EXPECTED METADATA OF GEODATA FOR GEOWEB SERVICES ORCHESTRATION

JAN RŮŽIČKA (INSTITUTE OF GEOINFORMATICS, VSB-TU OF OSTRAVA, OSTRAVA, CZECH REPUBLIC)

The article describes expected extent of metadata for geodata in a process of GeoWeb services orchestration. The expected extent is evaluated according to current stage of spatial metadata infrastructure in the Czech republic and current European legislature. The metadata are very useful in a process of GeoWeb services orchestration, but unfortunately they need to be very precise and accurate. This paper should show what metadata we can expect, when we decide to orchestrate general GeoWeb services in an enterprise environment.

GeoWeb services orchestration project

The research project is targeted to orchestration of GeoWeb services. The main goal of the project is a development of an independent architecture for the orchestration. There are not any rules available for development of independent web services orchestras in the area of GeoWeb (basic platform for SDI (Spatial Data Infrastructure) and SMDI (Spatial Metadata Infrastructure)).

We are going to prepare implementation rules for the orchestration. The first phase of the project analysed languages for planning business processes (such as BPEL, XLANG, ebXML), which are necessary for services orchestration. The second phase is targeted to building a knowledge base for the orchestration. We are going to analyse conditions necessary for running and building orchestras based on GeoWeb platform. The third phase should be focused on testing the prepared rules, architecture and pilot system.

- Sub-projects for 2007 year:
- Analyse INSPIRE requirements for orchestras and GeoWeb platform;
- Analyse Business Processing Languages;
- Analyse WS-CDL;
- Analyse Enterprise Integration Patterns;
- Analyse application server JBoss and SUN Application Server;
- Analyse services for support for crisis management;
- Extending WSCO for UDDI and CSW 2.0;
- Analyse WMS and WFS services of the public administration;
- Analyse application server Zope and CMS PLONE.
- Sub-projects for 2008 year:
- Services monitoring
- Set of orchestras and their monitoring
- Metadata for used geodata
- GUI design of client for orchestration
- Cache-proxy GeoWeb services server
- ESB (Enterprise Service Bus) design for orchestration.

We believe in a complex SDI that can cover services, geodata, metadata and of course SMDI. The research project should describe all the important parts of the SDI for the orchestration as it is shown in the following schema.

- Clients of the architecture use orchestras or simple chains of services and proceed (usually visualise) the results.
- Type of services (or whole orchestras) are searched by clients in the catalogues.

- The current bindings for the services are searched by orchestras in the catalogues.
- Geodata or services can migrate (replicate, duplicate) from one machine to another if it is needed for optimisation of the current network traffic.
- Usage of the orchestras can be free of charge or some orchestras can operate under some kind of licence payment policy.
- All transfers can be secure if this is necessary.

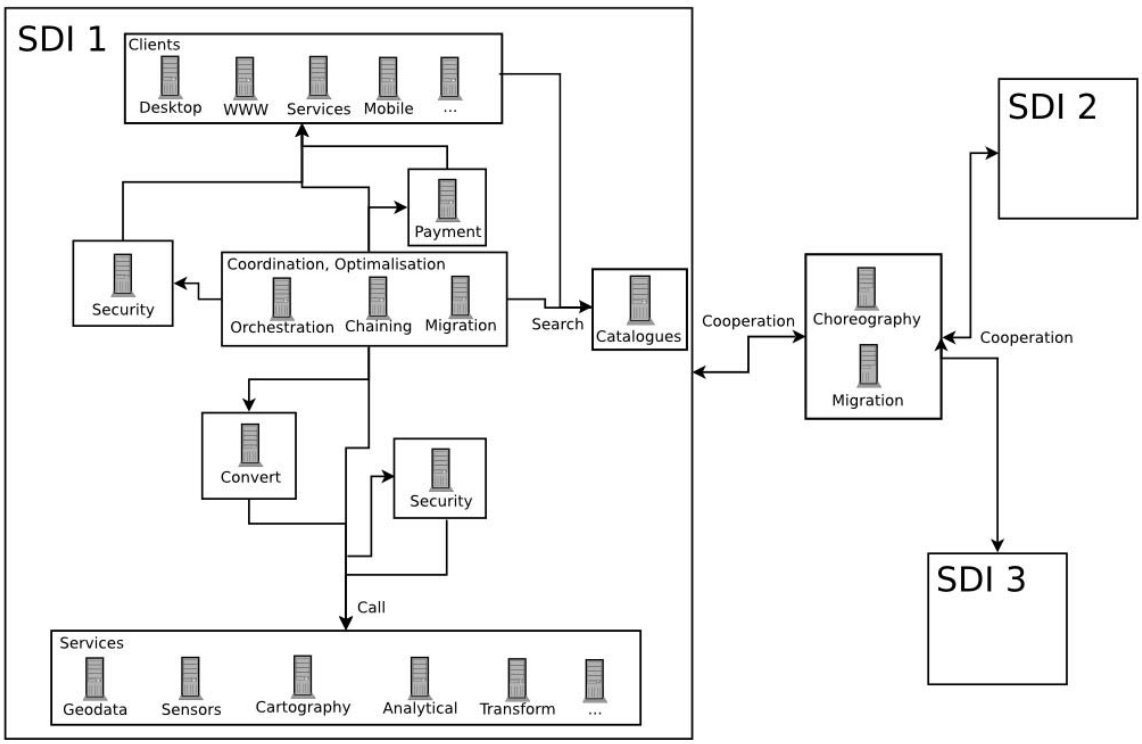

Figure 1: Catalogues in SDI for orchestration

Conversion services can be used for chaining services for the case when I/O structures are not compatible (e.g. GML 3 to GML 2, SVG to PNG, ISO 19139 to Dublin Core, CSW to UDDI, REST to SOAP, SOAP-RPC to SOAP).

