Experiences with Metadata

Sabine Timpf, Martin Raubal and Werner Kuhn
Department of Geoinformation
Technical University Vienna, E 127.1
Gusshausstr. 27-29
A-1040 Vienna
Tel.: +43 1 58801 - {3791, 3788, 3790}
Fax.: +43 1 504 3535
{timpf,raubal,kuhn}@geoinfo.tuwien.ac.at
http://www.geoinfo.tuwien.ac.at/

Abstract

The need to share and integrate spatial data has spurred an interest in metadata. This paper documents the acquisition and modeling of metadata from eleven digital geodata sources in Austria. It shows how the information was modeled according to the proposed CEN standard on metadata, how it was encoded in a database, and what problems were encountered during these processes. The paper concludes with a discussion of recent developments around metadata and of the option to make meta-databases available on the world-wide web.

Keywords

metadata, standards, OpenGIS, meta-databases, WWW

1. Motivation

The topic of metadata has recently received considerable attention [Blott, and Vckovski, 1995; Dorf, and Scholten, 1993; Fisher, 1993; Strobl, 1995]. While many discussions addressed conceptual, architectural, and organizational requirements, practical experience with producing metadata has scarcely been documented. This might be caused by the fact that not a lot of meta-databases exist because metadata is expensive and hard to collect. Also, the possibility of a distribution of geodata to unknown users over the network has only arisen in the last few years due to the technological advancement. This contributed to a shift in attitude towards sharing data collections. More and more users outside the traditional spatial disciplines need spatial data. Data providers need to tell users what they have and what it can be used for. Metadata is destined for this purpose.

Metadata are 'data about data'. For companies working with spatial data, good documentation of the datasets becomes extremely important to make sure that they can still be used after changes of employees, software and hardware [Strobl, 1995]. Metadata can facilitate research on the environment in Europe [Dorf, and Scholten, 1993]. Metadata are also necessary to insure multiple usage of datasets [Frank, 1992]: Spatial data are being collected everywhere and could often be of use to others if they only knew of their existence. Meta-databases are one possible solution to this dilemma.

Metadata describe spatial datasets in a way that one can infer the usability of a specific dataset for a specific task. Some important criteria for the use of a dataset are:

- reference systems and area covered by the dataset
- currency of the dataset
- quality parameters such as positional, thematic and temporal accuracy
- administrative metadata.

The metadata is either gathered during the data collection process itself or at some later time. Generating metadata later requires considerable effort and not all the information might be available.

This paper reports on the experiences that surveying engineering students at the Technical University of Vienna made with collecting and describing metadata and entering them into a database. The metadata was collected from eleven different sources of geodata in Austria, coming from the areas of geodesy, geophysics, hydrology, and geomarketing. The metadata was described using the proposed standard on metadata from CEN TC 287 [CEN, 1995]. The proposed CEN standard defines a minimum set of metadata that should be provided by data suppliers. We used the relational database Microsoft Access 2.0 to implement a metadata repository.

After a short introduction to related metadata work (chapter 2), we introduce the CEN metadata standard (chapter 3) and explain how we applied it to the geodata sources (chapter 4). We review the standard in its present proposed form (chapter 5) and describe how we implemented the meta-database in Microsoft Access 2.0 (chapter 6). Finally, we discuss the issue of standardizing metadata and its possible future (chapter 7).

2. Metadata projects

The first organization to consider data about data was the FGDC (Federal Geographic Data Committee) with its Spatial Data Transfer Standard (SDTS). This started the discussion on metadata and its organization.

In the United States, the national spatial data infrastructure (NSDI) encourages standards and information interchange.

The National Geospatial Data Clearinghouse links providers, managers, and users of information in a large network [FGDC, 1994]. In this system, users can search for metadata on the data they need. All federal agencies are required to make their data available to other agencies and to the public.

The Alexandria Digital Library[*] is a library for spatially indexed material. The library will enable users who are distributed over the network to access the information in the space they want. It will also be made available over the network.

