Data searching by metadata

Henri J. G. L. Aalders
Delft University of Technology
Geodetic Engineering
E-mail: h.j.g.l.aalders@geo.tudelft.nl

Katholieke Universiteit Leuven
Faculty of Engineering

Keywords: Information Technology, GIS, management, Spatial Data Infrastructure, metadata.

Abstract
More and more datasets become available in a digital way and so by the Internet one tries so to explore and achieve datasets being applicable for specific applications. Searching such data requires three steps, exploration (finding what data exists), discovery (interpreting the data, which is found) and exploitation (achieving the actual data from the provider and using it for the intended application).
All these steps require the description and definition of the data preferable in a digital way – called metadata - in order to make them interpretable by computer programs. In order to enable comparison and learn from other experiences this article describes the necessary content of metadata, available standards for metadata and projects undertaken in Europe using metadata for topographic, geological, hydrographic, cadastral, environmental and geo-statistical data.

Introduction

Many organisations start to implement datasets as 'layers', or 'coverages'; some organisations have 'series of layers' for which some elements, as spatial reference, entity and attribute definition and relationships and distribution information are the same. These elements are stored in the dataset only once and 'inherited' by the members in the layers or series of layers. Others specifications are concerned about the differences within a layer and are implementing some information (such as lineage) for components of layers. These components inherit other information from the layers and series of layers descriptions. Dataset containing descriptions of other dataset are called metadatasets. In other words metadata is data about data(sets).

In this way, the contents metadata sets follow the concept of set-theory: any combination of data can be assigned to a set that carries metadata, e.g. metadata description for datasets, dataset series i.e. 'supersets' and data in the datasets i.e. 'subsets'. So, metadata is just data about data or datasets.
Following the concept of set-theory, a geographic dataset defined for the purpose of giving its metadata can contain either several homogeneous geographic datasets or one or more occurrences of geographic entities or any set of attributes or occurrences of geographic entities and/or relationships. As such one could also speak of levels of datasets. Data shall be given to describe data, which occur in the dataset. This is usually done by the description of all **metadata elements**.

Metadata serve a wide range of applications in geographic information as [FGDC, 1998]:

- *organisation and maintenance* of data. As personnel changes and time passes new workers may have little understanding of the dataset's contents and consequently the information about the organisation's data will be lost and the data may loose its value. Complete metadata descriptions of the structure, contents and accuracy of all dataset is an important requirement for database design. Such descriptions also may provide protection for the misuse of the data in a dataset;
- *data exploration, discovery and inventory* as described below. Applications of geographic data often require many different themes of data. The collection of all these data is very expensive and often data from other organisations are used. By making metadata available through clearinghouses one can find the data required for a specific application;
- *quality* of the data of concern is, so that the recipient of the data can judge the fitness of the data for his or her intended use, prior to the transfer of the data.

Modelling metadata

A model for a metadata includes the definition of mandatory or optional metadata or which metadata is mandatory under special conditions (conditional metadata). Also the minimal metadata set should be clarified.

Metadata can be organised in the computer in many ways: as an incorporated part of data within the dataset, in a separate database or as a text file. One may choose the manner of storage of metadata according to a management strategy, budget or other institutional or technical factors.

There are also many ways to transmit, communicate and present metadata. Metadata elements will be valued in different ways by different users or by one user for different tasks. The Internet and other technologies are causing rapid change in means to provide information but the need for metadata on physical means - including paper - will continue to exist

The best time to develop and collect metadata is while the data are being developed, i.e. when the information needed for metadata is known. Probably waiting until after data collection is finished result in less accurate information lacking the description of what ought to be collected in the database. Besides, searching information, when lacking metadata may become cumbersome and increase the cost. However, documenting existing data can be daunting. Details may be long forgotten and costs can be high. Though these concerns are valid in well-considered decisions should be taken allowing documenting metadata for existing databases.
Though all metadata can also be described by free text in text fields, the understanding of text strings makes automatic interpretation difficult. The more elements of metadata are encoded in pre-defined fields the easier it will be for search engines to understand the metadata by automatic means. For the evaluation of the quality of metadata sets the automatic interpretation is a key concept.

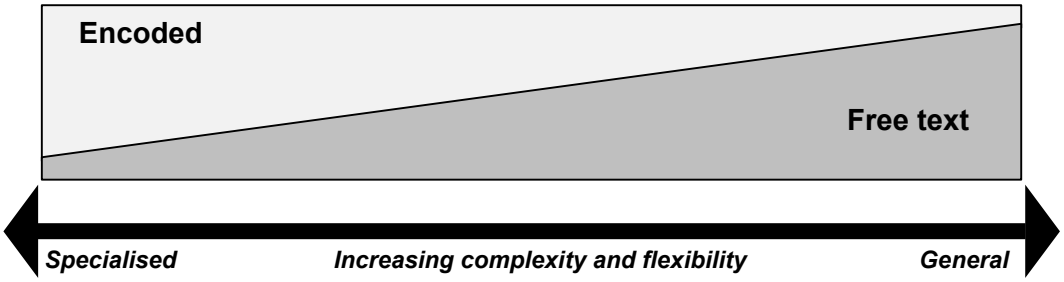

Fig. 1. The range of metadata standards.

Searching for data

In the GI context one can think of three levels of metadata. The first level deals with browsing over the web to find data resources, which could fit a given application (Discovery metadata – user's end). The second level is the metadata that is necessary to select the datasets, which will be acquired (Inventory metadata – user's end). The third level relates to the metadata that accompanies the dataset received from the data provider that enable the interface to load the geographical data into the application (Data model – provider's end).

The user's end data are understandable by human, the provider's end data are understandable by computer. Inventory metadata can also be named exploitation metadata. There are strong links between discovery, on one side and exploitation metadata on the other side, with overlaps in the concepts.