Cooperation of one SDI (e.g. company, national, regional) with another SDI can be based on a platform and a process independent choreography. Services or geodata can migrate between SDIs. the migration of orchestras is possible, but not expected, because of the dependency on the processes inside each SDI. We believe in a complex SDI, which can cover services, geodata, metadata and of course SMDI.

Expected metadata extent

Metadata extent belongs to the significant aspects. For geodata evaluation we do not need only general information, but in many cases we need detailed metadata.

We can expect that metadata available in the Czech republic are going to be prepared according to several different set of conditions (rules). This is necessary to know for geodata evaluation.

These different sets are:

- metadata according INSPIRE IR (INSPIRE, 2007),
- metadata according to ISO 19115 core (ISO/TC 211, 2003),
- metadata according to Dublin Core basic set (DCMI, 2007),
- metadata according to level of MIDAS database (CAGI, 2008) completeness.

Other alternatives are not expected.

Metadata according to INSPIRE

The list of items is used from draft implementation rules (INSPIRE, 2007).

Level 1 is a basic level, that will be required always (if the conditional rule does not define different options).

- Resource title.
- Temporal reference – in a case when information is meaningful.
- Geographic extent of the resource.
- Resource language – in a case when text is used.
- Resource topic category.
- Keyword.

Service type – in a case of a service.

Resource responsible party.

Abstract.

Resource locator – in a case if any reference exists.

The second level is extended level and we can not expect full implementation of this level for all catalogues (datasets or services).

- Constraints.
- Lineage.
- Conformity.
- Service type version – in a case of a service.
- Operation name – in a case of a service.
- Distributed computing platform – e.g. Web Services.
- Resource Identifier – e.g. URI.
- Spatial resolution.

INSPIRE specifies other metadata elements, that can be available, but their usage by data (services) provides is disputable. The same problem is with the second level of metadata, where usage is based on provider decision.

We can probably expect only following items:

- Resource title,
- Geographic extent of the resource,
- Resource language,
- Resource topic category,
- Keyword,
- Resource responsible party,
- Abstract
- Temporal reference (in some cases).

That level of detail is not enough for the orchestration, but it can be used for a basic services selection. The main problem is going to be with the item keyword, because providers can use different thesauruses.

Metadata according to ISO 19115 core

ISO 19115 is more detailed than INSPIRE requirements and is going to be better applicable for orchestration. But we are still missing quality reports, constrains of the usage and other items. Items in the core are Mandatory (M), Conditional (C) or Optional (O).

- Dataset title (M),
- Dataset reference date (M),
- Dataset responsible party (O),
- Geographic location of the dataset (by four coordinates or by geographic identifier) (C),
- Dataset language (M),
- Dataset character set (C),

- Dataset topic category (M),
- Abstract describing the dataset (M),
- Distribution format (O),
- Additional extent information for the dataset (vertical and temporal) (O),
- Spatial resolution of the dataset (O),
- Spatial representation type (O),
- Reference system (O),
- Lineage (O),
- On-line resource (O),
- Metadata file identifier (O),
- Metadata standard name (O),
- Metadata standard version (O),
- Metadata language (C),
- Metadata character set (C),
- Metadata point of contact (M),
- Metadata date stamp (M),
- Metadata according to Dublin Core.

Dublin Core is general standard and can be used for definition of own items, but we can not expect that providers will use such capabilities. They will probably use only simple metadata items list.

- Title.
- Creator.
- Subject.
- Description.
- Publisher.
- Contributor.
- Date.
- Type.
- Format.
- Identifier.
- Source.
- Language.
- Relation.
- Coverage.
- Rights.

Metadata according to level of MIDAS database completeness Core

We have analysed MIDAS database and we can probably expect same providers behaviour in the future. But some of the results are declined, because their completeness was controlled by system MIDAS (and used standard). The following table categorised metadata items according to completeness in the MIDAS database. MIDAS system contains metadata about 3400 datasets.

Mandatory and conditional items were always filled (was controlled by the system). Optional items were filled in a case, when list of options was available. Very interesting is completeness of alternate title, temporal extent (date from), reference data and dataset usage. Out of interest are quality elements (except lineage).

No metadata

We can expect that some of the services are not going to have metadata available. We can contact providers, but if there is not any response and the service seems to be useful we have to (for orchestration purposes) create (at least basic) metadata ourselves. Keywords,

categories and language can be derived from some documents, published by a provider and extent of the geodata have to be included in the GetCapabilities response.

