STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper no. 14
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Tallinn, Estonia, 25-28 September 2001)

Topic (ii): New technological solutions, including those based on online data access

A NEW ARCHITECTURE FOR THE GISCO REFERENCE DATABASE

Submitted by Eurostat[1]

Contributed paper

Abstract

In 1998 Eurostat and the Joint Research Centre in Ispra launched a project to modernise the infrastructure of the GISCO reference database. The future GISCO database should follow an object-oriented approach including the storage of the data in a relational database. This approach would make it possible to combine spatial and attribute data in one single database. Additionally, lifetime cycles for geometric objects will be introduced. Analyses are at the moment carried out to investigate the steps to be taken before a migration can be realised.

1. INTRODUCTION

1. The use of geographical information systems at the Commission began with the CORINE Land Cover project, launched in 1990. An internal working party was then asked to assess the Commission's requirements for geographical information and geographical information systems (GIS). The final report by the working party highlighted the usefulness of GIS as tool for integrating data from different sources (socio-economic data, environmental data, transport data etc.). This led Eurostat in 1992 to set up the permanent GISCO (Geographical Information System for the Commission) project. The Nomenclature of Territorial Units for Statistics (NUTS) forms the main link between statistical data and geographical information. Monitoring this classification over time, together with the corresponding geographical contours, is the core of the system set up by Eurostat.

2. The objective of this paper is to describe the future architecture of the GISCO reference database and the migration of the GISCO spatial data to a new database environment. The architecture of the GISCO reference database is currently built around Arc/Info. The structure of the software was developed in the late 1980s and has been continuously improved but without touching the core of the system. Fundamental technical developments in the field of geographical information systems have taken place in recent years. These developments prompted GISCO to reconsider the structure of its database. The idea behind the future architecture is to move from a proprietary environment to an open database system and thus combining non-spatial tabular data with spatial information. This will be a way to promote the usage of spatial data in Eurostat and the European Commission.

3. In 1998, Eurostat and the Ispra Joint Research Centre launched a project to modernise the database infrastructure (hardware and software).

---

[1] Prepared by Albrecht Wirthmann and Anette Björnsson

4. Major aims of the project are:

- to combine spatial and attribute data in one single database
- to include the management of spatial data in the database management
- to introduce an object oriented approach for modelling spatial data which includes the establishment of dependencies between spatial objects
- to open the usage of geographic information systems to less skilled users (no longer an exclusive expert domain)
- to make use of more common user interfaces (e.g. web browser)
- to introduce lifetime cycles for geometric objects by assigning start and end dates (allowing thus the production of maps for a certain point in time by selecting valid spatial objects in the database).

5. Migrating to a new environment gives the opportunity to assess the quality of the existing data and to improve it by applying relations and constraints inherent in the management of a database. Together with the spatial data, the metadata should also be migrated to a database. In this way, the three components: tabular data, spatial data and metadata can be stored in one system allowing interactivity between them.

II. THE CURRENT GISCO REFERENCE DATABASE

II.1 Contents and content related aspects

6. The GISCO reference database is trying to cover the common interest of the European Commission services in spatial data. It has been developed since 1992 and currently comprises 12 themes subdivided into 40 layers. The themes are:

- Administrative data,
- Community Support data,
- Infrastructure,
- Hydrography,
- Altimetry,
- Environment,
- Industry,
- Land and Nature Resources,
- Support and
- World data.

7. The geographic extent of the spatial data is at first priority the EU15 and then the accession countries. A third priority is given to Pan-Europe as defined geographically. Some data cover the entire world[2].

8. The main scales[3] of the data are between 1:1 Mio. and 1:3 Mio. The overall scales of the reference database vary between 1:100 000 and 1:20 Mio. The scale implies a certain value of spatial

[2] The GISCO database manual gives a detailed description of the contents.

