing any supplied model to be translated into the Registry Information Model.

The Open Archive Initiative Protocol for Metadata Harvesting - OAI-PMH - fits the 'low bar' criterion fairly well. A simple bounding-box-query extension to it could become a de facto standard, and it has some groundswell of use in metadata sharing initiatives. As a protocol design, OAI-PMH is somewhat "backward looking", to an era where one authoritative source of information published it; there is no contribution or synchronisation mechanism in it. There are tentative developments in the OAI community to establish a data exchange protocol which are worth keeping an eye on in the general context of exchange of scientific data between repositories.

Yet none of these candidates provided the basic facility - time-based updates, synchronisation of changes - that GeoNetwork ended up implementing a non-"standard" inter-node harvesting mechanism in order to do.

The harvesting protocol inside the GeoNetwork metadata management package holds implications for the spread of data and metadata through an SDI. Metadata is syndicated between nodes that correspond to social networks within an organisation. At the UN's CGIAR, each department runs its own GeoNetwork node, all are 'harvested' by an agency-wide one, this harvested in turn by a central FAO repository of shared data descriptions. If one needs to represent the structure of a hierarchy in a system design, this can be done; equally a set of peers can all synchronise with each other, using the same underlying interfaces.

This does not constitute a recommendation to reuse the current GN harvesting interface directly though. One element is lacks is a clean mechanism for doing remote authentication. One can "log in" with the credentials of a user or group, but only if one knows the local system aliases before hand; if one wants a user account on many GeoNetwork nodes, each one must be created locally, negotiated socially, out of band. This illustrates a situation in which integration of a metadata/data exchange network with a Certificate Authority or other means of providing credentials via a third party service. A future essay on *Data Access and Authentication* is planned to explore the rationale, prior art and implications.

Whether open or constrained in availability, data can be syndicated and the beginnings of a transparent remote archive for both data and metadata, made. A third-party-managed authentication between nodes on a data sharing network can help to resolve the question of how possible it is to usefully *evaluate* data found through a search/discovery service, on the basis of metadata, without being able to get at it.

There are two different kinds of exchanges going on here. Different interfaces are needed between nodes that have previously connected, and know what each other is likely to know about; and nodes encountering one another for the first time and trying to build a picture of a shared domain model, of useful stuff to query. Protocols that try to do both things via the same mechanism are suffering, if not in their original conception then in their current usage, from design failure.

Emerging concerns for future systems

The following topics are "emerging" in the literature on geospatial metadata; but not yet into widespread implementation in metadata tools. These issues are really generic to data coming from observation and research, that is being made available in an open access / open annotation context.

Data lineage

Tracking the **lineage** of data sets. Especially in a context where packages of data, metadata and configuration rules are being offered on-demand by a service, and where public domain sources are mixed in with proprietary ones; it becomes more important to model and to convey in a machine-reusable form:

- What data sets have in part gone into a combined data set
- What algorithms have been applied to an original set to create a processed version.

The INSPIRE Implementing Rules draft on metadata, at time of writing, stipulates a plain-text field for human-readable description of data lineage, which is likely to vary so much as to be useless, and to provide no assistance in being ale to trace back and re-run processes or separate out parts of data free for redistribution from parts that are not.

A representation of data lineage - the history and journey of data - comes under the heading of metadata. It is not typically easy to infer the sources of a composite data set without a lot of effort, and/or reliance on time-consuming steganographic techniques. The Commons for Geographic Data offer some thought-provoking suggestions on embedding lineage in data sets for distribution [CGD-L]. This approach may bear some fruit for raster data admit but for vector data, it's easy to filter out unwanted embedded attribution information, or to obfuscate by generalising geometries of original sources.

Usage reporting

Data providing agencies seeking to focus limited resources where they may be of most use to others, benefit from "monitoring" information on how their data is being reused. Managing metadata about contributions, rights to reuse and kinds of reuse, are all of great interest to the maintainer of a spatial data repository and index.

Domain Model Translation

Mapping between domain models - being able to express and reconcile the difference between similar data sets using alternate taxonomies or ontologies. This comes closer to the developing ideal of the "semantic web". This topic has so many ramifications that a separate essay On the Development of Geospatial Ontology is offered in which to explore it.

