Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

ONTOLOGY-BASED INTEGRATION OF DATA AND METADATA

Christine Giger, Christine Najar

Swiss Federal Institute of Technology (ETH) Zurich
Institute for Geodesy and Photogrammetry
ETH Hoenggerberg, 8093 Zürich, Switzerland

1. INTRODUCTION

1.1 GIS Applications and Ontologies

According to the SDI Cookbook [1] a Spatial Data Infrastructure (SDI) is a means to host spatial data and attributes as well as sufficient documentation (metadata). It should also be a means to discover, visualize and evaluate the data, and to provide access to the data. In order to develop such an infrastructure we have to face the necessity of integrating spatial data of different sources. Several other authors already stated the problems in handling spatial data in general. From a computer scientist's point of view, we have to face the usual problem to formalize a specific part of the information space in order to develop deterministic algorithms to process the resulting data. In context with spatial data, this information space usually describes parts of the real world and natural and/or artificial objects on its surface or as parts of that world. The occurring problems often cope with the amount of data used in one single application and several – in other IT application areas also common – problems of interoperability between software systems of different providers. Still, if all these problems could be solved, users of Geographic Information Systems (GIS) tend to act in contradiction to all other scientists. The latter often try to build common "single" and "true" models of the reality and try to agree on a common understanding of the reality by specifying models or an ontology. GIS users normally specify fairly different models for the same object depending on their notion and with regards to their specific application and point of view. The Open GIS Consortium promoted the helpful notion of an information community. According to [2] an Information Community "is a collection of people (e.g. a government agency, a profession, a group of researchers in the same discipline, etc.) who … share a common digital geographic information language and share common spatial feature definitions. This implies a common world view as well as common abstractions, feature representations, and metadata." When trying to develop an SDI we quickly realize (as in any application of spatial information) that there is a demand for integrating spatial data from different sources, which belong to different information communities. But, usually this integration is non-trivial because of the different semantics used for the spatial data in different information communities. So, even if we ask the seemingly easy question "What is a street?", we will receive fairly different answers according to the membership to a specific information community.

Obviously, there is no single geometric, topologic and/or thematic description of a street. One could state, that in contrary to the philosophic approach, there is no single ontology describing a street. Nevertheless, within an SDI you need services for data visualization, data transfer, data integration, data migration, quality checks, etc. for these data sets. The only possibility to provide them and cope with the effect is to provide formal descriptions of the different ontologies as well as deterministic algorithms to map the different ontologies onto each other. Here, we use a notion of ontology as it is common to computer science [3], [4], denoting the complete but formal description of objects which exist in a given model space, or more exact a specification of a conceptualization, as it is explained in detail e.g. in [5] and [6]. This includes all physical, independent, relative,

abstract and mediating aspects of the object and leads to e.g. processing, participation or historic behavior of the object.

Today, we know several semi-formal and formal methods to describe ontologies. As semi-formal we regard for example Entity-Relationship-Diagrams or models in the Unified Modeling Language (UML). Formal models would be specified for example in the Web Ontology Language (www.w3c.org), which is based on RDF and XML, or simply on RDF, or on other, algebraic [7] or logic-based [8] approaches. Of course, any formal language would be suitable but not necessarily elegant to deal with ontologies. In Switzerland, INTERLIS [8] was developed and used for describing conceptual models especially in official surveying. INTERLIS is a conceptual schema language specifically created for the conceptual modelling of spatial objects and it already solved many problems on data integration, migration and transfer for GIS applications. Recently, we used its newest Version 2.1 to specify ontologies formally. Since they include more general, process-oriented aspects of objects than existing conceptual models of surveying data, this was not a trivial task and even not possible with the former versions of INTERLIS.

In any case, the standardized use of a formal specification of ontologies has to be integrated into an SDI. On that basis the SDI has to support different information communities in

- finding appropriate data;
- accessing data;
- integrating data in existing solutions;
- data visualization and analysis;
- etc.

1.2 Integrating Data and Metadata to Support an SDI

Although the “Spatial Data Infrastructure” (SDI) became a common description of a specific understanding of providing geospatial information to users, recent approaches show that we should rather speak of a “Spatial Information Services Architecture”. This is a fact assuming that digital spatial data is of no value to the end user (consumer) without the corresponding software tools. Even business users can only profit in a limited way of spatial data, if it is not embedded in a framework of services and therefore providing added value.