Discovery Metadata - Metadata to find data resources
The Web is an important universal information tool, embracing vast stores of information with many purposes, multiple disparate sources, and quite a few unpredictable users. There is a clear need to improve access to this mass of information and for the development of better search, retrieval, and organisational tools.
Metadata is a fundamental part of the solution to these challenges. Effective use of metadata requires three things: a set of commonly-understood terms to describe the content of information resources (semantics); a standard grammar for connecting those terms in meaningful metadata sentences; and a framework that allows us to transfer and recombine those metadata sentences across different applications and subjects. These three elements - standardised semantics, a definitive syntax, and a framework for transfer - provide architecture for resource description that can work across all subject areas on the Web.
The mission of the Dublin Core Metadata Initiative (DCMI) is to make it easier to find resources using the Internet through the following activities:

- developing metadata standards for resource search and retrieval across different subject areas;
- defining frameworks for the interoperability of metadata sets;
- facilitating the development of community- or subject-specific metadata sets that work within these frameworks.

The Dublin Core metadata element set is intended to support cross-subject search and retrieval. It can be thought of as a simplistic or pidgin metadata language that helps the user navigate through disparate subjects, languages,

and cultures. Adoption of the Dublin Core by governments, libraries, museums, archives, publishers, environmental science repositories, prints and e-print archives, to name a few, testifies to its success in this role. There are emerging applications in the commercial sector, as well, with health care organisations and financial industries using the Dublin Core as the basis for organising and exchanging information.
Part of the mission of DCMI is to provide a vendor-neutral forum for the development of additional vocabularies that are interoperable within the broader architecture of the Dublin Core and other Web metadata schemes in general. At the recent 8th Dublin Core metadata workshop in Ottawa, Canada, a special interest group formed around the exploration of metadata issues that are of particular interest to the business community. The provisional charter for this group is to provide a forum:

1. to investigate metadata schemes used in commercial business models (Business-to-Business, Business-to-Consumer);
2. to promote the use of Dublin Core in internal and cross-company business environments;
3. to identify business sectors and commercial resources (e.g. information, services, catalogues, products) that could benefit from the use of the DC standard;
4. to highlight within the DC Community the commercial ramifications of DC developments;
5. to discuss the possible expansion of Dublin Core to accommodate information vital to commercial requirements and uses.

(http://www.xml.com/pub/2000/10/25/dublincore/)

Inventory Metadata - Descriptions of datasets that could be ordered
In its simplest form a central database is maintained, with data providers updating dataset descriptions (metadata sets) as necessary. Data providers are provided with a software update tool, which includes the current version of the central database. They can update information about existing datasets or add new datasets before returning the media to the central, when the new data are then transferred to the main central database.
On the other hand a Spatial Data Infrastructure (SDI) may link the centre to all metadata providers. Locally, the providers will maintain their datasets (both the metadata and the data itself). Users may consult the metadata sets for inventory by using the appropriate web browser, protocols and standardised catalogue services to support technologically the interpretation of the (meta-) data.
Web pages are derived automatically from the database, and the information is made available free of charge to all Internet users. In a way it can be described as an Internet gateway dedicated to Geographic Information.

Model metadata - Metadata for data transfer
Most of the existing transfer standards contains metadata on top of the geographical data that are transferred and make the data being described in metadatasets understandable and interpretable for computer programmes.

Three main developments have been undertaken to standardise metadata:

1. CEN/TC 287 resulting in CEN - ENV 12657 Geographic Information - Data description - Metadata;
2. ISO/TC 211 from 1994 till now resulting in ISO DIS 19115, December 12, 2000 also used by the OpenGIS Consortium;
3. Dublin Core Metadata Initiative - DCMI (http://purl.org/dc/).

The development for standards in the field of Geographic Information in Europe took place by the CEN/TC 287 from 1991 until 1999. Its basic objective was to enable geographic information to be accessed by different users, applications and systems, and from different locations.
This requires a standard way of defining and describing this information, a standard method for structuring and encoding it, and a standard way of accessing, transferring and updating via geographic information processing and communication functions, independent of any particular computer system. Suppliers and users of data and developers of GISs and GI-applications may use the CEN family of standards to enable databases and applications that are different in structure form and content to interconnect and inter-operate.
It shows that in the CEN concept datasets exist about:

- metadata, describing the dataset of concern, including quality data and;
- geographic data, collected in datasets.

The ENV 12657 Geographic Information - Data description - Metadata defines a conceptual schema for geographic datasets and their metadata separately. Since metadata is data about datasets, it includes information about the content, representation, extent (both geometric and temporal), spatial reference system, quality and administration of the dataset. The CEN pre-standard identifies those items that are mandatory for describing geographic datasets - the minimum set of metadata. It gives examples of how the standard may be applied but does not concern itself with the construction of databases for holding metadata. It is designed primarily for use

with digital geographic datasets, but the principles can also be used to describe geographic information in other forms such as paper maps or lists.
In this way, the CEN metadata standard is used for description of metadata that results in a dataset separate from the original dataset but is necessary for reading and interpretation of the original dataset.
Both national and pan-European clearinghouse services and projects - for example the Dutch NCGI, the Croatian, Czech, Danish, Irish and Polish clearinghouses as well as the ESMI and GDDD project - use the CEN ENV 12657 on Metadata.

The CEN standard was a model for the development of the Metadata standard in the ISO/TC 211 metadata standard: ISO 211 draft standard 19115 (WI 19115 - Metadata) which is not only the metadata standard for the international community but is also adopted by the OpenGIS Consortium.

Finally, activities such as the DCMI and ETeMII - European Territorial management Information Infrastructure (http://www.ec-gis.org/etemii/) recommendations may also provide useful input, as they propose a very simple and limited set of metadata, easy to comply with, and easily open to several sectors of data, while CEN and ISO are GI focused.