In order to establish consensus about the correspondence of meaning in different domain models, it helps to have a broad reference base to establish shared identifiers from. One very interesting effort arising outside "traditional" GIS systems and standards provision is that of geonames.org. Geonames provides an online gazetteer service and RDF feeds associating named places with URIs; it refines and increases the local accuracy of the US-published global gazetteer data set from NIMA. Geonames cross-references place indicators associated with geometries, both with URLs considered definitive for place names on Wikipedia, and with URI schemes published by national cadastral agencies where available.[Geonames] A link between data objects, expressed in an RDF graph, need only be made in one place on a network, its ramifications subsequently traversed and augmented by many others.

Internationalisation and localisation

Internationalisation and localisation both of data sets and data descriptions is a task whose complexity can only be partially known until more SDI components are in place. Yet language localisation will be the largest stumbling-block for building shared spatial data infrastructures across linguistic regions. Search services will suffer the burden of navigating data and metadata in many languages with user requests expressed in many different ways. No further commentary is offered other than this area befits more study than it has been given; more of the impact on search services is discussed, but not in much depth.

Next Steps

Metadata for geographic information is a well-worn subject and a dry one; yet still a "hot topic" in SDI discourse. Why is it still a subject of interest, and irritation, to so many? For "metadata" is used to describe all elements of data in a spatial data system that aren't spatial data. As GIS mutates into SDI and "interoperability" becomes more of a byword than a buzzword, these are new concerns that current geospatial metadata standards aren't designed to handle or express.

The "next generation" of metadata challenges - data lineage, auditing, both human- and machine-language translation issues - provide some justification for a continuing debate; these are solutions that weren't worth designing until the problems became real. Meanwhile, "getting metadata right" provides a solid basis on which to solve data description and search problems in potentially new ways.

Currently, a heavy burden of expectation is being placed on taxonomy, keyword-based classification, and fulltext search of abstracts, to find data. A potential data user must *evaluate*, on the basis of metadata, whether a piece of data is relevant or useful. The next essay in this series, On the Development of Geospatial Ontology, considers "semantic interoperability" and the extent to which it is possible to extend the "minimal abstract model" approach to specifying data domain models.

Even partial access to original data - a small sample or a generalised version of a whole data set - lessens the dependency on "compliant" metadata. [CGD-Metadata] The management and maintenance of data "catalogs" is likely to be a resource-sink to the data provider, without proven utility to data search and discovery services. Means by which to bypass or to complement classification-driven catalog systems are covered in the third essay, On Spatial Data Search.

References

Writing and Legislation

- [INSPIRE-MD] http : //www.ec-gis.org/inspire/reports/ ImplementingRules /draftINSPIREMetadataIRv2_20070202.pdf INSPIRE Directive Draft Implementing Rules on Metadata
- [INSPIRE] http://inspire.jrc.it/directive/l_10820070425en00010014.pdf INSPIRE Directive, particularly Article 11 describing minimal spatial data search service capacity
- [NAP] http://www.fgdc.gov/standards/projects/incits-l1-standards-projects/NAP-Metadata/napMetadataProfileV101.pdf North American Profile of ISO19115 - Geographic Information, Metadata. (Draft)
- [UNSDI] http://www.ungiwg.org/docs/unsdi/UNSDI_Compendium_13_02_2007.pdf UNSDI Compendium
- [GSDI] http://www.gsdi.org/pubs/cookbook/ GSDI Cookbook Chps. 3 and 4
- [CGD-Metadata] http://geodatacommons.umaine.edu/wpapers/CGD%20Metadata%20White%20Paper%20v3.pdf "CGD Metadata White Paper"
- [IJSDR] http://ijsdir.jrc.it/articles_under_review/najar_giger.pdf "Spatial Data and Metadata Integration for SDI Interoperability", Christine Najar, C. Giger.
- [PR-Google] http://geotips.blogspot.com/2007/02/google-sdi.html Paul Ramsey, "Google SDI"
- [PR-Metadata] Paul Ramsey, "Everyone loves metadata"
- [Platial] http://platial.typepad.com/news/2007/04/best_practices_.html "Best Practices for Emitting KML and Giving Props to the Geoweb's Authors", Chris Goad, April 13 2007

Standards, Specifications, Services

- [ISO19115] Geographic information -- Metadata
- [ISO19139] Geographic information -- Metadata -- XML schema implementation
- [CSW] http://portal.opengeospatial.org/files/?artifact_id=20555 Catalog Services for the Web 2.0.2
- [OAI-PMH] http://www.openarchives.org/OAI/openarchivesprotocol.html The Open Archives Initiative Protocol for Metadata Harvesting 2.0
- [DCLite4G] http://wiki.osgeo.org/index.php/DCLite4G Minimal abstract model for metadata
- [Geonames] http://www.geonames.org/ Geonames - open data global gazetteer