According to the services idea, any SDI would have to follow some basic principles. Such as

- Mainstream IT has to be used.
  This means that accessible technology has to be used. Technologies under development have to be taken into account for midterm and long-term planning, but have to be continuously evaluated and adapted if necessary, the planning has to be adapted to the actual state of development.
- Standards and the international situation for implementing SDIs have to be taken into account.
  On every implementation level of the infrastructure outside and inside a specific region or country it is necessary to guarantee that other infrastructures can be integrated or at least connected.
- Different information providers within or without public administration will exist, and will be (and stay) owners of the data.
  An infrastructure will have to support decentralized organizations as well as technical implementations (servers) of data and metadata sources. Tools must be available to integrate additional units.
- The technical infrastructure has to support technical innovation and must not avoid or interfere with it.
  This implies the specification of rules and constraints by means of standards as far as possible and useful. At the same time there should be freedom to select components of the infrastructure from different providers.

- The usability of the infrastructure for end users (consumers) as well as providers of services (business users) should be in the focus of the development. This demand implies to strive for a services infrastructure instead of a data infrastructure. Any technical component has to support this goal.

These principles lead to a variety of different technological consequences for implementing the infrastructure. One of these consequences, which is the core of this paper, is the idea that metadata can only be a means of communication for detecting the best data for a specific service. This leads to the hypotheses, that

Metadata should not be treated as descriptive but independent data from the original data set. Instead Metadata must be an integral part of any data set.

This would not only provide flexible interpretations of what is data and what is metadata, it would also lead to processes for data and metadata (as part of the data) acquisition which naturally produce consistent data and metadata.

2. CURRENT STATE-OF-THE-ART ON METADATA ACCESS

The existing approaches for implementing an SDI focus on the development of sufficient data documentation (metadata) in the first step. For the first metadata specifications it was usual to build up huge centralistic servers. These servers followed the idea of one access point and the provision of huge and as complete as possible data catalogues describing a variety of different data sources of all possible data providers. Recent implementations [10] take already into account that different data providers are also metadata providers and are parts of a decentralized network of servers. Nevertheless, there is a clear distinction between data and metadata servers. In Figure 1 the current situation is illustrated, using the ISO 23950 (Z39.50) approach for the protocol and the ISO 19115 for metadata specification.

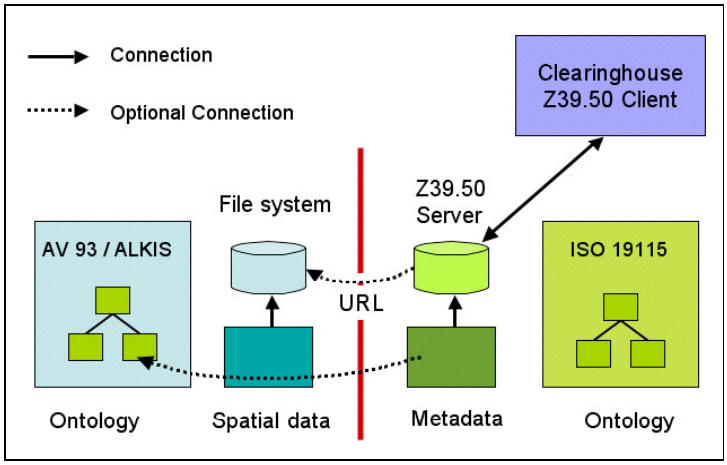

Fig. 1 Current distinction between spatial data and metadata

In the case of the *Final Draft International Standard ISO 19115 Geographic Information – Metadata* [11] the metadata schema is fairly huge (about 300 attributes). A link to the data set may be given through a URL, but is optional as well as the link to the data model or even a formal specification of the ontology. Usually, the metadata will provide only a link to a feature catalogue that consists of a list of all possibly contained objects.

So, a question “does the data set contain streets?” cannot be answered definitely without downloading the whole data set. For example, if you would specify a query on metadata describing GDF files (car navigation data) and asking for pedestrian traffic lights, you would receive all available GDF data sets in that network (in the world). This happens, because the format description and therefore any derived feature catalogue or ontology would contain pedestrian traffic lights. Nevertheless, none of the companies, who currently use GDF (Tele Atlas, NavTech), ever acquired these data anywhere in the world, at least to our knowledge.