In Europe, the MEGRIN (Multipurpose European Ground-Related Information Network) is an initiative of the Comite Europeen des Responsables de la Cartographie Officielle (CERCO). It provides an information system with metadata on data-sources of the members of CERCO [Salgé, Smith, and Ahonen, 1992]. The information system is currently made available over the network.

3. The proposed metadata standard

The proposed European Standard on metadata (Geographic Information - Data Description - Metadata) has been prepared by the Technical Committee (TC) 287 of the European Committee for Standardization (CEN). Currently CEN TC287 is in the process of soliciting comments on the draft of the metadata-standard.

The standard defines a conceptual schema for metadata, based on two related standards: the proposed standard for quality and the proposed standard for positioning. The main reason for developing this metadata standard is to encourage the widespread use of geographic information. It is explicitly stated that the standard is not concerned with implementation details and therefore the construction of meta-databases. This is also made

---

[*] http://alexandria.sdc.ucsb.edu/

clear in the choice of EXPRESS as the language for the formal definition of entities. EXPRESS is a language for defining an information model, not a database model.

EXPRESS is the data description language of the Standard for the Exchange of Product Data (STEP), which has been developed by the International Standards Organization [ISO, 1992]. EXPRESS has been selected by CEN TC287 as the standard for the exchange of geoinformation. EXPRESS defines entities and relations in schemas. A schema is a context, which contains several entities and their relations as well as rules for their interaction. For example the context of metadata forms a schema, the context organisation forms a second one. Schemas can be used in other schemas. This avoids the redefinition of entities. For example the schema *organisation* is used in the schema *metadata* for the definition of the organization that manages or supplies the metadata. EXPRESS-G illustrates the EXPRESS definitions. Efforts are made to derive the EXPRESS-G diagram directly from the EXPRESS language.

The proposed standard consists of six parts. Each of the six parts is described below. We give an example from the standard for the verbal description, the EXPRESS-G description, the tabular description, and the EXPRESS description of the item Organisation.

3.1 Introduction to the proposed standard

The first three chapters of the document contain the scope of the project, references to other standards or draft standards, and definitions.

3.2 Verbal description

This most detailed part of the standard contains verbal descriptions of the metadata (see table 1), including EXPRESS-G graphics of each group of metadata (see figure 1).

Organisation and organisation role
Organisation name - the name of the organisation
Abbreviated organisation name - the short name of the organisation
Organisation address - the postal address, telephone, telefax number, electronic mail address of the organisation
Role - the responsibility of the organisation in relation to the dataset, for example, the creator, owner, administrator or distributor of the dataset. An organisation shall have one or more roles
Alternative organisation name - another name of the organisation which is either in the same language or another language
Function of the organization - description of the overall role of the organisation.

Table 1: Verbal description of the metadata

The groups are:

- Dataset Identification,
- Dataset Overview,
- Dataset Quality Parameters,
- Spatial Reference System,
- Geographic and Temporal Extent,
- Data Definition,
- Classification,
- Administrative Metadata, and
- Metadata Reference.

The metadata items are defined by a name and a short description of what is meant by this name. For example, the name *Role* means the *responsibility of the organisation in relation to the dataset, for example, the creator, owner, administrator or distributor of the dataset. An organisation shall have one or more roles* (see table 1).