Content of Metadata

Metadata includes three types of data:

- *metadata of metadata*. In order to understand the metadata of a dataset it is required to state the characteristics of the metadata such as the language and coding system of the metadata as well as the reference and coding systems that the metadata uses to define the geometrical and temporal extend;
- *directory data* describing the identification of the database, as name, origin, names and addresses of owner, distributor and manager, the spatial extend of the database in semantic, geometric and temporal sense, quality, security measures, etc.;
- *dictionary data*, including the data definition, referring to semantic, geometric and temporal definitions of the data, data organisation by conceptual schemes for homogenous datasets and quality information including the quality conceptual schema and the description of the parameters for the data as idealisation, accuracy, reliability, up-to-dateness, completeness, consistency and currency.

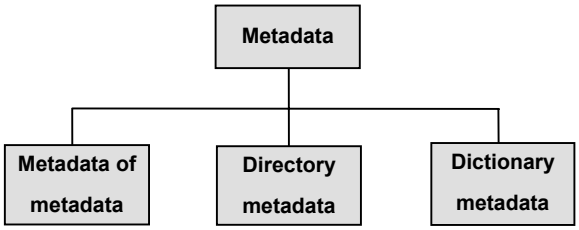

Fig. 2. Subtypes of metadata.

Metadata of metadata

In order to understand the metadata of any dataset the organisation of the metadata, the language used and the reference system for semantic, geometric and temporal extend should be given. (Metadata may refer to a homogeneous dataset series, a dataset or a set of occurrences of attributes of entities and/or relationships.)
These characteristics can be given referring to other standards or by a description:

- the *character set* used for encoding the data in the metadata set (using e.g. ISO 8859-10);
- the *language* for textual statements in the description of geographic datasets may follow a coding system as defined by ISO 639 or can be given by a text like 'Danish', 'Dutch', 'Dutch preferred', 'English (AUS)', 'English (UK)', 'English (VS)', 'French (F)', 'French (B)', 'French (CAN)', 'German (D)', 'German (Swiss)', 'Italian', 'Portuguese (P)', 'Portuguese (Braz.)', etc.;
- one need to choose a standard way of describing *dates* by a selection of EN 28601, ISO 10303, part 41 (STEP).
  The advantage of using ISO 10303 part 41 is that it also standardises the description of *persons, organisations* and *addresses;*
- for the definition of a *geometric reference* system geodetic standards could be applied. CEN ENV 12762 give directives what data should be given to define a geodetic reference systems, as well for planimetric geometry, as for the vertical component;

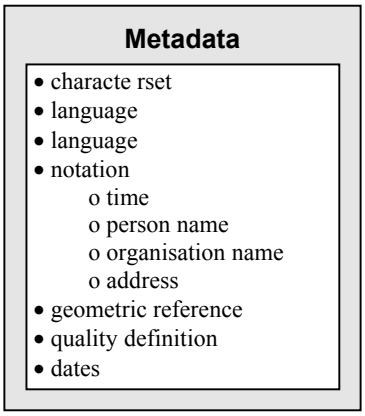

Fig. 3. Metadata elements for metadata.

- for the *quality* description one can find the method for describing in the quality schema (see chapter 5) e.g. about the original intended purpose of production by the producer, the usage prior to the present intended use, the lineage of the dataset, describing the process history, the thematic, positional and temporal accuracy, the completeness of object types and consistency between them;
- *dates* for the creation of the metadata, the last check and update and the future review(s).
- *administrative data about metadata*, being:
  - *point of contact* in the organisation that created or menages the metadata;
  - *date*s of creation, last check and update future updates of the metadata;
  - *constraints and security* measures of the metadata.

The metadata element language and quality description are mandatory elements. The positional reference system whether it be geodetic or non-geodetic (as addresses including postal codes, parcel identification, road networks, administrative subdivision of a country, etc.) are required when available in the dataset. Here, conditional metadata may also exist: a dataset with geographic information may contain semantic geographic data only where no positional data is available. Then no positional reference systems need to be given in the metadata).

Directory metadata

The elements of metadata comprise at least the dataset's

- *identification*. This can be done by a code or a name that defines the dataset uniquely and clearly amongst other datasets including the version number of the dataset when appropriate. Besides other names can be given as *alternative title*(s). This may be the dataset's name in other languages or an *abbreviation* of the dataset's name. Also the dataset *overview,* giving an overall description of the dataset including a *summary* (in text) and describing the content of the dataset can be given;
- *administrative metadata.* In order to acquire a dataset information regarding where and how the dataset is held as well as its procurement is given in the following (For descriptions of addresses, persons and organisations the definitions can be used from ISO-10303, Part 41):
  - *organisation*: name and abbreviations (if available)of the organisation , the address (postal address, visiting address, telephone number, telefax number and/or electronic-mail address, home page address on the World Wide Web;
  - *organisation role*: as responsible authority for the dataset, producer organisation;
  - *point of contact*: in many cases it is evident that personal contacts lead to the best way of transferring data from provider to user. In such cases direct contact between persons is necessary. However personnel basis is a extreme aid for mutual use of data the organisations should still appoint officials for this purpose. Here the name, function and addresses of the point of contact can be given;
  - *distribution* includes descriptions of restrictions on use, copyrights, units of distribution (e.g. per tile, per square kilometre, per administrative unit, etc.), pricing and discounts of the data (types) per unit, data media on which the dataset can be recorded and retrieved or on-line access, delivering data formats, procurement, giving instructions for ordering the data and the delivery service and services for processing the data;
- *origin* of the data:
  - *producer purpose* of the dataset, describing the original purpose by the producer for which the data in the dataset was captured. It may include the *original intended application scale*. Much of the digital geographic data is collected for direct data presentation on maps. The content definition of the dataset (selection and abstraction in the definition of the Universe of Discourse) is usually based on what should and can be represented on those maps and is very dependent on the original purpose of the dataset;
  - *capture method* and *type of semantics* describing the way of original collection and the type of data that can be found in the dataset, e.g. spacial data, aerial data landsurvey data, etc.;
  - *potential usage* give the provider's view on the potentials of the dataset, i.e. the possibilities of the data for different applications. Also the *usage* can be described to give the future users an idea for which applications the dataset was already used. Important in these cases is also to give an impression of