In addition, there are many examples for the problem of clearly distinguishing between data and metadata at all. For example:

- Does the reference co-ordinate system belong to the data or to the metadata?
- Does the accuracy of single point co-ordinates belong to the data or to the metadata?
- Does the bounding rectangle or volume belong to the data or to the metadata?
- and so on….

3. INTEGRATING DATA AND METADATA UTILZING ONTOLOGIES

3.1 Approach

When we analyse existing ontology-based or model-based approaches in order to specify geospatial data sets, as it is already practised in some Scandinavian countries and in Switzerland, we realize that there is always an explicit link between the data set itself and its corresponding data model, from which a formal ontology can be derived. This does not necessarily mean that the ontology is always transferred together with the data set. Often enough, the target (GI) system does not provide the functionality to read and interpret the formal ontology. In this case only the pure data will be transferred. The ontology – in this case – has to be implemented beforehand by specifying explicit data models and data structures for the specific GIS.

We analysed the following datasets:

- for Germany:
  cadastral data of the states of Hessen and Bavaria according to the following data models: Automatisierte Liegenschaftskarte (ALK), Automatisiertes Liegenschaftsbuch (ALB), ALKIS
- for Switzerland:
  cadastral data of the city of Zurich and of the federal government according to the following data models: AV-93, DM.01-AV

The analysis of these data sets and their corresponding ontologies leads to the discovery that already a large percentage of the needed metadata according to e.g. ISO 19115 can be automatically derived from that information. A short example is given in chapter 3.2.

This is obviously another hint that it might be useful not to distinguish too strictly between data and metadata at all. Our approach therefore tries to interpret metadata as a **view** on the original data and its ontology.

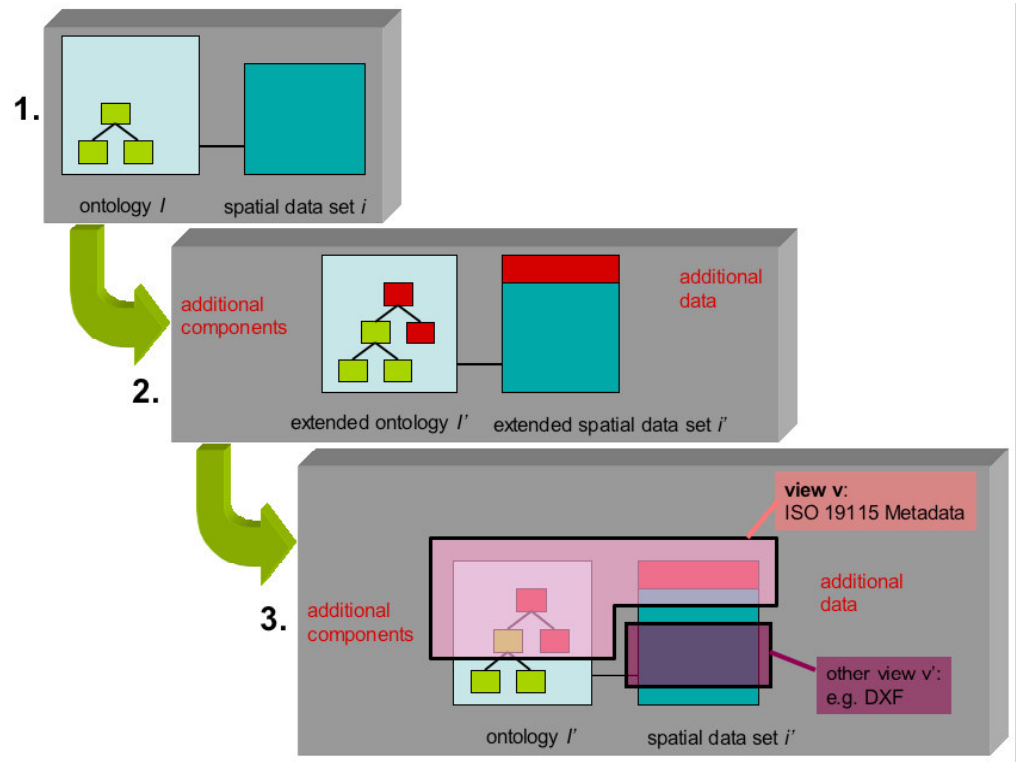

Fig. 2 Extending existing data sets to carry metadata and producing metadata views

