INTEGRATIVE GEOSPATIAL DATABASES

A. Garagon Dogru [a, *], G. Toz [b]

[a] BU, Kandilli Observatory and Earthquake Research Institute, Geodesy Department, 34684 Cengelkoy Istanbul, Turkey - garagon@boun.edu.tr
[b] ITU, Civil Engineering Faculty, 34469 Maslak Istanbul, Turkey - tozg@itu.edu.tr

KEY WORDS: Integration, GIS, Databases, Geospatial Data, Spatiotemporal Data

ABSTRACT:

Modern geographical information systems make it possible to integrate information from multiple sources in different forms for querying and analysing information. So the importance of geospatial databases is growing from day to day, and their applications extend beyond the traditional applications of geospatial databases and GIS. One of the most recent challenges is integrating geospatial data from heterogeneous sources with varied formats and positional accuracy into databases for GIS applications. Another research is required to integrate time into the databases because geospatial data and its applications are not always static. The aim of this study is to provide an overview of current status of researches and future trends in geospatial databases.

1. INTRODUCTION

Geospatial data is a subset of spatial data which indicates where things are within a given coordinate system. It includes details about characteristics, relationships to other things or ideas, and the dimension of time as it relates to all of these. Geospatial data, which is also called geodata, has terrestrial coordinate system that can be shared by other geospatial data. And there are many ways to define a terrestrial coordinate system and also to transform it to any number of local coordinate systems. Geospatial data comes in many formats from various resources, and its structure is more complicated than non-spatial data.

Geospatial datasets allow users to integrate data with time as well as space, concept as well as location, relationships as well as values. Geospatial data includes dynamics over time. Physical characteristics of geospatial datasets include projection, datum, scale, accuracy, topology, data model, attributes, and metadata. They should be standardized among all datasets and these datasets should be stored in the most economical and functional format possible as well.

2. DATA INTEGRATION

Today GIS, which contain geospatial databases, is used in the applications of:

- *monitoring* to keep an eye on the state of earth systems using satellites and monitoring stations (water, urban development, security), data are available in near-real-time on the Internet,
- *analysis and risk assessment* to find the problem areas and analyse the possible causes (soil erosion, earthquake, pollution, flood…),
- *modelling and simulation* to predict consequences of human actions and natural processes, and
- *planning and decision support* to provide information and tools for better management of natural resources.

However, the expected benefits of GIS usage were overestimated, since the regional co-ordination of risk management was blocked by incompatible data formats and inconsistent datasets. Lack of institutions cooperation in spatial data management resulted in different spatial object classification and visualisation, thus generalisation of maps to the regional level was almost impossible. (Merson, 2004) In the past, 80% or more of a project's resources were consumed for converting paper records to digital databases. Today, online digital data is available and it increases the power of GIS. But it is still to spend 80% or more of a project's resources on searching, discovering, retrieving, and reformatting data. Today, people who need geospatial data must visit many web sites, each having their own appearance, and format. The data is distributed across many servers, many data centres, and managed by many different organizations.

Data integration is the capability to connect different datasets together. The aim of data integration is to make users access distributed, heterogeneous, and administered datasets more easily. If users need access to datasets that have been acquired by others or that are stored under different administrative controls, they may face with a range of barriers. Overcoming these barriers calls for a number of responses. For example, it requires datasets to be described in sufficient detail for other users to be able to understand their meaning; it requires agreement on standards for description (metadata) and for the adoption of compatible computing infrastructures, and it requires organizations holding datasets to agree on common policies for sharing and access. Data integration is technically difficult because there are many different formats in which data can be encoded and different types of database software, data may be held on many different computers which belong to different organizations (administrative domains), each with

[*] Corresponding author.

their own mechanisms for user authorisation and authentication. (“What is”, 2006) These barriers prevent easy access to geospatial data. Fortunately data integration is about removing such barriers. A GIS which involves the integration of data also provides significant efficiency by enabling integrated spatial analysis of geoscientific data.

Single data producer can not produce useful datasets and information without integrating data from others. So data must be interchangeable. This means that data from different sources can be integrated. Mechanisms are needed to integrate data with different spatial reference systems, different semantics, and different formats. (Peng, 2003). Datasets come from different sources are problematic because no dataset is perfectly accurate, and different data entry approaches may cause different results. Another problem is with the geodetic base (i.e. projection, coordinate system, datum) of each data being different. All of these problems can be solved by forming metadata. Metadata is needed to automate the process of search for data, to determine the fitness of a data set for use, to handle data effectively and to identify notable data contents.