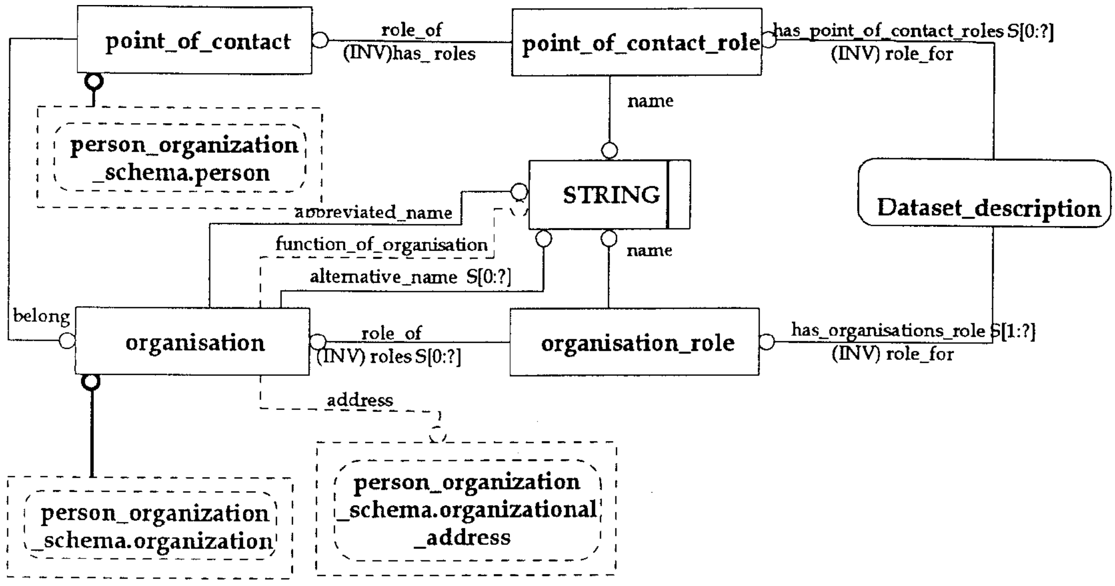

Figure 1: EXPRESS-G schema of organisation

The EXPRESS-G diagram illustrates the metadata items and their relations graphically. For example, the *role of* the *organisation* can be found in the field *organisation_role*.

The EXPRESS-G schema also shows sub- and supertypes (dotted lines). For example the item *organisation address* needs to be represented as defined in the item organization_address in the schema person_organization_schema.

3.3 Annex A: Metadata Table

Annex A is a table with the metadata items including constraints, cardinality, and types (see table 2). Types may be string, numeric, picture, enumeration, address etc. Constraints define if the metadata item is mandatory or optional. Cardinality defines the allowed number of occurrences of an item. For example, the Alternative organisation name is an optional item of type string, which can occur N-times in a dataset.

Administrative metadata				
Organisation and organisation role		*Con-straints*	*Card-inality*	
• Organisation name	The name of the organisation	M	1	string
• Abbreviated organisation name	The short name of the organisation	M	1	string
• Organisation address	The postal address, telephone, telefax number, e-mail address of the organisation	M	1	address
• Role	The responsibility of the organisation in relation to the dataset, for example, the creator, owner, administrator or distributor of the dataset.	M	N	enumeration
• Alternative organisation name	Another name of the organisation which is either in the same language or another language	O	N	string
• Function of the organization	Description of the overall role of the organisation.	O	1	string

Table 2: Tabular description of the metadata

3.4 Annex B: EXPRESS

In this part a formal description of the metadata items using the EXPRESS language is given (see table 3).

```
ENTITY Organisation
SUBTYPE OF (organization);
      abbreviated_name            :STRING;
      address                     :organizational_address;
      alternative_name            :OPTIONAL SET OF STRING;
      function_of_organisation    :OPTIONAL STRING;
INVERSE
      roles                       :SET OF Organisation_role FOR role_of;
END_ENTITY;
```

Table 3: The formal description in EXPRESS

In this example the item organisation is defined as an entity in the schema. Its supertype is *organization*, which comes from a different schema. The attributes abbreviated_name, address, alternative_name, and function_of_organisation are defined. The relation *roles* is an inverse of the relation role_of in the entity Organisation_role.

3.5 Annex C: EXPRESS-G

In Annex C, the complete Express-G schema for the standard is given.

3.6 Annex D: Examples

In Annex D, two examples are supplied of how metadata should be described according to the proposed standard. The structure used is the table of metadata from Annex A.

4. The setting

Students in the course "Sources of Geoinformation" at the Technical University of Vienna* described eleven digital spatial data sets using the proposed CEN standard on metadata. The goal was to test the usability of this proposed standard for a variety of datasets and to get experience with modeling and describing metadata.