Figure 2 shows three steps:

- Use or generate an explicit (formal) description of an ontology *l* corresponding to a given data set *i*
- generate new data set *i'* by adding metadata to the data set *i* as additional data, extend the corresponding ontology and generate new ontology *l'*
- Generate view on data set *i'* and ontology *l'* in order to generate a metadata set *v* (or any other data set *v'* according to a given structure or format)

In order to implement the three steps, we have to fulfill the following preconditions:

- We have to be able to generate a formal ontology for a given data (metadata) set. In our example, we need an explicit representation for the corresponding ontologies for:

  i. cadastral data (DM.01-AV)
  Its data model is specified on the basis of INTERLIS 1.0. Using the INTERLIS 2.1 compiler, we are able to produce a formal ontology using an object-oriented modelling language. This is done by restructuring the model and adding inheritance information.

  ii. metadata (ISO 19115)
  Its data model is specified in the ISO document as a UML model. Using existing tools (UML Editor (Eisenhut Informatik) or Rational Rose with INTERLIS module) we are able to add the needed structures to receive a complete representation of the corresponding ontology.

iii. Any other data format we want to provide.
For the most commonly used GIS the corresponding ontologies are known and even available as INTERLIS specifications. In general, there has to be a complete specification of the format and the semantics of the used elements and attributes as well as encoding rules for the generation of the format. If this is available to the public, a formal description of the ontology can be generated and used in this process.

- We have to know, which (meta) data should be added to existing data sets and how the ontology has to be extended in order to fulfill our needs. In general, this may very well be a specific task for every (formal) ontology and its accordingly specified data sets. We assume that existing data models will have to be extended and that we will have to apply general rules for data and metadata acquisition in related processes.
- We need tools for reading and mapping (in other papers called "merging" or "integrating") ontologies, in order to generate the needed views.
- We need tools for quality checking. At least we have to be sure that a given data set corresponds to the given ontology. Other quality aspects, such as topologic correctness, completeness, consistency, etc. of the data are also important.
- We need tools for incremental update of data (and metadata) sets in order to support acquisition and updating processes.
- Search engines have to exist, which are able to read and interpret data (and metadata) provided through ASCII-based files (e.g. GML [2]).

The process of viewing (step 3) can be implemented as follows (see also figure 4). The basis for the implementation is one or more specified formal methods *M1, M2, …* to describe the ontologies *I, I' and V* provided by different information communities. It is also necessary to have a clearly defined set of encoding rules *R* for each formal method. These rules provide the information of how algebraic, logic or language elements of the methods are mapped to physical representations (i.e. file formats) of the data describing the ontology *I* itself as well as data *i*, which corresponds to the ontology. These rules can then be used to "translate" and/or "map" ontologies onto each other. This has to be done in order to interpret, re-arrange and restructure the data in Service *S* for producing the view *v*, which can be interpreted as a subset of data *i'* but represented according to ontology *V*.

3.2 Preliminary Results

In our prototype implementation, we chose for ontology *I* an adaptation of the official surveying model DM.01-AV in Switzerland. As a corresponding data set *i* we used the public available test data set of the Swiss Federal Surveying Office. We performed step 2 (illustrated in figure 2 and 3) manually by comparing the existing ontology *I* and data set *i* to the ISO 19115 core metadata ontology and data set and extending *I* to *I'* and *i* to *i'*.

Figure 2 illustrates an excerpt from the basic data schema of that data set (DM.01-AV), which was used to specify the ontologies *I* and *I'*.