[3] There are two different kinds of scales inherent in the spatial data. The **source scale** describes the scale of the source map that was used to digitize the spatial data. The **application scale** is the scale that corresponds to the accuracy of the spatial data. Usually the application scale is smaller than the source scale because the

accuracy (resolution) of the data. The variety of scales in the reference database influences the types of operations that can be applied to the data, e.g. when overlaying different spatial datasets the resulting accuracy corresponds to the lowest scale data.

9. The spatial data comes from different independent sources. As a consequence, spatial features coinciding with each other may not share the same co-ordinates. The lack of logical spatial consistency restricts the types of applications.

10. As the spatial data usually originate from national sources, it must be transformed from national co-ordinate and projection systems to a European one. As the exact transformation parameters are not always known, this might introduce additional spatial errors. The current projection system used for GISCO (Lambert-Azimuthal) represents an application point of view. Some drawbacks of this system are that it is not suitable for universal storage of the spatial data and it introduces errors when reprojecting between different co-ordinate systems.

II.2 Current Database Structure

11. The structure of the database is closely related to the GIS software used for its implementation. The spatial datasets are stored as Arc/Info coverages for vector data and Arc/Info grids for raster data. The attribute data are kept in INFO tables. The format of the data is proprietary, the data can be interchanged via export formats.

12. Storing spatial data as coverages means applying the concept of a defined topology. Arc/Info distinguishes between point, line and polygon features, each of which has a predefined type of attribute table. The topology concept reduces redundancy when storing polygons. Polygon borders are only stored once, line and polygon geometries share the same storage. The concept of stored topology facilitates spatial analysis, e.g. when querying for neighbouring features or establishing routing systems, but there are also important drawbacks. Point and polygon features cannot be stored in the same coverages. Complex polygons that consist of several non-contiguous polygons are represented with multiple datasets. Overlays of different non-harmonised datasets result in a large number of sliver polygons that complicate the management and the analysis of the newly created dataset.

II.3 Applications and usage of the GISCO reference database

13. The database contents reflect the common needs of the Commission concerning spatial data. Therefore the data are mainly used as reference for additional thematic data held by the different Directorates General. The Nomenclature of Territorial Units for Statistics (NUTS) forms the main link between statistical data and geographic information. Monitoring this classification over time, together with the corresponding geographic contours, is the core of the system set up by Eurostat.

14. The most important objective for GISCO is to disseminate the reference database to the users in the European Commission services. Currently this is done by copying the files from a commonly accessible server in the computing centre and via CD-ROM.

15. The second important activity is the production of thematic maps for the various publications of Eurostat. The demands on spatial accuracy for producing thematic maps on a European extent are low because the scale of most of the maps is rather small.

---

digitization process introduces additional distortions. Unless mentioned explicitly the term scale refers to application scale in this document.

16. A GIS system provides the tools for doing spatial analysis, i.e. linking different thematic spatial data by their location. Spatial analysis requires higher resolution data and a more homogenous scale than mapping. Currently there are only two layers, the commune boundaries and the land cover map that are in a scale range of 1:100 000 - 200 000 that allow performing simple spatial analysis.

II.4 GISCO database meta information

17. The existing meta information for the GISCO database is divided into the database manual and the data dictionary. The data dictionary contains the attribute definitions of the different layers and information on the lineage of the datasets. The metadata is stored in a database, a visual basic application is used for data entry and retrieval.

18. The database manual contains additional information on the source, the type, the contents, the spatial reference system, and the distribution, etc. of the spatial data. The information is stored as HTML pages. Thus the information cannot be queried directly.

III. THE NEW GISCO REFERENCE DATABASE

III.1 Principles

19. The future GISCO database should follow an object-oriented approach. This approach makes it possible to define real world entities with their attributes, behaviour and relationships, mapping them to simple features according to the OpenGIS definitions[4] and finally to tables, relationships and procedures in a relational database. The Arc/Info Geodatabase concept represents this approach. The modelling of the database is done with the Unified Modelling Language (UML).

