SERVICES FOR DATA INTEGRATION

Catharina Riedemann[1]* and Christian Timm[2]

[1] Institute for Geoinformatics, University of Münster, Robert-Koch-Str. 26-28, D-48149 Münster*
Email: riedemann@ifgi.uni-muenster.de
[2] ICF Informationstechnik Consulting Forschung GmbH, Lahnstr. 1, D-48145 Münster
Email: timm@hansaluftbild.de

ABSTRACT

The fact that many decisions need a combination of information sources makes easy integration of geospatial data an important data usability issue. Our vision is to achieve automated just-in-time integration. As a foundation, we present a system architecture with distributed data and services. Existing and evolving standards and technologies fitting into this architecture are presented along with their scope and shortcomings. A major point is the appropriate definition of data and operation semantics. Further research is needed here to make the automatic formation of service chains for data integration possible.

Keywords: Data integration, System architecture, ISO, OGC, Metadata, Ontology, Semantics, Data wrapping, Service chain, Web services

1 INTRODUCTION

Many decisions are dependent on information that can only be obtained by combining various data sources. Finding locations for industrial plants, for example, requires topographic, infrastructure, environmental, and demographic data. Integrating these data so that they can be queried seamlessly for extracting the information wanted is not a straightforward task. It needs a good understanding of the question that shall be answered, the data that contains information supporting the answer, and the systems that are capable of delivering this information. But data users' business is to solve problems and make decisions, not to handle and process data. For them, data usability means that data can be easily integrated to reveal inherent information without demanding technical expertise.

1.1 Motivation

When the amount of valuable data captured in isolated systems grew and technical progress made it possible to link these isolated systems, the wish to exchange and share data arose and became more and more important. But the combination of data - done on different levels that we subsume under the term data integration (Bömelburg, 1996) - requires resolving heterogeneities, which still poses research questions (Koch, 2001). Database people have dealt extensively with architecture alternatives (Dadam, 1996). However, they mostly focused on the distribution of data and metadata and neglected distributed operations (Conrad, 1997). The distribution of operations plays an important role in the geospatial domain, because the additional spatial component requires many more operations than traditional database applications. Albrecht's collection of universal geospatial operations (Albrecht, 1996) contains many of them. Different specialized versions of such operations are likely to be distributed in the Internet rather than being compiled in a single system. Much research has been and is still done in order to develop and improve them, for example in the areas of feature matching (Gabay & Doytsher, 1995; Walter & Fritsch, 1997; Sester, Anders & Walter, 1998) and generalization (Weibel & Jones, 1998; Lamy, Ruas, Demazeau, Jackson, Mackaness & Weibel, 1999; Cecconi & Weibel, 2001).

Integrating datasets today does not follow standard steps or procedures and is mostly done manually with static results: it is a craft. This leads to costly isolated case solutions, that are poorly or not not documented at all, and not transferable to other situations. The repeated effort makes this approach even more expensive. Furthermore, the result of integrating existing datasets is mostly a new integrated dataset, which means that the original data are duplicated. Such an additional dataset is difficult to update when the origi-

nal data changes. These difficulties often lead to datasets not being used at all or at least not as often as possible.

In search of solutions for these problems, the present work discusses data integration in the light of recent developments in information technology that promise a more dynamic and automated approach.

1.2 Our vision

Our vision is to enhance data usability by promoting easier access to integrated and current data. First, we want to simplify the technical work presently demanded from data users by automating the integration procedure. Second, we seek a way to integrate just-in-time data (dynamic data integration on-the-fly) and avoid producing persistent additional integrated datasets. This will eliminate the updating problem, because at any time the most current original data will be accessed.