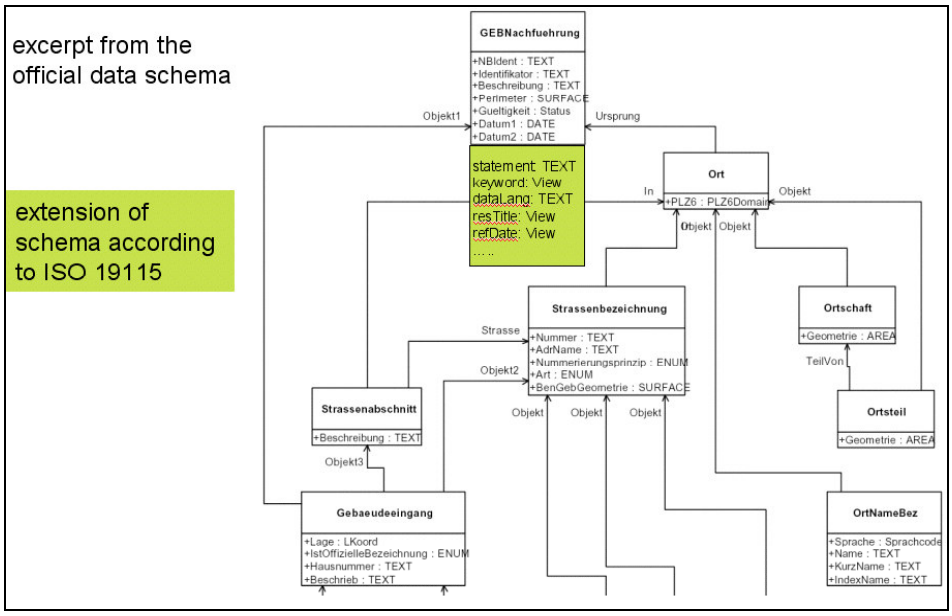

Fig. 3 Excerpt from the official schema for DM.01-AV (Swiss cadastral data, here update of buildings), corresponding to ontology I, and its extension to carry metadata, corresponding to ontology I'.

We then implemented step 3 (see figure 2) by using the general approach illustrated in figure 4.

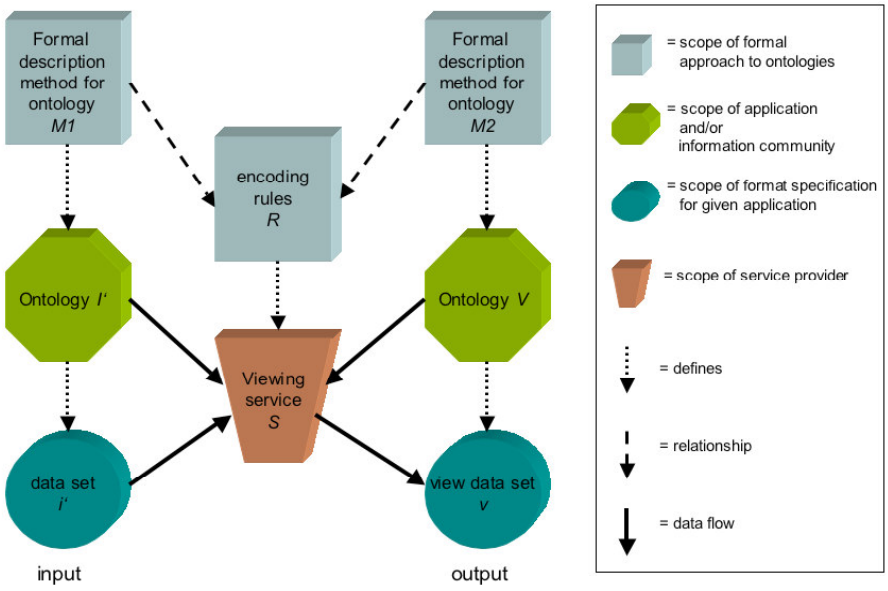

Fig. 4 General overview on the process of producing a view on a given data set, corresponding to step 3 in figure 2

We used INTERLIS in step 3 as formal method *M1 = M2* and received the encoding rules *R* through the INTERLIS compiler. Our ontology *I'* was the extended DM.01-AV data model and data set *i'* was the extended test data set for DM.01-AV, which we received from step 2. Ontology *V* was derived from the UML model, which is provided by ISO 19115, and which we described formally using INTERLIS. We then used a commercial software package (infoGrips conversion system ICS, www.infogrips.ch) to implement a first prototype of the viewing service *S*. The resulting data set *v* is provided on an XML Schema basis (currently INTERLIS does not support GML, although a respective compiler option was already announced). *v* can be checked (by using for example iG/Check www.infogrips.ch) to be consistent to ontology *V* and it can be used and represented in a variety of existing GIS (ARCView, GeoMedia, MapInfo), ODBC enabled data bases or ORACLE spatial.