Datasets for testing the proposed standard came from four different areas: geodesy, geophysics, hydrology and geomarketing. Students worked with the following databases:

- Coordinate database of Austria
- Parcel database of Austria
- Administrative Boundaries of Austria
- Digital cadastral map
- Terrain database
- Database of the multi-purpose city map Vienna
- Database of the city information system Linz
- Leveling and gravity database
- Fresh water wells database
- Water management database - ground water Vienna
- Geocoding database of Vienna.

The data sources were analyzed in two different ways: The necessary information could partly be extracted from flyers, papers, or brochures that were given out by the collecting organizations. More detailed information was collected by interviewing the responsible people at each organization. The second part of the information gathering process often

* http://www.geoinfo.tuwien.ac.at/Department/Courses/GIQ.html (in german)

took considerable time. Data were described by filling in a table taken directly out of the standard (Annex A).

Usually, the contacted persons were very interested in the project and liked to be selected as data providers. However, the students noted that some persons were not very interested in the proposed standard itself. This is, at least partly, due to the fact that the importance of metadata is still underestimated.

When formal descriptions of the datasets had been finished, students entered the metadata into a common database. For this implementation the relational database Microsoft Access 2.0 was used.

Our experience with the proposed standard is given in the next chapter, the implementation is critically reviewed in chapter 6.

5. Comments on the proposed standard

A data description using the proposed CEN standard provides a useful overview over a data set. The students got a generally positive impression of the proposed standard. Major problems were not reported, though some details created obstacles in the processes of modeling or implementing. After describing their datasets, the students prepared individual comments on the proposed standard. As the datasets were very heterogeneous, the results and especially the difficulties encountered were also of a great variety.

5.1 Criteria used

A standard should be clear, comprehensible, consistent, complete, flexible, and simple to use. In other words, it should be easy for the user to apply the standard to a dataset. Only if a standard is easy to use will it be used.

The documentation should be well structured and should give *clear* instructions how to derive metadata from geodata.

It should be *comprehensible*, explaining each step in the definitions, and connecting the different representations.

The various representations of the data description and the use of terminology throughout the document should be *consistent*.

The representations should be *complete* but also complement each other in the description. It is especially important that the examples be complete.

The standard should be *flexible* enough to accommodate different datasets, proveniences, complexities, or cultural differences.

The *ease of use* criterion requires a deep understanding of the process of creating metadata. Ease of use should therefore benefit the most from the experiences made in such a metadata project. It is also the all encompassing criterion in which all other criteria are reflected.

Generally, the standard should clearly point out the purpose of metadata and of meta-databases so that users can see the benefit and be prepared to produce and use meta-databases.

5.2 Application of criteria

- **Structure and clarity**

The description of the proposed metadata standard is not very homogeneous. A verbal description cannot easily give the full information that is detailed in a formal language. This means that the verbal description should be complemented with EXPRESS and EXPRESS-G descriptions. At least, the verbal description should be structured the same way as the EXPRESS description is.

The proposed standard describes how metadata should be defined. It does not give instructions for its use. This is a serious impediment to the use of the standard in its current form. This shortcoming may be alleviated with the forthcoming reference framework and other explanatory documents from CEN TC287.

- **Comprehensibility**

The relations of the different representations are partly made clear in the verbal description, where the connections between definitions and the graphical representations are given. The representations should be placed side by side to clarify their equivalence.

The use of some expressions or examples obscures rather than helps the proposed standard. For instance, the item Spatial_Reference_of_Metadata presumably means: the location, where the metadata will be available, though this is not evident from the textual definition. It is necessary to know EXPRESS or to consult a manual to understand these expressions and their meaning.

Some expressions are not clear, even when provided with examples. For instance, three types of structure primitives are given without further explanations of their meaning. The meaning of these structure primitives needs to be extracted from another standard. Again, this problem may be somewhat reduced once the complete family of standards is available.

When technical terms are used, they have to be defined very carefully. For instance, the parameters required when describing the reference ellipsoid are not clearly stated.

- **Consistency**

The proposed standard has some inconsistencies in type definitions and with missing or redundant relationships. For example, a picture is of type image in one annex and of type string in another.