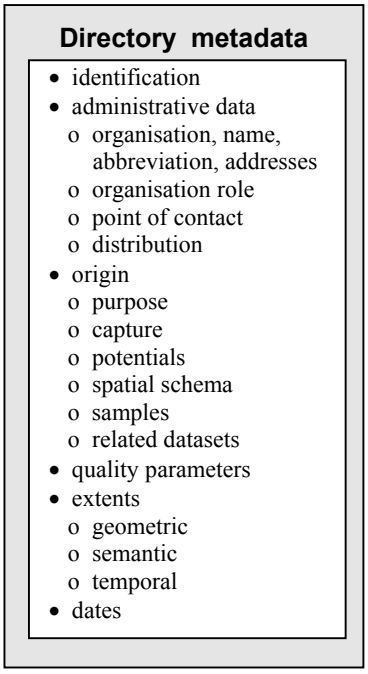

Fig. 4. Elements for directory metadata.

the successes and failures that are experienced in the specific applications of the dataset. *References* to other relevant published documents or public available additional documentation may be provided;
  - type of *spatial schema* used in the dataset. This may be a standard or a user defined spatial schema, describing the main characteristics and components of the schema. Also the definition of the *spatial reference system* should be given in the case of use of a spatial schema. The metadata may also describe geographical data that does not contain any spatial reference. In this case no spatial schema has to be defined in the metadata
  - one or more examples taken from the dataset and being representative for the whole dataset can be provided as *samples*, e.g. as a browse graphics either in raster or in vector format;
  - title and/or code and provider of *related datasets* of possible interest to a potential users can be given for further information about the possible uses of the dataset.
- dataset *quality elements*. The dataset quality describes the difference between the dataset and the user's and producer's Universe of Discourse. By describing the quality of the dataset the user can determine whether the data has enough potentials for the intended application prior to the transfer of the data. The dataset quality elements follow the designed quality conceptual model and comprises:
  - *source* describing the producers'-, owners'-, managers'-and providers' as well as the point of contact within the providers' organisation;
  - *homogeneity,* giving a description of how well the data in the dataset follow the overall uniformity of the data;
  - *usage* gives an overview of the applications for which the information in the dataset has been used previously and how well the data fitted in these applications. Also the *potential use*, is an important aspect in these which gives an indication of the possibilities of the data seen from the providers' perspective;
  - *lineage* giving a description of the origin of the information contents in the dataset and everything that has happened to the data since, until the moment of transfer;
  - the quality parameters as *positional accuracy, thematic accuracy, temporal accuracy, logical consistency, completeness*;
  - *up-to- dateness* dating the information in the dataset as to how well the data is kept up to date;
- *extents*. Data about geographic objects are positioned in space and time. In order to judge for the user whether a dataset is suitable for the intended application the planar, vertical and temporal extent of the data in the dataset should be available.
  This can be done by one or more 'bounding range(s)' of the dataset giving the maximum and minimum co-ordinates appearing in the dataset or by one or more 'area boundary(ies)' delimiting the area(s) covered by the dataset. Also the currency of the extent, indicating the status, completeness and validity should be given. For the temporal extent an indication of the range (from start to expiring date) should be given (this may still be 'on-going', i.e. having no expiring date yet), but also a descriptive text may indicate such as 'mediaeval period', '20th century', 'annually update', 'continuously update', 'last update' etc.

Dictionary metadata

The dictionary data describe the semantics of the data in the dataset as well as the conceptual schema that has been used for data modelling. This allows search engines to not only access the data in the dataset but also interpret the semantics for determining the requested information and gives the user access both to the data and its semantics.

Dictionary metadata consists of the following elements:

- *spatial reference* systems. The spatial reference system may be either a direct, geodetic reference system based on co-ordinates, or a non-geodetic reference system, e.g. based on addresses. A dataset can contain objects referenced by a multiple of geodetic and/or non-geodetic spatial reference systems.
  Elements for the definition of the non- geodetic reference systems are: type (e.g. the system in which the references are given as country, county, municipality, etc.), title and owner and version date of the non-geodetic reference system.
  For geodetic reference systems the name of the reference system, the co-ordinate representation and its units (for both planimetry and height), the name of the datum, ellipsoid, map projection, etc.;

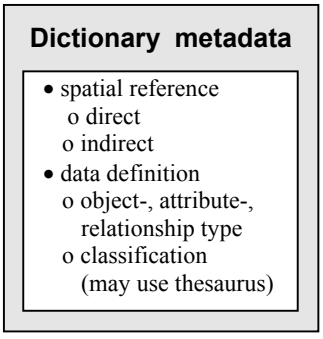

Fig. 5. Elements for dictionary metadata.