1.3 Technology for realizing this vision

The technique to achieve this is by wrapping data in suitable services. Services are focused on answers (information) instead of delivering data. We envision an Internet environment with data and service providers, where services can be coupled with data on demand to form a "wrapped object" exposing the desired information. Promising technologies subsumed under the term "Web services" are available and under development (for example Web Services Description Language=WSDL, Universal Description, Discovery and Integration=UDDI, and Simple Objects Access Protocol=SOAP) that help to build the necessary infrastructure. The challenge is to ensure that only suitable (with respect to the user's task that is supported by the information) and permissible (with respect to the data) operations are performed on the data. As data integration often involves more than one step, the correct interconnection of services is another issue. With both issues the exposure and evaluation of task, data and operation semantics play a crucial role.

The service idea has another appealing effect. Splitting data and software functionality into smaller units (small pieces of information instead of a whole dataset, respectively isolated operations instead of a whole GIS) provides the technical prerequisite for the business model of pay per use. The user only pays for what he really needs and not for the additional data and functionality that comes with monolithic datasets and software packages. Payment systems like the Web Pricing and Ordering Service (Wagner, Gabriel & Holtkamp, 2002) are based on these services.

2 SERVICE ARCHITECTURE FOR DATA INTEGRATION

This section describes the architecture depicted in Figure 1. An Internet environment is shown. It focuses on data integration, but it could be extended on the one hand by a data cataloguing and search functionality and on the other hand by a search facility for analytical operations. In this example we assume that from all available distributed data sources suitable ones have already been identified and that the needed analysis functions are already at hand.

To make the technical description more tangible, we will assume a user task, which is to determine a bicycle route that links two cities and leads through forest and meadow areas. Two datasets, topography (in German) and street network (in English), shall be integrated for joint analysis by a German user.

2.1 Integration steps

At first, a common model must be defined on which the datasets are mapped. This is like a database view: the original data are not changed, but shall only be transiently presented in this model.

In a second step, the datasets are checked concerning their correspondence with the common model. By comparing the common model to the original models the need for transformation is determined. Two categories of transformation operations can be differentiated: operations performed on isolated datasets (for example select area, normal face in the figure) and operations needing all datasets as input (for example match geometries, bold face in the figure).

A chain of transformation steps is formed for each dataset, converging to the integration operations of the second category.

The chain of operations is used as a middleware when loading the data. Thereby, the data are transformed on-the-fly and the two datasets appear as one integrated source in the common model, where they can finally be analyzed.

For our example of determining a bicycle route this means that in the topographic data the relevant area has to be selected as well as the needed themes (cities, forests, and meadows). The street network only contains the thematic information needed, so selection is restricted to the relevant area. But for this dataset, the spatial reference system has to be converted into Gauß-Krüger and texts have to be translated into German to accommodate to the German user.

2.2 Architecture components

The whole process is controlled by an application that consists of various parts specialized in tasks related to the steps described in the preceding section.

The part mentioned first, serving the definition of a common model, consists of a dialogue with the user and does not refer to any other component of the architecture. Data analysis, the part mentioned last, is done as usual; the diversity of data sources is transparent to it. The architecturally interesting parts lie in between and are explained in the following.

2.2.1 Metadata

The analysis of datasets needs information about their contents and structure. Extracting this directly from the datasets would take a long time and would therefore be inconvenient. Hence, for each dataset metadata are provided, which might but need not be stored together with the data (for example in the same database). The content metadata describe how a dataset sees the world, which objects it knows and how objects relate to each other. This is what object catalogues of various standards like the Digital Geographic Information Exchange Standard (DIGEST) or the German Authoritative Topographic-Cartographic Information System (ATKIS) do. In other words, at least part of those metadata can be called an "explicit specification of a conceptualization" – which is Gruber's definition of an ontology (Gruber, 1993). The ontology provides information about the meaning of data, in other words their semantics.

These metadata are only useful if they represent the actual state of the dataset. It is therefore important that they are automatically synchronized with the dataset to reflect possible changes, for example the insertion of a new feature class. And vice versa, the ontology can be used to create the initial structure of the dataset.

To make the metadata easily accessible regardless of their internal implementation, they shall be wrapped in services just like the data. Standardized interfaces enable access for components adhering to this standard.