Table 2. Completeness of the metadata items in the MIDAS database

Completeness	Metadata items
80 – 100 %	Title, abstract, coordinate system for metadata, metadata update, spatial schema, lineage, horizontal spatial accuracy, update frequency, data structure, format, language, classification, direct coordinate system, responsible party.
60 – 80 %	Alternate title, temporal extent (date from), planar extent (by coordinates), reference data.
40 – 60 %	Dataset usage
20 – 40 %	Memo, planar extent (by description)
5 – 20 %	Abbreviated title, version, purpose of production, temporal extent (by description), metadata language, spatial coverage, scale, temporal extent (date to).
< 5 %	English title, English abstract, update date, fees, metadata update plan, vertical spatial accuracy, logical consistency, completeness, homogeneity, resolution, quality, vertical extent, distribution units, medium, indirect reference system, vertical reference system, features description

Metadata created automatically

From many years of running MIDAS system [CAGI 2008] we know that metadata must be created automatically during the process of geodata creation. This can help with usability of metadata.

The whole architecture expects that metadata will be automatically stored in a catalogue, directly from the source and it must be done every time when geodata are updated or created. The following summary shows what kind of metadata items could be filled automatically.

Identification

The system can simply use place of publishing (URL, URI) or generated unique code (quite common).

Title

Title can be derived from the name of the file, directory, table, or database. In this item the user should probably do some correction, but the correction is not always necessary.

Spatial schema

Basic schema, such as point, line, polygon, grid, tin can be identified directly. More complicated schema rules, such as topological rules, can be identified, but not for all cases, especially when the rules are not directly coded in the system.

Sample

Static or dynamic preview of the geodata is simple to produce.

Coordinate system

In these days a new created dataset has usually defined the coordinate system.

Geodata extent

Spatial extent is simply defined. Temporal extent is a little bit more difficult, but can be derived in many cases.

Quality

The most extended part of metadata (information about geodata quality) must be (in many cases) generated automatically. This is not so technically difficult when are geodata produced, but there is problem with rules of software developers.

There is a problem with closed source software. Part of the quality report must be description of the used algorithms (way of implementation, used parameters) for geodata manipulation (e.g. line generalisation). Closed solution usually does not describe such items in a detail and source code is not available.

We believe that there will be strong impact of users, who are going to ask their software vendors for open source version of their software.

Data dictionary

There can be a problem with ambiguity of data types, but this can be solved by GML or simple XSD.

Classification

A new created geodata should be created in some semantic context (ontology), but there will be probably needed user interaction. At the beginning of the geodata creation user should select context of the data from some ontology.

Administrative metadata

That is an easy part (when we are talking about basic administrative metadata). A data creator is usually identified by the operation system (network - LDAP, Active Directory) and other metadata are necessary to fill in only once.

Metadata of metadata

In this case metadata must be created automatically.

Related items

These items can be generated partly. There will not be usually problems with related datasets and services. Others should be defined by the semantic context (ontology). If metadata are created in some context we can find in OWL or RDF documents relations to other items such as legislature, events, people or other documents.

Current state

It is clear that GIS can not automatically create all metadata for all geodata, but even if the metadata are generated for 80% of the datasets, we are closer to full interoperable SMDI. Unfortunately we are far from this 80%.

Conclusion

Results are not so optimistic, because we can not expect in any potential case that metadata are enough detailed for the efficient orchestration. The situation may not be better in next years. The change must come from GIS software developers and they will not probably do this in efficient way. We are going to find another alternative ways, how to evaluate served geodata.

We have decided to test in our project a way that is not based on metadata. Our simple solution that will be tested this and next year is based on evaluation of results that are produced by orchestras. The evaluation will be based on user (expert) point of view. His satisfaction (dissatisfaction) with result will be stored in a knowledge base for further evaluation. We are working on methodology for back tracing of the orchestras that can help with better specification of dissatisfaction with orchestras.

The metadata will play role for basic geodata evaluation, but the main weight will be on knowledge base and its evaluation.

References

1. **CAGI.** (2008). MIDAS. 2001- 2008. at http://gis.vsb.cz/midas, [accessed 2 July 2008].

2. **DCMI.** (2007) *Dublin Core Element Set v. 1.1. – Reference Description*. at http://dublincore.org/documents/dces/, [accessed 12 April 2008].

3. **INSPIRE.** (2007). *DT Metadata – Draft Implementing Rules for Metadata*. at http://www.ec-gis.org/inspire/reports/ImplementingRules/draftINSPIREMetadataIRv2_20070202.pdf, [accessed 12 April 2008].

4. **ISO/TC 211**. (2003). *ISO/FDIS 19115:2003*. ISO/TC 211 Secretariat, Oslo, Norway, 152 p.
5. **Růžička, J., Kaszper, R.** (2007). Opět o metadatech v geoinformatice. *Proceedings 1. národní kongres v Česku – Geoinformatika pro každého, May 29-31 2007, Mikulov, Czech Republic,* at http://mikadapress.com/prednasky/Ruzicka.pdf, [accessed 2 July 2007].