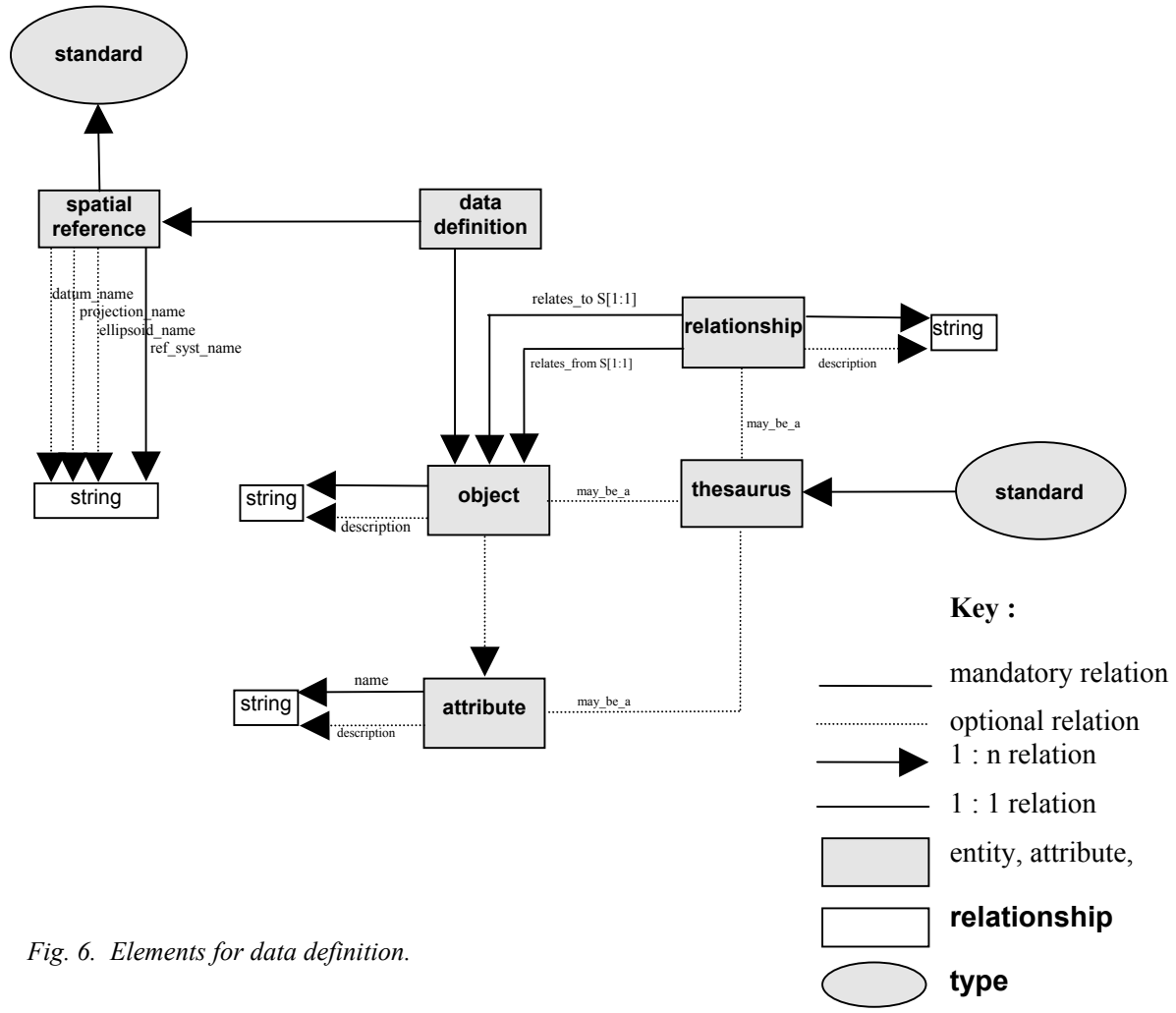

Fig. 6. Elements for data definition.

- *data definition.* Objects may be defined in different ways in different datasets. To enable comparison between them, data describing the characteristics of objects are required. Data definition and classification, describing the differences between classes in order to distinguish between objects in different classes and to define the relationships between the classes do this.
  For this purpose, data is provided giving the definition of object types, attribute types and relationship types, wherever they exist in the dataset. Together with the application schema all objects that are represented by data in the dataset should be defined and become understandable for the receiver of the data. Sufficient description of the data definition is given by:
  - object type, attribute type and relationship type *name and definition* should be given completely. If object types, attribute types or relationship types are indicated in the dataset by codes also the coding should be part of the definition. Relationship types should also indicate the object types they relate (a 'from' and 'to' object type);
  - *classification*. The object, attribute and relationship types of a dataset can be described in a classification system, in which the object, attributes and relationship types may take part in an given hierarchical organisation. These hierarchies should be given in the form of a thesaurus according to the standard ISO 2788: 1988.
    In the case of using a thesaurus the name, version number or version date, the thesaurus administrator's point of contact and references to other publications about (use of) the thesaurus should be available to the user. If not all the thesaurus elements are used in the classification system then those thesaurus elements that are used shall be listed in the classification part of the directory data within the metadata set.

When a thesaurus is provided in the metadata set, besides the definition (if it does not belong to a 'de jure' standard) of each thesaurus element also the code, and synonym should be given; this can be accompanied by related, narrower and broader terms as well as pictures.

Minimal metadata set
Minimally a dataset should have some metadata to stay understandable for any enquiry by users. It is noted that elements have descriptive text to convey a common semantic understanding of the elements. However, controlled vocabulary may be required to promote global understanding of the element values. Comparison to bibliography (that has much experience with metadata to structure the vast amount of publications for consulting) learn the following minimal metadata set:

1) *identification* of the dataset by a *name* given to the dataset by the original producer or publisher, being unique to distinguish the dataset from others. This might also be done by an unambiguous *code* for the dataset;
2) *publisher*, providing the organisation responsible for making the dataset available in its present form, such as a publishing house, topographic service, private company, municipality, etc.;
3) *author* or *original producer* which is the organisation responsible for the original capture of the data in the dataset;
4) *other contributors*. A dataset may consist of several subsets for which the data is gathered by different organisations; their source should also be mentioned;
5) *reference systems* for spatial reference, semantic definition by thesaurus and time definition;
6) *extent*, describing the geographical, semantic and temporal coverage of the valid data in the dataset and whether or not spatial attributes are carried in the dataset. The geographical coverage may be a description by a set of quadrangles or geographically defined circumferencing boundaries. The semantic coverage may be a list of the type of objects by semantic attributes and the temporal coverage indicates the time period of the valid data in the dataset;
7) textual *description* of the content of the dataset; the description may include samples in pictures, video or sound;
8) *date*, indicating the date of validity of the data in the dataset. Many ways of writing dates are possible but if used they should be written in a clear and unambiguous manner, e.g. as EN 28601 or ISO 10303 - 41.
9) *language*, of the metadata
10) *format* of the dataset transfer, used to identify the software that will be required for reading the data. Standard formats should be used as indicated in the tables 3.2 and 3.3 or others as CEN ENV 12658, ISO 10303 - 21;
11) *quality*, describing the spatial, semantic and temporal quality parameters for the dataset. In the quality definition of larger datasets information should be available of the meta-quality indicating the quality indicators e.g. for spatial quality a relative standard error should known as 'average', 'maximum', 'minimum', 'expected', 'required', etc.;
12) *relation* to other datasets that in itself may be datasets too;
13) *rights* and *management* indicating the copyrights, constraints on use, way of distribution (e.g. by tiles, by square kilometres, etc.).

Apart from the minimal metadata set the format for the dataset should be declared. Since the 13 core elements are a subset of the total metadata set, they follow the description of the full metadata set.

European projects funded by the EU

The European geographical information market is merely nationally focused, and this means that users requiring pan-European datasets face a number of obstacles in acquiring the information they require, including:

- lack of awareness of the datasets which are available, their quality, and means to obtain them;
- datasets from different organisations are often available with different licence terms (which may even be contradictory);
- data itself is unlikely to be consistent across borders.

The European Union stimulates the use of digital information by the INFO 2000 (http://europa.eu.int) program supported by EUROGI.
INFO2000 aims at stimulating the emerging European multimedia content industry, encouraging the use of multimedia content and exploiting new business opportunities.
The central theme of INFO2000 is the development of a European information content industry capable of competing on a global scale and able to satisfy the needs of Europe's enterprises and citizens for information content leading to economic growth, competitiveness and employment and to individual professional, social and cultural development. The programme aims to achieve this through four main Action Lines:

- stimulating demand and raising awareness;
- exploiting Europe's public sector information;
- triggering European multimedia potential;
- Support Actions.

The INFO2000 programme has a four-year work programme from 1996 until 1999.
One of the EC funded projects from the fourth and the fifth framework programmes (FP-4 and FP-5) is ETeMII (European Territorial Management Information Infrastructure) aiming at organising a network of excellence, bringing together most of the stakeholders of the Territorial Management Information market, coming from research, industry and public sector. Territorial management means any management activity, related to the territory; it covers a wide scope of activities, including agriculture, transport, utility management, land planning, environment, fisheries, geo-marketing, etc. Particular attention will be given to user participation within this project, so that all tasks are based on users' needs. Awareness activities are an important component of the project.
Such a network will build a consensus on the most important technical issues that are the foundation of ETeMII:

- reference data, data access policy;
- interoperability, standards implementation, including metadata;
- research and development challenges;
- integration of space tools: positioning systems (GLONAS, GPS), Earth Observation and telecommunications;
- active participation into global initiatives: ISO, GSDI, OGC, etc.

To move a step toward the creation of a European Information Infrastructure three themes of ETeMII (http://www.ec-gis.org/etemii/) are promoted: to be able to advertise best practice and promote contribution to and use of GI infrastructure issues.

EUROGI (http://eurogi.org), the European Umbrella Organisation for Geographic Information, was set up in November 1993, as a result of a study commissioned by DG Information Society (DG XIII) of the European Commission to develop a unified European approach to the use of geographic technologies.
Realising present-day economic development, modern countries demand accurate and detailed geographic information to maximise the value of promising new geographic technologies such as global positioning systems (GPS) and geographic information systems (GIS). To realise all the potential benefits of geographic technologies at the European level, EUROGI stimulates the harmonisation of required geographical data by European co-operation, complementing the same efforts at the national level, trying to improve:

- transfer and integration of geographic information;
- sharing of experience gained by the numerous national and international geographic technologies research initiatives,

to contribute to the reduction of the costs of geographic information and geographic technologies and to their more widespread use.

The mission of EUROGI is to maximise the effective use of geographic information for the benefit of the citizen, good governance and commerce in Europe and to represent the views of the geographic information community. EUROGI achieves this by promoting, stimulating, encouraging and supporting the development and use of geographic information and technology.
EUROGI tries to achieve this mission by the following objectives:

- to raise awareness of the value of GI and its associated technologies;
- to encourage the greater use of geographic information in Europe;
- to work towards the development of strong national GI associations in all European countries;
- to facilitate the development of a European Spatial Data Infrastructure;
- to represent European interests in the Global Spatial Data infrastructure.

The following projects are stimulated by EUROGI and funded by the EU and the national public and private sector:

1. **LA CLEF**

MEGRIN (Multi-purpose European Ground Related Information Network) represents, and is funded by, a group of 19 European National Mapping Agencies (NMAs). It aims to bring a European perspective into NMAs national activities. MEGRIN aims to meet the increasing demand for pan-European data by improving international users access to national datasets.
It does this by providing information about the digital data available now in 23 countries, and is creating harmonised pan-European datasets. Although MEGRIN comprises only 19 full members, its everyday partners are all European NMAs, the 30 plus CERCO members (Comité Européen des Responsables de la Cartographie Officielle).
Recognising the growing pan-European market demand, MEGRIN has been established to focus on two areas of activity: providing metadata (through GDDD) and creating new harmonised digital datasets.