2.2.2 Service chain

The services available for assembling a data-integration service-chain are distributed throughout the Internet. To make them detectable, they are registered in catalogues (service registries), which are similar to the metadata catalogues used for finding datasets (not part of the architecture in Figure 1). Again, metadata are needed, but this time they describe operations. These metadata can also be referred to as ontologies. Figure 1 shows the metadata for the coordinate transformation service only. The service registry (in fact there might be more than one, or a meta-registry referencing other registries) offers search functionalities that the assembling component of the application deploys. But it needs more intelligence to arrange a suitable service chain. For example, an execution order has to be established that considers results and optimizes performance.

2.2.3 Data loading

The services of the chain are instantiated to represent the "glasses" through which the data are viewed. As similarly described for the metadata in Section 2.2.1, the data are not directly accessed, but through services that constitute a standard interface. The integration services are compliant with this standard and consequently can access the data without knowing about their internal implementation.

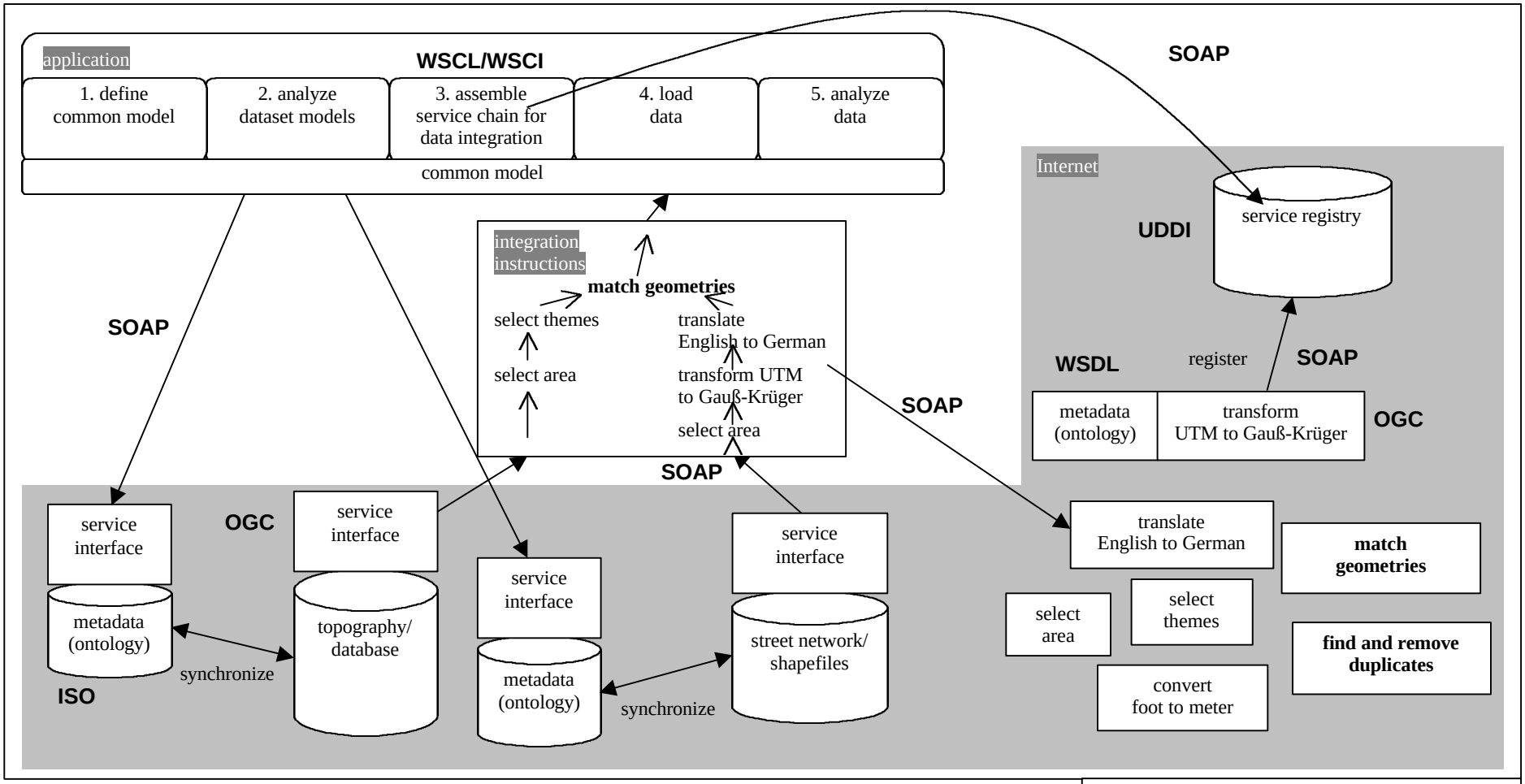

Figure 1. Internet-Based Service Architecture for Data Integration

3 TECHNOLOGIES AND STANDARDS

Having shown the architecture, the next interesting question is what can be realized with the means existing and where can we find gaps in today's developing or available technologies. This work does not intend to give a complete overview. We focus on major technologies and standards. Mentioning a specific standard as an example does neither mean that it is the only one for this purpose nor that it is the most suitable one and will "survive" the ongoing standardization process.

3.1 Metadata

Metadata describing datasets and services are required. They are the clue to what integration steps must be performed and to the services that can do it.

The International Organization for Standardization (ISO) has developed a draft standard for describing digital geospatial data (International Organization for Standardization, 2001a). It is intended to "serve the full range of metadata applications", including the "use of digital data" (International Organization for Standardization, 2001a). Information necessary for data integration, like spatial reference systems, feature catalogues, units of measurement, and of course a pointer to the online resource is indeed contained in the schema. The OpenGIS Consortium (OGC), which leaves the definition of metadata elements to the International Organization for Standardization (ISO) but deals with their relationship to data elements and their uses, also envisions the processing and use of data as a key metadata issue. "Find information that may assist a process that is about to be attempted (such as conflation, generalization, symbolization, projection, ...)", where the processes can be seen as the building bricks of an integration procedure, and "using certain metadata to exploit the related feature data in the context of a specific mathematical model" are two scenarios in the abstract specification for metadata (Kottmann, 1999b).