20. The current GISCO database needs to be redesigned by means of UML for introducing spatial objects and features. The redesign of the model should consider certain aspects, such as the spatial accuracy of the data, possible relationships between different feature classes, the co-ordinate system for storing and representing the spatial data, a new concept of topology, the problem of generalised feature classes, data dissemination, the time aspect, the issue of updates and metadata.

III.2 New Database Structure

21. The GISCO data will be stored in a relational database. Newly developed extensions of existing software allow the storage of spatial data in relational databases following the "Simple features specification". The Arc/Info Geodatabase model complies with these specifications[5].

22. The use of a relational database for storing and retrieving spatial data also means abandoning the concept of stored topology. In a database environment each feature is stored as a separate object in the database. Topology is established on the fly by spatial queries determining the spatial relationship between two objects. This means that spatial constraints can be checked and verified before inserting objects into the database. This also implies that objects sharing the same borderline have to be stored twice or even more times as neighbouring borderlines of polygons are stored as part of each polygon. If the border also represents a line feature, e.g. a river or a road, it has to be stored as a line feature separately. Thus co-ordinates have to be stored redundantly. As a consequence the relationships between different feature classes have to be clearly defined and checked when inserting, changing or deleting features in the database. The advantage of this concept simplifies overlay procedures

[4] See OpenGIS Abstract Specification, Topic 1 – Feature Geometry.
[5] see http://www.opengis.org/info/newsletter/2001/20010529ogc_news.htm

because spatial accuracy can be considered in spatial analysis without needing to create new feature classes and to remove sliver polygons.

23. By applying the mentioned database technology it will be possible to introduce lifetime cycles for geometric objects by assigning start and end dates. A map for a certain point in time can be created by selecting spatial objects valid at that time.

24. The GISCO database must be redesigned starting from the current structure that represents the user's point of view, deducting a logical model with defined objects and relationships, implementing a physical model that matches the objects to physical tables and relationships.

III.3 Content related aspects

25. The introduction of the new database architecture has consequences for the data. In principle, it requires more harmonised data aiming at a certain application scale. This is a prerequisite for introducing spatial constraints and relationships. The aim is to build a geographic information system that allows users to do spatial analysis by overlaying different layers of spatial information. A scale of 1:100 000 - 1:250 000 is proposed for this purpose. The spatial objects represented on this scale have a resolution of approximately 30-m. The suggested medium scale is a trade off between amount of data and acceptable generalisation applied on this level. All data used in a spatial analysis project should be represented in this common range of scale.

26. A second demand when applying spatial analysis is the topological correctness of one layer compared with another, e.g. a railway that runs in reality on the left side of a river should also be located on the left side in the database. To define and verify these spatial relations it is necessary to introduce a reference layer. As it serves as a reference for a large variety of themes the contents of that layer should be very generic. The most generic reference layer is a remotely sensed image of the earth. It contains all information needed to assess topological relations between different layers. On the other hand it is too generic having almost no information that can be used directly by the computer, e.g. the computer cannot determine if a railway line represented as a vector in a digital map is located on the correct side of a river, as it is appearing as pixels in the image.

27. An aspect of an image is that it is not generalised, so it can be used for quality assessment based on random sampling and for digitising new information layers (e.g. validate thematic information and geometrically correction of the location of spatial objects).

28. A vector database as reference layer can be utilised for determining spatial relationships, e.g. between a river network and the railway network. The disadvantage of this approach is that each dataset might already be generalised to a certain scale and that it cannot be used in a generic way. The themes that constitute the basic dataset have to be defined to limit possible spatial interactions between the different layers.

29. To conclude, the optimal solution would be to combine a basic thematic vector-dataset with a remotely sensed image that can be used for quality assessment and as final decisive element.