Geographical Data Description Directory (GDDD)
MEGRIN's GDDD metadata service provides information about 250 digital datasets available from the NMAs of 23 countries of Europe. The GDDD was also the first pilot implementation of such a scale of the pre European metadata standard CEN ENV 12657 of the CEN/TC287.
Its current structure has existed since its introduction in November 1994. In 1996 it became more accessible as the widespread use of World Wide Web (WWW) browser became common place.
Metadata information in the GDDD falls into the following categories:

- *Overview*: Short abstract, including contact address (organisation, web-site address and person of contact);
- *Commercial information*: Contains some commercial details of coverage, copyright, format, price and other conditions related to the use of the datasets;
- *Technical information*: describes the technical specifications of data sources, features & content, updates, data accuracy and other data quality parameters;
- *Descriptions of the provider*, held in organisation details.

Users require data which is easy to access, and would ideally like to use a 'one stop shop' to view, purchase, and be supplied with a wide range of geographical data, including topographic data, remotely sensed imagery, geological and demographic data. So, LaClef must be a four-sided solution covering:

- semantic issues,
- distributed architecture,
- a wide range of services related to the metadata offered on LaClef,
- e-commerce facilities.

It is likely that this situation will be created incrementally developing such metadata service by:

1. distributed metadata systems linking initially various (existing) national databases of metadata;
2. access to databases to enable on-line data sales.

LaClef will use the XML standard to enable easy metadata transfer from the data provider to the central database, or alternatively offer a linked service to another meta database. The data producer will be able to extract XML-formatted metadata from his own metadata in his own local metadata base. The XML-formatted metadata will be sent to the LaClef metadata service where they will be imported with an import tool. From LaClef, they are available by dynamic HTML-pages to the users.

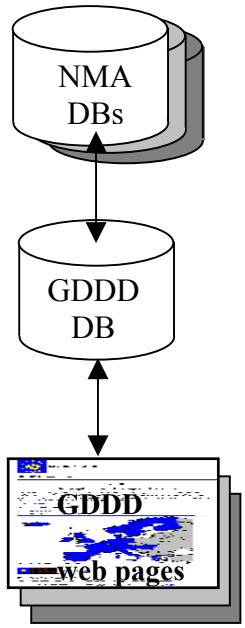

Fig. 7. As example: Central database at MEGRIN, which is updated with data collected from NMAs and from which static web pages are derived.

2. **GEIXS**
This Geo-Scientific Electronic Information Exchange System will help in all European languages to find out, whom to approach for information on minerals, oil- and gas, groundwater, geology, natural disasters and geotechnics everywhere in Europe (http://www.geixs.brgm.fr/).
GEIXS is the European Geological Data Catalogue. It gives a dataset description through:

- the geographic coverage of the data;
- key words from lexicons;
- free text.

Its aim is to provide a single point access to geological metadata because geology crosses borders. It also aims at reducing language difficulties. It is useful for land use planning, minerals industries, civil engineering, waste disposal, energy sectors, water industries, environmental sectors, insurance and banking, pollution control, government infra-structural strategy, coastal flooding, global change, health studies (e.g. radon), schools, colleges and universities.

3. **AVID**
AVID (Added Value Information Dissemination from Hydrographic Data Sets) will develop a prototype on-line service to provide access to this information to both general and specialised users. Data on bathymetry (depth measurement), coastal topography, sedimentology, waves, currents, tides, landmarks, buoys and beacons, lights and sea-limits will be made available.
AVID aims to demonstrate the effectiveness of a European information service based on hydrographic data. A key objective is to add value for potential users by providing the integration of different sources of hydrographic data.

4. **CLEAR**
When infrastructure developments or business projects cross-national boundaries there can be particular problems in assembling all the geographic (and environmental protection) information needed. It is in the interests of both providers and potential users of this data that it is easily accessible, up-to-date and comprehensive.
This CLEAR (spatial data CLEARing house) project focuses on the Saar-Lor-Lux region, which includes the German Länder of Rheinland-Pfalz and Saarland, the Grand Duchy of Luxembourg, the French region of Lorraine and the Belgian Province Luxembourg, an area with some 8 million inhabitants. The project team will develop a system giving information about available geographical data (a metadata system) including ownership, price and technical quality held in the participating countries. It will also provide a means to access the data itself.
The key objective of CLEAR is to develop a central, bi-lingual French/German information system for geographical data in the region (a metadata service) as well as a functional delivery system to provide direct access to information resources held by the public sector.

Project name	Data type	Metadata Standard	Conceptual language	Legal aspects	Pricing	Organisational aspects
MEGRIN/GDDD	topographic	ENV 12657	XML	national	national	NMA's
GEIXS	geologic			national		NGA's
AVID	hydrographic	Web-based		yes		NHA/regional
Clear	property	Web-based			Web	Bi-lingual
ESMI	geographical	ENV 12657	XML	national	national	4 partners
Geoserve	geo-brokering	ENV 12657			private	
GESIDI	geo-data/trade	ISO 19115	XML/EDI	national		9 partners, multi-lingual
Madame	Geo-statistical cadastral	Web-based	Text/ HTML			4 country organisation;

Table 1. Comparison of aspects for different European projects using metadata.

5. **ESMI**
The use of geographic information in all parts of European society is growing. Private business, government, research and educational institutions as well as individuals are increasingly using geographic information as a key component in their activities. Increased international co-operation and competition means that there is a need to determine what geographic information is available in other countries or organisations and how to obtain it. The proposed European Spatial Metadata Infrastructure (ESMI) is an initiative set up by several European public and private organisations (i.e. CNIG in Portugal, Geodan in the Netherlands, LISITT (University of Valencia) in Spain, MEGRIN from France, Ordnance Survey of the United Kingdom) to establish a framework for the distribution of geographic information by creating a universal metadata service.

There are already a number of existing metadata services in Europe and elsewhere. However, these systems are specific, nationally oriented and rely on central servers which may be separate from the actual geographic data. They do not communicate, so there is some duplication which gives rise to inconsistency. ESMI has the objective to link these existing and future metadata systems.