It remains open to question whether all the information necessary for data integration can be mapped onto the ISO standard. Another issue is how the (extensive) metadata will be collected and fed into the computer. Only tools that automatically "harvest" as much metadata as possible from the datasets can sensibly do this. The Environmental Systems Research Institute (ESRI), for example, provided such a tool very early. Their latest developments in this area aim at automatic synchronization (Environmental Systems Research Institute, 2001), which is essential to keep the metadata up-to-date. The question of automatically deriving and updating metadata is especially interesting if datasets are stored in a database. The database maintains its own metadata and creating another metadata schema means creating redundancies.

ISO also "support[s] the development of a service catalogue through the definition of service metadata" (International Organization for Standardization, 2001b). At this point, notice that the focus lies only on parameters. Further operation descriptions are optional and as character strings difficult to use for automatic interpretation. This issue will be picked up in Section 3.3, where other standardization efforts for service descriptions are presented.

3.2 Geospatial services

The idea of geospatial services is promoted by both ISO and OGC. It is about decomposing monolithic geospatial information systems (GIS) into smaller building blocks. Researchers such as (Albrecht, 1996) have considered the idea of geospatial services from a content and user interaction point of view. ISO and OGC work on the technical basis of standardizing interfaces for "pieces of software that can play in different operating systems, networks and application frameworks". This results in interoperable services, which can be (dynamically) assembled "in unpredictable combinations" to "create whole applications from reusable software parts" (Kottmann, 1999c).

Services for data integration are mostly the usual geoprocessing services similar to those specified by OGC. Examples are coordinate transformations (Open GIS Consortium, 2001) or topological operators (Kottmann, 1999a): the equal operator is one possible way of finding duplicate features in two datasets. It is questionable if all the functions needed for data integration will be specified by OGC, at least in the near future. Matching geometries is for example a complicated, seldom-needed and poorly-understood task, in comparison with tasks such as coordinate transformation, which does not make it a high priority.