30. Figure *1* shows the hierarchical structure and the relation between basic and derived layers. The satellite image constitutes the basic reference layer. The vectors of the subsequent basic layers fit into the basic image layer. The basic vector layers may share common lines or points. The vectors of the derived layers are extracted from the basic vector layers or cannot be perceived in the basic image layer.

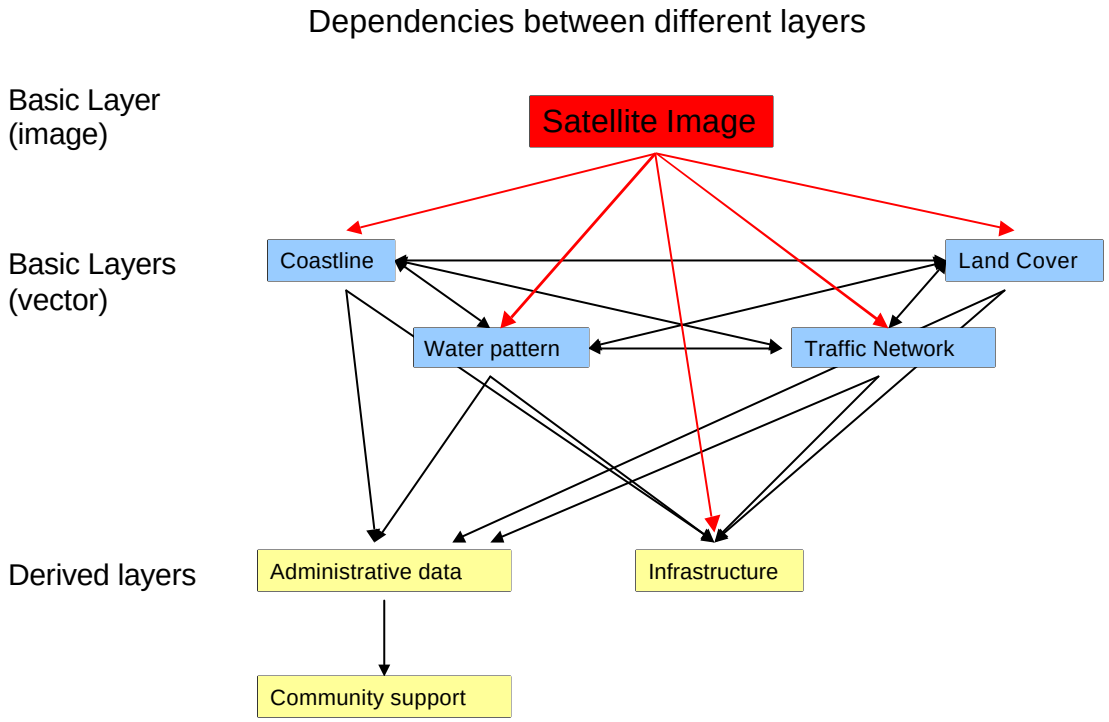

Figure 1 : ***Interactions between reference layers and derived layers***

31. Besides the quality control mentioned above it is necessary to ensure that updates will be available on object basis and that data exchange between GISCO and the data supplier is possible. If GISCO corrects geometry of data it should be possible to integrate the corrections in the original dataset in order to receive correct updates.

32. The new object oriented approach allows the integration of behaviour into the database. This includes the definition of drawing symbols or the selection of generalised features depending on the mapping scale.

33. Depending on the output scale this involves the production of generalised maps for cartographic purposes. Generalised maps have to be generated from the base-scale features. In order to keep the topological consistency between different layers the geometrical relations between different layers have to be taken into account during the generalisation process.

34. It is foreseen to produce generalised features for mapping scales ranging from 1:500 000 to 1:20 Mio. The concept of generalised maps could be applied when viewing the spatial data. At European level it is sufficient to display the small-scale data with limited geometrical and attribute information. When zooming in, the software can automatically switch between different levels of generalisation and display features with more geometric and attribute information.