Data integration challenges can be categorized as follows:

- Syntactical Heterogeneity - heterogeneous data format (e.g. 30-03-2006 vs. 30/03/06)
- Structural Heterogeneity - heterogeneous data models and schemas (e.g. 30-03-2006 is saved as three columns or one columns)
- Semantics Heterogeneity - fuzzy metadata, terminology
- Others - different platform, database systems, data type, units, accuracy, resolution

2.1 An Application Example

Thousands of earthquakes have been recorded worldwide and scientists try to learn from these events to construct analytical and numerical models and predict the future distribution of earthquakes in space and time. This requires a careful understanding of historical earthquake events and a combination with field data. Integration of spatial information plays a prominent role in this process and GIS is used as a powerful tool to integrate earthquake related data.

The North Anatolian Fault is one of the most seismically active faults of the world. It runs along the northern part of Turkey about 1500 km, from the Karliova to the North Aegean (Figure 1).

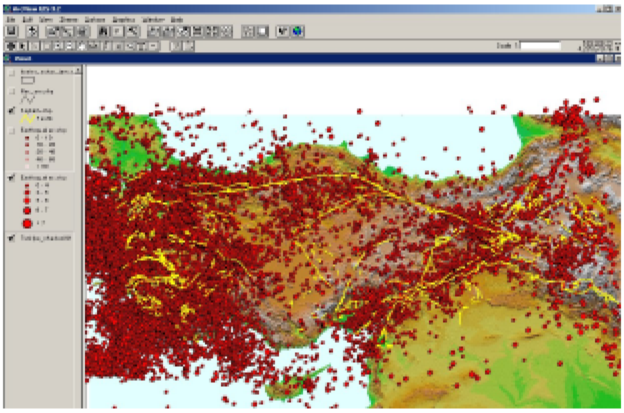

Figure 1: Seismicity of Turkey

A study, called Interactive Earthquake Information on the Internet for Turkey, was realized in 2002 for the integration and dissemination of geoscientific datasets (Figure 2) The data integration phase of the study was the most time consuming. For the preparation of geospatial data ArcView, ArcInfo, Erdas Imagine, and Microsoft Photo Editor Software were used. Digital Elevation Model downloaded from the US Geological Survey’s web site was exported to cell based grid format. ArcView was used to classify grid data according to its height values. In order to use raster data in IMS application, data were converted from grid format to JPEG image format. Since the TIFF format is displayed more efficiently than JPEG format, a conversion was performed using Microsoft Photo Editor Software. For population density map, ArcView and its scripting language (Avenue) were used to calculate areas of the boundary polygons and colorize shape file according to the population density values. Population values of districts were added to districts dBASE table. Density image was exported to JPEG format. And then JPEG file was converted to TIFF format. Vector data (downloaded from Digital Chart of the World’s web site) were acquired in shape file format. For other ASCII text files which include coordinate information of stations and earthquake parameters information, ArcView scripting was used to convert text file to shape file format. All of these data were in geographic coordinates relative to the WGS-84 and ED-50 datums. Datum transformations and coordinate conversions were performed. Data were projected to Lambert Conformal Conic map projection. ArcView can not export maps to georeferenced images. So topographic map image and population density map image was georeferenced to the Lambert Conformal Conic projection by using image processing software ERDAS Imagine. (Garagon Dogru, 2005)

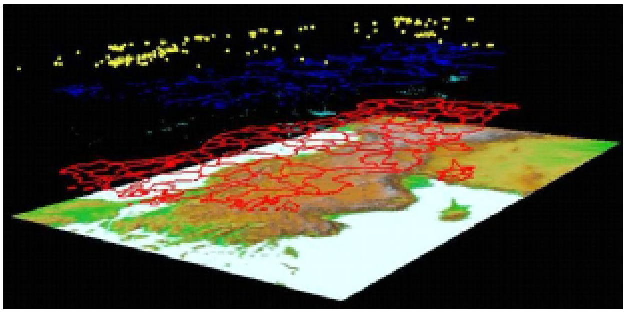

Figure 2: Datasets which are on the same spatial reference system

In this study, a database of over 30.000 earthquakes in Turkey from the years 1900 through 2005 has compiled. The earthquake data include information such as date, location, time, and magnitude. In order to disseminate earthquake information to the public, this non-spatial database is transformed into a map layer that can be displayed along with the other geospatial information. The end result is a web-based GIS that can be used for analysis. Utilizing the map server interface, the end-user can browse the earthquake information by clicking on an event. Event symbols are graduated and coloured depending on earthquake magnitude and depth. Locations of seismic stations are also displayed. The end-user can activate and query other map layers such as faults, topography, and population information.

3. SOLUTIONS

The development of GIS and the interoperability between these systems demands new requirements for the description of the underlying data. The exchange of data between GIS systems is problematic and often fails due to confusion in the meaning of concepts. (Visser, 2002) Standards facilitates seamless sharing of data and products, moreover, it enables smooth integration of disparate datasets. Uniformity in models, datasets, and rendering are essentials to the success of enabling the reuse of valuable and important datasets. (Chhatkuli, 2005) According to The Federal Geographic Data Committee (FGDC) Standards Reference Model (March 1996), the four basic categories of the information engineering standards are data, processes, organizations, and technology.