There are data integration tasks that are not specific to geospatial data, like converting units of measurement. This emphasizes the fact that geospatial processing should be compatible with and integrated into the general world of information technology (IT). The next section presents such a general IT development and discusses its applicability to GIS.

3.3 Web services

Web services are first of all services. As such they endow data with computational elements making the needed information accessible and exposing it in the desired form, which is known as wrapping. For many applications a variety of services must be combined (chained), because one service only performs a limited set of tasks (Kottmann, 1999c). Ideally, there should be a pool of interoperable services that can be put together in any combination desired. Here the Web, representing a large distributed repository not only for data, but also for software, comes into play. Existing and upcoming standards based on the Extensible Markup Language (XML) and standard protocols such as the Hypertext Transfer Protocol (HTTP) ensure interoperability. A suite of several such standards is forming the specification of what is called Web services. The capabilities of a Web service are described in the Web Services Description Language (WSDL) (Chinnici, Gudgin, Moreau & Weerawarana, 2002). The availability of services is announced in a registry following Universal Description, Discovery and Integration (UDDI) (Bellwood, Clément, Ehnebuske, Hately, Hondo, Husband et al., 2002). The communication among clients, registries and servers is done via the Simple Object Access Protocol (SOAP) (Mitra, 2002). This suite is complemented by workflow, transaction, security, and user interface description standards. The Web Services Conversation Language (WSCL) (Banerji, Bartolini, Beringer, Chopella, Govindarajan, Karp et al., 2002) and the Web Service Choreography Interface (WSCI) (Arkin, Askary, Fordin, Jekeli, Kawaguchi, Orchard et al., 2002) are examples for the task of controlling the interactions between various Web services. Both work together with WSDL.

The main challenge for using Web services as building blocks for automated data integration is the dynamic combination of single services into a suitable chain. The currently available techniques for exposing (UDDI), describing (WSDL, ISO service metadata) and chaining (WSCL, ISO service chaining) Web services all have deficiencies in the area of handling semantic issues associated with Web services. Today the description of Web services in WSDL and UDDI is limited to the straight syntactical level. For automated chaining and thus automated data-integration, semantic information about the Web services must be available. For example, a service transforming feet to meters is described with WSDL as a service, which expects a number and returns another number. This is enough for an environment knowing the services that have to be used. In a scenario where unknown services are to be dynamically connected on demand this information is not sufficient. The component calling must be informed about the expected and returned units of measurement in a form that can be interpreted by a machine. This problem has been described by (Frank & Kuhn, 1995).

Apart from the necessary improvements concerning the service side, the determining whether a Web service is suitable for solving a specific task and the question of how it can be integrated into a chain are also unsolved problems.

4 CONCLUSIONS AND OUTLOOK

Automating data integration will free the user from difficult and lengthy pre-processing steps. Instead of spending time on data preparation, information can immediately be accessed, as it is exposed, in the form needed. But apart from saving time this approach also prevents errors. A metadata (ontology) driven approach ensures that only permitted operations are performed on data.

Fully-automated just-in-time data-integration using distributed data and software is a vision. But with the development of technologies like Web services, which provide the basic functionality for this scenario, this vision is coming closer to reality. In the preceding sections we have shown that a major obstacle for making the scenario really work is the lack of exposure especially of operation semantics. Sensible automatic service chaining is only possible if operation descriptions go beyond mere technical interface definitions in the form of signatures comprising input and output parameters. This had already been stated some time ago for the isolated application of services, and it has subsequently proved to be highly relevant for (automatically) chaining services. The first steps have been made to address the semantic shortcomings by enriching the

service descriptions (Janowicz, Kuhn & Riedemann, 2002), but further research is needed into case studies and prototypical implementations to fully understand data integration issues and come up with solutions that the current standards do not offer. New technologies like the Resource Description Framework (RDF) (World Wide Web Consortium, n.d.) and the DARPA Agent Markup Language Web Services (DAML-S) (Ankolenkar, Burstein, Hobbs, Lassila, Martin, McDermott et al., 2002; DAML.org, n.d.) must be examined.