35. Another point to consider is the definition of data dissemination. Together with the new structure, the possibility for online dissemination of the database could be introduced. Depending on the user, in or outside the Commission, data could be directly accessed using either a direct tcp/ip connection or via the ARC Internet Map Server (Arc/IMS). A web application could allow searching the database for specific spatial data by means of the stored meta information, displaying the selected data as a map and, for users within the Commission, downloading datasets.

36. For this purpose the new GISCO concept distinguishes between 3 groups of users depending on their skills in GIS techniques and their demands to the system.

37. The first group consists of GIS specialists. They are able to use the GIS system as it is without prior customisation. They are doing spatial analysis and are creating new datasets or modifying existing datasets which are integrated into the GISCO reference database. They are familiar with the internal data structure. These are typical Arc/GIS and Arc/INFO users.

38. The second group is occasional GIS users, therefore they access the spatial data via a graphical user interface. They are also performing spatial analysis and are producing maps. The results are not integrated into the GISCO reference database but are stored in their local environment. These are typical Arc/View users.

39. The third group consists of inexperienced GIS users. They are performing basic analysis, the emphasis of their work lies on data visualisation and identification functions. There is no need to modify or copy the data locally. This type of user is not yet considered at GISCO. We want to introduce this new user community that is not trained in GI systems and software. Therefore the spatial information shall be accessible through a web browser, generating maps on an ad hoc basis. Browser interfaces for accessing and displaying the spatial and non-spatial data have to be specified and designed.

40. The above-described concept is illustrated in the following figure 2.

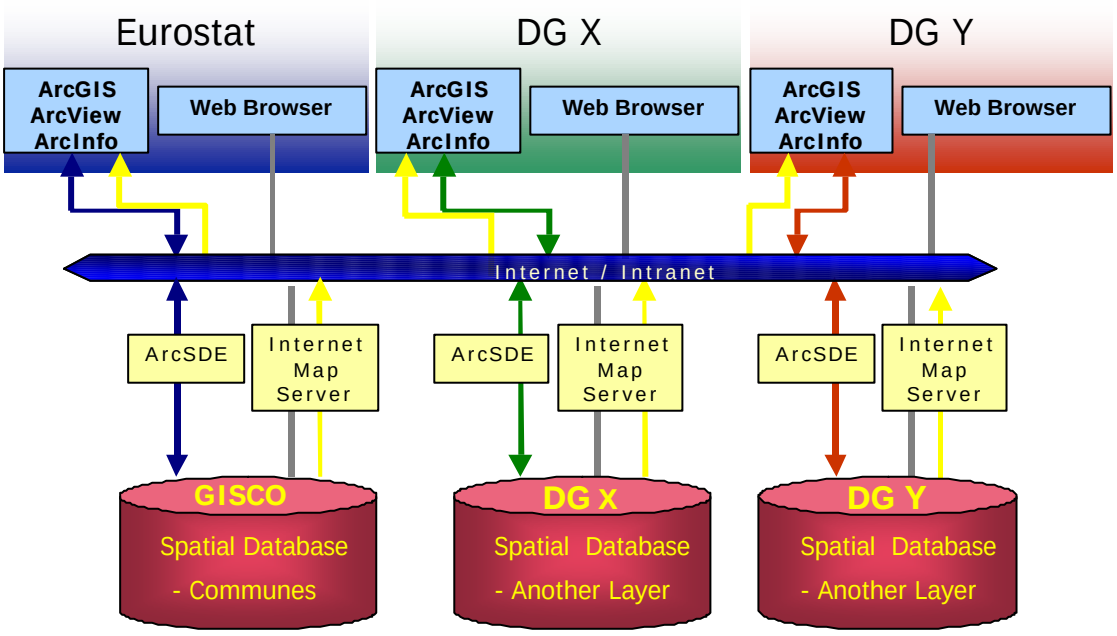

Figure 2 : New GISCO environment

III.4 Meta Information

41. The data dictionary and the database manual should be converted into one application. As for the spatial data, the meta data will be stored in a database too. The newly published ISO 19115 standard shall be applied when creating the new metadata structure. The ISO 19115 include a UML model that can be used for expressing properties, relationships and functions of the metainformation.