For INTERLIS users, it is a usual approach to produce "views" on the data set itself. Although most of the users initiate the viewing process for format conversion to facilitate data transfer from one GIS to another, which is the main purpose for which INTERLIS was created. In our case the situation is a little different. We want to produce new data *v* out of the original data set *i'* AND the original ontology *I'*, because some parts of the metadata are not included in the data set *i'* itself but (only) in its corresponding ontology *I'*. The latter is currently not supported by ICS but can be implemented on that basis.

4. SUMMARY AND OUTLOOK

We were able to implement a first simple prototype and example to integrate spatial data and metadata. Furthermore, we generated in three steps a usable formal specification of an ontology for a given data set, we extended the ontology and the data set to carry its own metadata and we produced a view on the ontology and the data in order to support existing standards for metadata description and access. Based on the first results we see evidence that the general approach might be useful and usable to facilitate the implementation of an SDI. Nevertheless we believe that several additional aspects must still be investigated:

- Can we provide general rules on how to specify data "completely" in the sense, that it carries its own metadata? Do we need common approaches (common ontologies, data models, etc.) for global, regional, local levels?
- What about the integration of service metadata? Does it make sense to treat it the same way as metadata for spatial data, e.g. by creating "self describing" services?
- The view concept provides a very flexible way of handling spatial data and metadata. How does a data server support this flexibility? Is it useful to generate all metadata attributes during runtime (when the information is requested) or should we better preprocess data sets and store/offer redundant information (e.g. bounding rectangles)? How far can the data set be revealed to a search engine (in order to protect still the valuable spatial data)?
- How can a search engine be implemented to support the offered flexibility? Is ASCII-based data description (as it is currently used in international standardization efforts for SDI) really suitable to implement a search engine providing adequate performance for data, service retrieval and access?

5. ACKNOWLEDGEMENTS

The authors like to thank Joseph Dorfschmid, Claude Eisenhut and Hans Rudolf Gnägi for the fruitful discussions and the valuable advices on the usage of INTERLIS and its specific features. We thank also the administrations of the city of Zurich, of the Swiss Federal Office of Topography, and of the two official surveying offices in Hesse and Bavaria (Germany) for kindly providing example data sets and the corresponding descriptions of the

data models. We owe special thanks to the interdepartmental GIS Coordination Group KOGIS (www.kogis.ch) of the Swiss Federal Administration due to their continuous support of our research.

6. REFERENCES

[1] Nebert, D.N. (Ed.), Developing Spatial Data Infrastructures: The SDI Cookbook, V1.1, 2001

[2] Open GIS Consortium, www.opengis.org

[3] Sowa, John F., Knowledge Representation – Logical, Philosophical, and Computational Foundations, Brooks/Cole, Pacific Grove, 2000

[4] Sowa, John F., Ontology, Metadata, and Semiotics, in: B. Ganter & G.W. Mineau (Eds.): Conceptual Structures: Logical, Linguistic, and Computational Issues, Lecture Notes in AI#1867, Springer-Verlag, Berlin, 2000, pp.55-81

[5] Gruber, T. R., A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993

[6] Gruber, T. R., Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, appeared in an edited collection by Nicola Guarino

[7] Frank, A.U., Tiers of ontology and consistency constraints in geographic information systems, in IJGIS, 15 (7), pp: 667-678, 2001

[8] Hakimpour, F., Geppert, A., Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach, FOIS'01, October 17-19 2001, Ogunquit, Maine, USA

[9] INTERLIS, www.interlis.ch

[10] Göbel, S., Lutze, K., Giger, Ch., InGeoForum - Information and Co-operation Forum for Geospatial Data, Published in: J. STROBL and C. BEST (Eds.), 1998: Proceedings of the Earth Observation & Geo-Spatial Web and Internet Workshop '98 = Salzburger Geographische Materialien, Volume 27. Instituts für Geographie der Universität Salzburg

[11] International Organization for Standardization, Technical Committee 211 Geographic Information – Geomatics, ISO/TC211, www.isotc211.org