5 ACKNOWLEDGEMENTS

The work reported here is supported by a grant from the European Commission under grant number IST-2001-34386 (BRIDGE-IT). The authors gratefully acknowledge valuable discussions with Werner Kuhn and Krzysztof Janowicz about Web services and semantics.

6 REFERENCES

Albrecht, J. H. (1996) *Universal GIS Operations - A Task-Oriented Systematization of Data Structure-Independent GIS Functionality Leading Towards a Geographic Modelling Language*, PhD thesis, University of Vechta.

Ankolenkar, A., Burstein, M., Hobbs, J. R., Lassila, O., Martin, D. L., McDermott, D. et al. (2002) DAML-S: Web Service Description for the Semantic Web. *The First International Semantic Web Conference (ISWC)*. Retrieved February 14, 2003 from the DAML.org website: http://www.daml.org/services/ISWC2002-DAMLS.pdf

Arkin, A., Askary, S., Fordin, S., Jekeli, W., Kawaguchi, K., Orchard, D. et al. (2002). Web Services Choreography Interface (WSCI) 1.0. Retrieved February 14, 2003 from the World Wide Web: http://www.w3.org/TR/wsci/

Banerji, A., Bartolini, C., Beringer, D., Chopella, V., Govindarajan, K., Karp, A. et al. (2002). Web Services Conversation Language (WSCL) 1.0. Retrieved February 14, 2003 from the World Wide Web: http://www.w3.org/TR/wscl10/

Bellwood, T., Clément, L., Ehnebuske, D., Hately, A., Hondo, M., Husband, Y. L. et al. (2002) *UDDI Version 3.0. Published Specification, 19 July 2002*, Retrieved February 14, 2003 from the UDDI.org website: http://www.uddi.org/pubs/uddi-v3.00-published-20020719.pdf

Bömelburg, J. (1996) ATKIS-Datenintegration. *Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung. 3. AdV-Symposium ATKIS*, Koblenz (pp. 199-204).

Cecconi, A. & Weibel, R. (2001) Map Generalization for On-demand Mapping. *GIM International 15*(5), 12-15.

Chinnici, R., Gudgin, M., Moreau, J.-J. & Weerawarana, S. (2002). Web Services Description Language (WSDL) 1.2. Retrieved February 14, 2003 from the World Wide Web: http://www.w3.org/TR/wsdl12/

Conrad, S. (1997) *Föderierte Datenbanksysteme. Konzepte der Datenintegration*, Berlin: Springer

Dadam, P. (1996) *Verteilte Datenbanken und Client/Server-Systeme. Grundlagen, Konzepte und Realisierungsformen*, Berlin: Springer

DAML.org (n.d.). DAML-S 0.7 Draft Release. Retrieved February 14, 2003 from the World Wide Web: http://www.daml.org/services/daml-s/0.7/

Environmental Systems Research Institute (2001) *Creating a Custom Metadata Synchronizer*, Redlands: Environmental Systems Research Institute (ESRI), Retrieved February 14, 2003 from the ESRI website: http://arconline.esri.com/arconline/whitepapers/ao_/metadata.pdf

Frank, A. U. & Kuhn, W. (1995) Specifying Open GIS with Functional Languages. *Advances in Spatial Databases, 4th International Symposium*, (pp. 184-195)Portland, ME, USA.

Gabay, Y. & Doytsher, Y. (1995) Automatic Feature Correction in Merging Line Maps. *ACSM/ASPRS Annual Convention & Exposition Technical Papers*, (pp. 404-411), Charlotte, North Carolina.