42. For migrating the meta information, the ISO 19115 UML model has to be adapted to the contents of the current metadata applications and should be implemented with the selected Arc/Info software.

III.5 Migration of the Reference Database

Project on new GISCO architecture

43. For the third year of the joint project with the JRC on the new GISCO database structure, it was decided to concentrate on the conceptual work for the introduction of the new architecture. The new version 8 of Arc/Info has been identified as being suitable to implement the new data model. The spatial and attribute data will be stored in an Oracle 8I database applying the Geodatabase concept of Arc/GIS. The Arc Spatial database engine should act as the interface between the Arc/Info software and the spatial database. The Unified modelling language will be used for data modelling. For the preparation of the migration of the current GISCO database, several preparative activities have been identified.

44. The existing logical data model will be converted to UML and implemented as Geodatabase. The logical model represents the core part of the database and has to be enhanced by the different user's views of the database in relation to its application.

45. In order to tackle the problem of temporal changes and updating the database, a study on updating the SABE[6] administrative boundaries will be performed. The study includes descriptions on updating strategies with an applied example.

46. The issue of spatial generalisation will be the content of another study. The study should be used for implementing procedures for automatic rule-based generalisation of the administrative boundaries dataset to smaller scales.

47. The logical datamodel needs to be physically implemented in a database. Depending on the type of data, its structure, the perceived applications, the hardware and the applied software, there are multiple ways for a database implementation. In order to discuss these issues, a review of strategies for handling large databases was planned. The results of this study will serve as a reference for implementing the spatial database.

48. Another issue is the dissemination of spatial data within the current dissemination structure of the European Commission. A report will analyse the capabilities of Arc/Info for data access on the web and make suggestions for an implementation within the current intra- and internet applications of the Commission.

49. Finally the current meta data structure of the GISCO database will be examined and analysed. Based on the ISO 19115 standard for metadata, a new structure for the GISCO metadata will be proposed and ways to migrate the current metadata to the new structure will be described.

Follow up of the project

50. As a follow up of the joint project on the new database structure, a prototype will be developed of the future production environment for the GISCO reference database. Based on the proposed data model and on the results of the previous activities, a prototype database will be implemented that covers the various activities of GISCO. The Geodatabase concept of ESRI will be the basis for the implementation. The prototype will include the conversion and storage of a selected number of spatial datasets (feature classes and related attribute data) as well as the implementation of specific functions to be applied on the spatial database. The purpose is to prove the feasibility of the proposed architecture and its various components.

---

[6] SABE : Seamless Administrative Boundaries of Europe. Datasets containing the administrative boundaries of Europe down to the local level (NUTS5)