6. **GEOSERVE**
Europe has a wealth of high quality digital geographic information. The variety of data formats, reference systems, projections and quality standards reflect the wealth of Europe's history; but it is an obstacle to cross-border transfer and integrated use of geo-data. The access mechanisms are poorly developed. The GeoServe (Geo-Data Access Services) project develops a brokering system that allows the user to identify the geo-data from a data catalogue of many providers in Europe and to order it in formats required by the users' applications. Additionally geographic services can be ordered.
The system is based on distributed meta databases that implement the CEN TC287 standard. Communication is done on Internet and Intranets. For casual users simple Web clients and information kiosks are provided. Interfacing a variety of geo-viewers and GISs supports Professional users.

7. **GISEDI**
The GISEDI Europe project (Electronic Trade for Geographic Information) aims to develop a European commercial and technical network infrastructure which will both facilitate and accelerate the transfer and trading of geographic information at local-, regional, national and international level. This requires the development of web based GI query, view and retrieval functionality, integrated with secure transactional procedures (EDI and e-commerce solutions), which will be tried in four European countries. The principal output of the project will be the development of a Book of Specifications, which will provide the technical basis for development of GISEDI systems in other European and world-wide locations.
GISEDI is a project involving nine organisations from seven countries in Europe (i.e. European Umbrella for Geographical Information EUROGI in France, Cara Broadbent & Jegher from France, Walter Research Centre from the United Kingdom, Instituto Engenharia de Sistemas e Computadores in Portugal, Ususimaa Regional Council in Finland, EDI Hellas SA in Greece, Indra SSI SA in Spain and URBA 2000 from France).

8. **MADAME**
The MADAME project (Methods for Access to Data and Metadata in Europe) will identify solutions and best practices for making public sector data available across Europe. Focusing on geo-statistical and cadastral (1) information in particular, it will evaluate current services providing access to data and metadata (2) from the perspective of current and potential users. The overall objective is to move from services that are producer-oriented to user-oriented services.
A key strength of this project is that it will examine data provision at (and across) three levels: European, national and local. This approach will be critical to developing transferable guidelines of best business practice. It will inform public sector agencies in their approach to making their datasets more accessible to others.

Conclusion
In conclusion, there are many issues to be dealt with when developing metadata services. Generally spoken they are of four levels:
1. data model standardisation;
2. semantic standardisation;
3. languages;
4. search queries.

Data model standardisation
It is likely that implementing databases of metadata according to standards will not only provide more reliable information to users, but should also help the data providers by:

- reducing the duplication of effort which currently exists in different databases;
- increasing access to their data descriptions, so increasing their possible sales (distributed systems need to agree on a common standard transfer format for the on-line transfer of information);
- ensuring that suppliers are able to store the same information (whether they choose to do so is another matter), possibly as a common core set of metadata.

Unfortunately, there are many metadata standards to choose from! These include the European standard ENV 12657 from CEN/TC287. Internationally, ISO's TC211 is developing the metadata standard ISO DIS 19115 which will ultimately replace any European standard.

Semantic standardisation

The use of standard data structures will not necessarily ensure that the metadata is of consistent quality, completeness and accuracy. Indeed, the quality of the metadata may be more important than the data structure, since this is what users will see and use. Without reliable information, users are unlikely to use the service.
All data providers and system users must have the same understanding of terms. Without this, the search results, which are presented to users will be meaningless and/or incorrect. This means that all data providers must either attach standard keywords to their descriptions, or a thesaurus mechanism needs to be created to provide a common view of individual implementations. The task of harmonisation is likely to be complex, as there is considerable scope for differences to occur (for instance, between disciplines, within the same discipline in different countries, and even between different organisations in the same country).
The need for semantic harmonisation is widespread: for example, there is not yet a standard for location references, with which geo-coded data and geographic information can be tied together. Where standards do exist, they are not always implemented!
Europe's experience indicates that a reliable metadata service depends on the quality of the data descriptions. This is particularly significant in a multi-national environment. The initial data collection exercise raised problems because different organisations had used different keywords to describe datasets based on the same data model. This disagreement in choice of keywords was significant for complex themes, e.g. 'topography' and 'land cover', while straightforward keywords, e.g. 'road network' and 'railway network' were generally interpreted consistently. This clearly illustrated the importance of having the same understanding of the terminology.

Languages

To propose a real solution for semantics consistency, the services aiming at covering all Europe, cannot be satisfied with only one language. There are even more critical reasons to point out the inadequacy of a uniquely English language service:

- it may be expected that a fair number of technical experts are sufficiently fluent in English, but that will not necessarily be the case for all potential users of GI. As it is the wish of both the European Commission and European organisations to make GI, and particularly public sector in GI, more available to a larger range of users, it is desirable for the service to develop a friendly interface available in a number of national languages;
- for the best semantic comparability between countries using different languages, it is essential that multilingual thesauri and keywords are developed and tested. Above being tools for ensuring cross-border (and cross-language) metadata consistency, it will allow the easy creation of multilingual interfaces. The automatic on-line extraction and translation of nationally based metadata are investigated.

Search queries

Ideally, the user who is looking for data would like to locate the data he is interested in, by pointing at the area covered by the project he is working on. Solving the main semantic and language issues should allow this possibility.
A text based query such as "FIND objects = 'roads' IN area = 'Benelux'" would be close to requests expressed in natural language. There are a number of ways to develop such a relatively user friendly search query. Knowledge about regions or geographical locations can be stored by the way of geographical keywords. Different relatively common solutions for look-up tables have to be considered, when on the pan-European level, each having its strengths and weaknesses according to specific approaches:

- geographical co-ordinates, based on a unique geodetic system, such as EUREF;
- administrative units;
- postal codes;
- addresses.

Literature

1. WWW
   Several web addresses as indicated in the article itself.

2. FGDC
   Metadata standards development. Content Standard for Digital Geospatial Metadata, Version 1.0
   FGDC Secretariat: Email F.G.D.C.