Gruber, T. R. (1993) Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Workshop on Formal Ontology*, Padua, Italy. Retrieved February 14, 2003 from the Stanford University website: ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-93-04.ps

International Organization for Standardization (2001a) *Geographic Information - Metadata. Draft International Standard ISO/DIS 19115*, Retrieved February 14, 2003 from the ISO website: http://www.isotc211.org/protdoc/DIS/ISO_DIS_19115_(E).pdf (restricted)

International Organization for Standardization (2001b) *Geographic Information - Services*, Retrieved February 14, 2003 from the ISO website: http://www.isotc211.org/protdoc/DIS/ISO_DIS_19119_(E).pdf (restricted)

Janowicz, K., Kuhn, W. & Riedemann, C. (2002) XMeta: Wie aus Geodaten Geodienste werden. *Angewandte Geographische Informationsverarbeitung XIV. Beiträge zum AGIT-Symposium*(pp. 206-211), Salzburg .

Koch, C. (2001) *Data Integration against Multiple Evolving Autonomous Schemata*, PhD thesis, Technical University (TU) Vienna. Retrieved February 14, 2003 from the TU Vienna website: http://www.dbai.tuwien.ac.at/staff/koch/download/thesis_20010516_1500_final.pdf

Kottmann, C. (Ed.) (1999a) *The OpenGIS™ Abstract Specification. Topic 1: Feature Geometry*, Version 4, Wayland, Massachusetts: Open GIS Consortium (OGC), Retrieved February 14, 2003 from the OGC website: http://www.opengis.org/public/abstract/99-101.pdf

Kottmann, C. (Ed.) (1999b) *The OpenGIS™ Abstract Specification. Topic 11: Metadata*, Version 5, Wayland, Massachusetts: Open GIS Consortium (OGC), Retrieved February 14, 2003 from the OGC website: http://www.opengis.org/techno/abstract/01-111.pdf

Kottmann, C. (Ed.) (1999c) *The OpenGIS™ Abstract Specification. Topic 13: Catalog Services*, Version 4, Wayland, Massachusetts: Open GIS Consortium (OGC), Retrieved February 14, 2003 from the OGC website: http://www.opengis.org/public/abstract/99-113.pdf

Lamy, S., Ruas, A., Demazeau, Y., Jackson, M., Mackaness, W. A. & Weibel, R. (1999) The Application of Agents in Automated Map Generalisation. *19th Int. Cartographic Conference*, Ottawa, Canada (pp. 160-169). Retrieved February 14, 2003 from the AGENT Project website: http://agent.ign.fr/public/ica/paper.pdf

Mitra, N. (2002). Simple Object Access Protocol (SOAP) 1.2 Part 0: Primer. Retrieved February 14, 2003 from the World Wide Web: http://www.w3.org/TR/soap12-part0/

Open GIS Consortium (2001) *OpenGIS® Coordinate Transformation Services Implementation Specification*, Version 1.0, Wayland, Massachusetts: Open GIS Consortium (OGC), Retrieved February 14, 2003 from the OGC website: http://www.opengis.org/techno/specs/01-009.pdf

Sester, M., Anders, K.-H. & Walter, V. (1998) Linking Objects of Different Spatial Data Sets by Integration and Aggregation. *Geoinformatica 2*(4), 335-357. Retrieved February 14, 2003 from the Kluwer website: http://www.wkap.nl/article.pdf?193781

Wagner, R. M., Gabriel, P. & Holtkamp, B. (2002) GIS Meets E-Business. First Steps towards a General Architecture for Geo-data Markets. *GIM 5*(1), 24-27. Retrieved February 14, 2003 from the Geoinformatics website: http://www.geoinformatics.com/issueonline/issues/2002/01_janfeb_2002/pdf/24_27_gise.pdf

Walter, V. & Fritsch, D. (1997) Matching Strategies for Integration of Spatial Data from Different Sources. *International Workshop on Dynamic and Multi-Dimensional GIS*, (pp. 215-228), Hong Kong.

Weibel, R. & Jones, C. B. (1998) Computational Perspectives on Map Generalization. *Geoinformatica 2*(4), 307-314.

World Wide Web Consortium (n.d.). Resource Description Framework (RDF). Retrieved February 14, 2003 from the World Wide Web: http://www.w3.org/RDF/