Inconsistencies between the EXPRESS and the EXPRESS-G model will not occur, when the EXPRESS-G will be directly derived from the EXPRESS model. All other relations between the representations must be checked for consistency.

- **Completeness**

The representations should be complete in themselves but they should also complement each other. This criterion is not yet satisfied.

For instance, some definitions like feature type and attribute type in the introduction are missing and the examples given in Annex D are incomplete.

Definitions should be given in the same document and be as concise as possible. Only if they need to be complex, they may be left out and the user referred to other documents. Currently, missing definitions and the need to look them up at different places is a serious impediment to the ease of use of the standard.

- **Flexibility**

Considering the diversity of the data sets described, the proposed standard proved to be very flexible. It was sometimes hard to find out from the document if or why certain items are mandatory or optional. In these cases, the decision should be in favor of the more flexible solution. For example, the item *support services* is mandatory. But many organizations do not have further services available, so this item should be optional.

- **Ease of use**

From the standard document, it is easy to understand the idea of the standard and its purpose. It is hard to understand the details of applying it to a real dataset. The tricky parts are in the details of the specification language. For example, one needs a better knowledge of EXPRESS to understand all the implications of the formal specifications.

Ease of use also requires simple words and ideas. The user of a standard should be incited to work with it. Simpler terms would often have done a better job. For example, the expression *spatial reference of metadata* presumably means simply the location of the meta-dataset.

6. Implementation

People of various disciplines want to use geodata for different applications. To find out if a certain dataset fits their needs, users have to browse the data. One way to do so, is querying and looking through a database. With a meta-database, users will find the desired information within a shorter period of time. This is a major improvement over looking through all kinds of flyers and brochures.

Another advantage of using a database is the possibility to look for specific sets of metadata. This can be done with queries. In SQL the query 'SELECT dataset_title FROM meta-database_Austria WHERE Administrative_Metadata.Abbreviated_organisation_name = BEV' would retrieve the metadata of all datasets that belong to the Austrian Federal Mapping Agency (BEV, Bundesamt für Eich- und Vermessungswesen).

This project used the relational database Microsoft Access 2.0 to implement the sets of metadata. Entities in the form of tables containing various attributes had to be created and relationships with cardinalities between the tables defined. The dataset title was used as key (see figure 2). While implementing the metadata we found a major problem: EXPRESS uses some concepts of object-orientation, which could not easily be implemented into a relational database. Therefore, the tables do not exactly correspond to entities of the EXPRESS model.