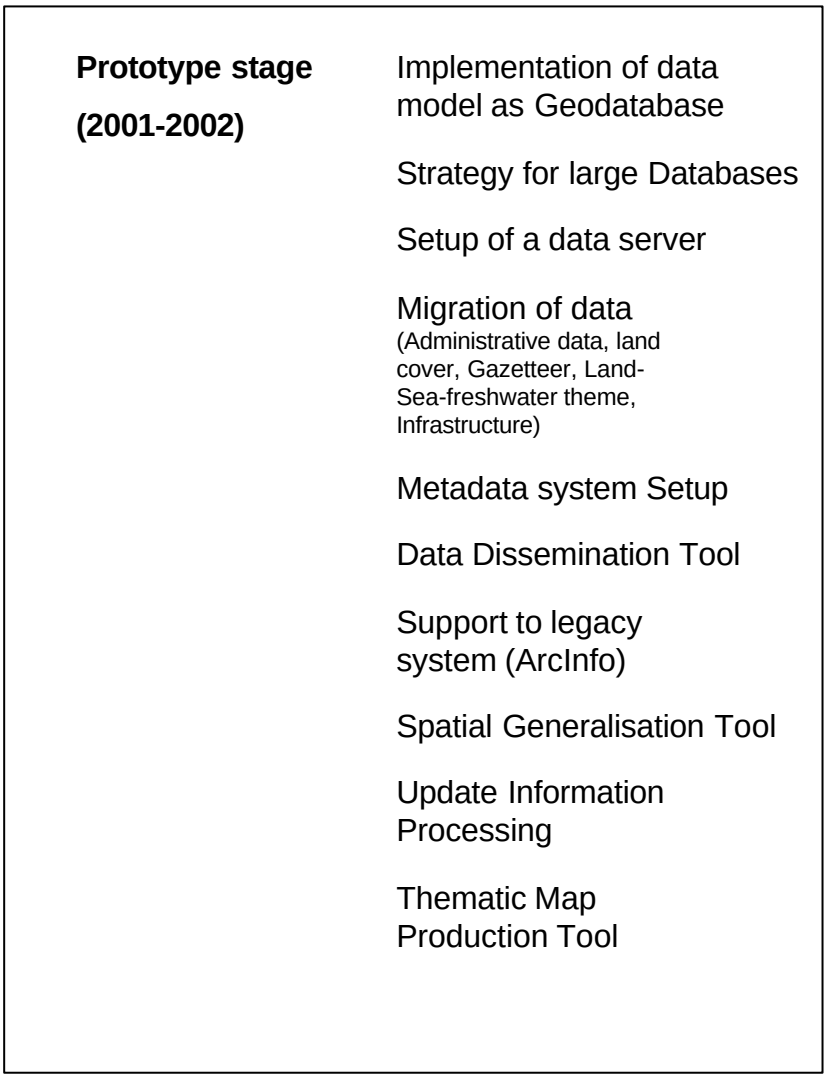

Figure 3 : ***Description of the prototype stage***

Data issues

51. The Pan-European coverage of satellite images from the IMAGE 2000[8] project could be used as a generic reference layer. As these images will be the reference for mapping land cover changes it can be expected that both datasets will fit geometrically. The prototype of the database system can be used for specifying requirements concerning quality aspects, integrating new datasets into the database, specifying spatial relationships between different layers as well as methods for implementing and maintaining these relationships.

52. The CORINE[9] land cover data is currently updated. The first step of the project consists of producing a Pan-European satellite image coverage, that can serve as a generic spatial reference for

[7] Arc Macro Language
[8] IMAGE2000 is a project of the European Environment Agency in the framework of the CORINE Land Cover update to create satellite image coverage of Europe.
[9] CO-oRdination of INformation on the Environment

data in a source scale of 1:100 000. The first satellite images will be available at the end of 2001. The updated land cover data will be available in 2002 or 2003.

53. The system can be dynamically expanded for new datasets that conform to the reference scale 1:100 000 – 1:250 000, like a newly acquired digital elevation model (DEM). A DEM can be used to specify relationships between thematic raster data, the DEM, and vector data, e.g. the water patterns.

54. EuroGeographics[10] presented its priority projects at the Eurostat Working Party in October 2000. Amongst these were the EuroGlobalMap which is a 1:1 million topodatabase. This database will be an extension to a pan-European dataset based on the development of the MapBSR for the Baltic Sea Region. The first release of the MapBSR was in 2000. The first version of the EuroGlobalMap is foreseen in November 2002. A further development of the EuroGlobalMap would be the Euromap in a 1:250.000 scale. The development of this map is conditioned on a funding from the European Commission of 50% of the total costs.

[10] EuroGeographics came into being 1 January 2001. The organisation is a result of a merge of CERCO (Comité Européen des Responsables de la Cartographie Officielle) and MEGRIN (Multipurpose European Ground Related Information Network).