FGDC is an interagency committee, organized in 1990 to promote the coordinated use, sharing, and dissemination of geospatial data on a national basis. FGDC is responsible for the development of the National Spatial Data Infrastructure (NSDI), an organization composed of 17 federal agencies, state, local and tribal governments, the academic community and the private sector. NDSI is charged with the coordination of policies, standards and procedures that help organizations to cooperatively produce and share geospatial data. Designed to reduce the cost of geospatial data collection and minimize duplication among federal agencies, the NSDI serves an important role in the management of geospatial datasets. Common elements of the NSDI are metadata, clearinghouse, standards, framework, geospatial data, and partnerships. Each of these components serves as a cornerstone in establishing consistency and structure when it comes to documenting spatial data for everyday applications, as well as in building a distributed network of producers and users that facilitate data sharing. (“The Federal”, 2005) We need SDI in a national scale because it integrates distributed providers of data and makes the related institutions share costs of data creation and maintenance.

Any third party users might require datasets from different specialist agencies to obtain a meaningful solution to their problem. But the situation at present is that it is very difficult to find out whether data of a particular kind existed or not and if exited whether such data is useful for a particular application or not. The prospective user might then start creating datasets on its own, when such datasets might be actually in existence somewhere. The clearinghouse concept might be the only answer to such a situation. Metadata of data from the producers be created and can be served to the users through a clearinghouse. Such an effort requires a consensus among the data producers in the first place to make available a well-documented metadata to the clearinghouse administrator. The problem would be there if some or the other organization do not prefer to participate. Such circumstances need intervention through well-developed policy guidelines. (Chhatkuli, 2005)

Another important organization in this field is the OGC, which is a consortium of over 250 companies, agencies and universities working toward a world in which everyone benefits from geographic information and services made available across any network, application, or platform. The OGC is focused on geospatial information and services. Here "systems" means software processes, services and other components, including data, semantics, hardware, and networks. The OGC works to make geospatial information and services a fluid part of the World Wide Web, and to likewise enable interoperability across networks, systems and enterprises. (“The Open”, 2006) Many agencies in the world use OpenGIS specifications to establish a common understanding of an emergency situation and coordinate incident response, bringing together different data from different platforms.

4. CONCLUSIONS

The purpose of data integration is to combine data from heterogeneous and multidisciplinary sources into one coherent dataset. The sources of data typically employ different resolutions, measurement techniques, coordinate systems, spatial or temporal scales, and semantics. (“IT Roadmap”, 2006)

Data in different databases and multiple sources may come in different resolutions, precision and data formats. Overlay of this data requires a pre-processing stage to make sure that the data layers from the different sources fit spatially. To achieve this, there is need for mechanisms to detect and adjust inconsistencies in the spatial data. Chaotic distribution of available data sets, lack of documentation about them, lack of easy-to-use tools to access them, and lack of communication among organizations (non-interoperability, redundancy) are the main issues in this field.

Many research disciplines are involved in data integration: from computer science and database research to artificial intelligence, the Semantic Web and Description Logics. We can certainly learn from these other disciplines in trying to improve the structural and semantic interoperability of geospatial resources of different organizations and agencies involved in emergency response. (Vries, 2005) We highly need cooperation between organizations in Turkey to use geospatial information during and after emergencies. This cooperation is required for increased awareness of seismic hazard in the general public and increased scientific understanding of seismic hazard as well.

REFERENCES

Chhatkuli R.R and Kayastha D.M., 2005. Towards A National Geographic Information Infrastructure: Overcoming Impediments To The Development of SDI in Nepal, *FIG Working Week 2005 And GSDI-8*, Cairo, Egypt, April 16-21.

Garagon Dogru A., Toz G., Ozener H., 2005. Spatial Information Retrieving Using Internet, *SETIT 2005, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications*, Susa, Tunisia, March 27-31.

Merson, M.E., 2004. *Manage Data - Manage Hazards*, MSc Thesis, International Institute for Geo-Information Science and Earth Observation, The Netherlands.

Peng, Z. and Tsou, M., 2003. *Internet GIS*, Wiley, USA

The Federal Geographic Data Committee. http://www.fgdc.gov/ (accessed Dec. 2005)

The Open Geospatial Consortium, Inc. http://www.opengeospatial.org/ (accessed Jan. 2006)

Visser U. et al., 2002. Ontologies for Geographic Information Processing, Computers & Geosciences, Volume 28, Issue 1, February.

Vries M., 2005. Recycling Geospatial Information in Emergency Situations: OGC Standards Play an Important Role, but More Work is Needed, http://www.directionsmag.com/ (accessed Apr. 2006)

What is Data Integration? http://www.ncess.ac.uk/insight/tutorials/datagrids/data_int/what_is_d_int (accessed Feb. 2006)