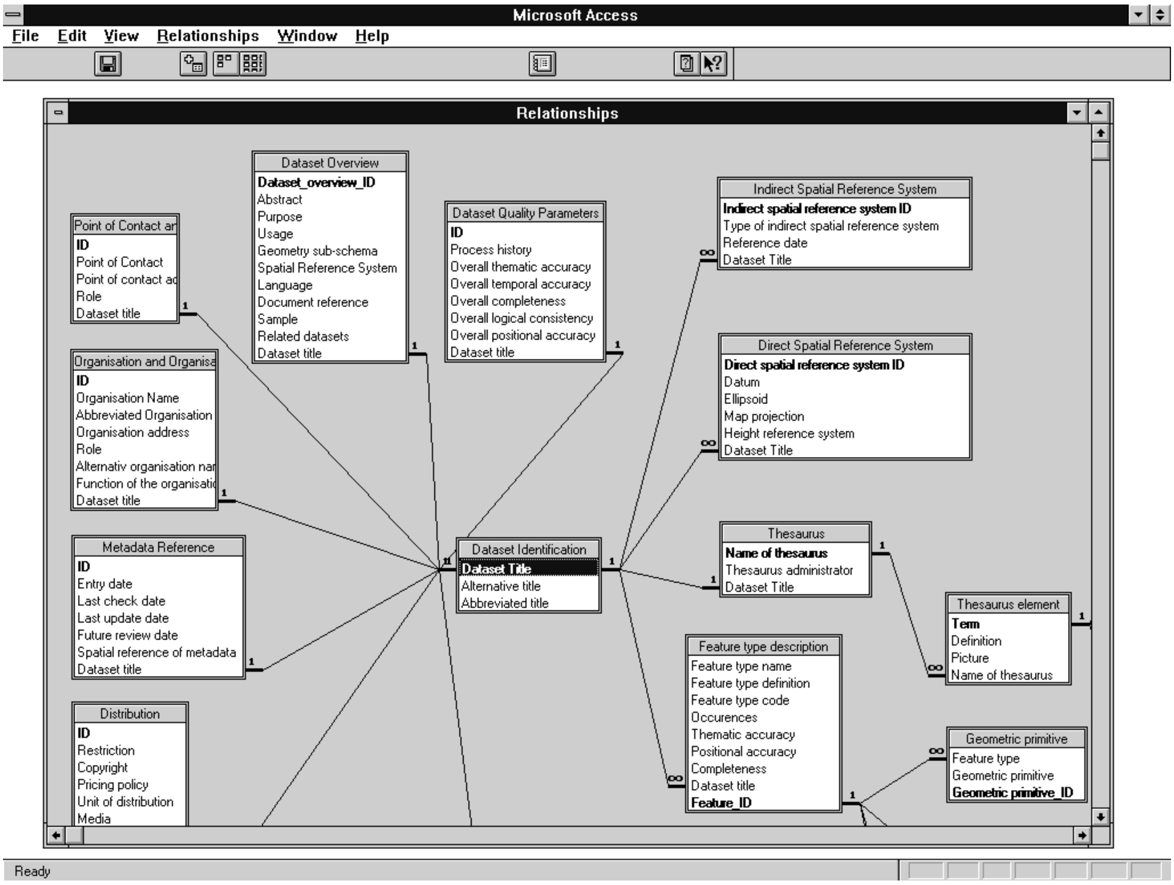

Figure 2: Part of the relationships in Microsoft Access 2.0

It was also not possible to implement data types like lists or sets with Microsoft Access 2.0. These data types had to be implemented by recursive decomposition to conform to the normalization rules. There is a general problem to describe standards with some sort of object-oriented language, when there are basically only relational databases on the market.

Another problem is the impossibility of using two different data types for the same attribute. The proposed standard states that when information is not available, then this shall be described as "no information available". When trying to write this into a column of a table which demands an integer type the user will not succeed. Some students tried to avoid this problem by inserting null values. But null values in databases have to be handled with care and lead to additional problems: what is the value of an undefined value in mathematical expressions, how is an attribute with a null value handled in joins, is there any difference between null values? These questions arise because a null value can be seen either as an unknown or as an arbitrary value [Duerr, and Radermacher, 1990]. These kinds of difficulties have to be overcome before a database like Microsoft Access 2.0 can be used.

The next step of the implementation process will be to connect the meta-database to the World Wide Web (WWW). This should happen as soon as we have revised the metadata and received permission from the data suppliers. Access to the meta-database will be provided through a web-based query form.

7. Discussion and Future perspectives

Metadata is the current approach to supply information to users on the purpose and usability of spatial data collections. Metadata is data about data and helps to decide if the data from specific sources are usable for another purpose than the one they were originally collected for.

Discussions on this topic have led to various definitions of standards and to the creation of some meta-databases. The process of defining metadata is often described abstractly and documented experiences with the collection, definition, and storage of metadata are missing.

In this paper we described how a group of students collected and described several metadata sets. We commented on the currently proposed CEN metadata standard and critically reviewed the modeling and implementation issues that arose.

The standard was found to be easy to understand, but not very easy to use. This is mainly due to the hidden implications of EXPRESS. It is necessary to know EXPRESS to understand the standard as well as to know how to translate EXPRESS models into implementations. An improved structure of the standard with a better distribution of the different representations of metadata (text, formal language, graphics) would contribute to comprehensibility and ease of use.

It is now necessary to find out if such a meta-database is really useful. This could be done by logging who queried the database and matching with those who used or bought the data. One could also prepare a questionnaire to find out why datasets are used or not used.

In general, the project left us with the impression that we have described the data sources in great detail without necessarily enlightening potential users. With all the metadata available in a database, it remains quite unclear whether anybody can use them to assess the fitness for use of the data sources for an application.

In order to assess the fitness of data for a given application, users need more and different information than the current kinds of metadata provide. Most importantly, they need to know what operations are supported by the data [Kuhn, 1994]. Such an approach to metadata represents a major step ahead from the current state of the art. It has been taken by the OpenGIS Consortium [Doyle, 1995] and has, for example, found modeling support in the form of functional languages [Frank, and Kuhn, 1995].

One of the unresolved issues in any metadata approach is to find out where metadata are located. It could require a database on meta-databases, i.e. a meta-meta-database. To some extent, this problem can be solved in the larger context of the World-Wide Web. Geodata are increasingly provided on WWW servers by government agencies (e.g., the USGS) or commercial providers (e.g., ImageNet), though these activities are still in an experimental stage. The main advantages of distribution via the Web are the low cost of distribution, high currency of information, potential market for geodata to millions of users worldwide and 24 hour access [Vincent, 1995]. When geodata are made available on the World Wide Web, they can serve as distributed repositories of their own metadata.

Acknowledgments

The authors thank the many organizations who provided the metadata and all participants of the course "Sources of Geoinformation" for their contributions: Karin Amesberger, Bernhard Gaupmann, Damir Medak, Hartmuth Schachinger, Franziska Silwester, and Ines Stanzer. Special thanks to Joachim Heinzl and Wilhelm Jachs for helping to integrate the different views from the participants.

References

Blott, Stephen, and Vckovski, Andrej. “Accessing Geographical Metafiles through a database storage system.” In *Advances in Spatial Databases (SSD'95) in Portland, ME,* edited by Egenhofer, Max J., and Herring, John R., Springer Verlag, New York, 117-131, 1995.

CEN. *Geographic Information - Data description: Metadata*. CEN/TC 287/WG2, 1995.

Dorf, Josh, and Scholten, Henk J. “Meta-Catalog: a necessary requirement for european environmental research.” In *EGIS'93 in Genoa, Italy,* edited by Janjaap Harts, Ottens, Henk F.L., and Scholten, Henk J., EGIS Foundation, Utrecht, 636-643, 1993.

Doyle, Allen. “Software that plays together - a systems integrator's viewpoint.” *Geo Info Systems* (May 1995).

Duerr, M., and Radermacher, K. *Einsatz von Datenbanksystemen*. Berlin, Heidelberg: Springer-Verlag, 1990.

FGDC. *The national geospatial data clearinghouse*. FGDC, 1994.

Fisher, P.F. “Conveying Object-Based Meta-Information.” In *11th International Conference on Automated Cartography in Minneapolis, MN*, ACSM, 113-122, 1993.

Frank, Andrew U. “Acquiring a digital base map - A theoretical investigation into a form of sharing data.” *URISA Journal* 4 (1 1992): 10 - 23.

Frank, Andrew U., and Kuhn, Werner. “Specifying Open GIS with Functional Languages.” In *Advances in Spatial Databases*, ed. Egenhofer, Max J., and Herring, John R. 184-195. LNCS 951. Springer-Verlag, 1995.

ISO. *The EXPRESS language reference manual*. ISO TC 184, 1992. Draft International Standard ISO/DIS 10303-11.

Kuhn, Werner. “Defining Semantics for Spatial Data Transfers.” In *6th International Symposium on Spatial Data Handling in Edinburgh, UK*, IGU, 973 - 987, 1994.

Salgé, François, Smith, Neil, and Ahonen, Paula. “Towards harmonized geographical data for Europe; MEGRIN and the needs for research.” In *5th International Symposium on Spatial Data Handling in Charleston, SC,* IGU Commission on GIS, 294-302, 1992.

Strobl, Josef. “Grundzüge der Metadatenorganisation für GIS.” *Salzburger Geographische Materialien - Beiträge zur Geographischen Informationswissenschaft (Proceedings AGIT 1995)* (22 1995): 275-286.

Vincent, S. “CyberGIS - Providing Geodata via the World Wide Web.” *GIS Asia Pacific* 1 (5 1995): 28-31.