GEODEZJA I KARTOGRAFIA
t. LI, z. 1–2, s. 55–68 (2002)

Elżbieta Bielecka

Institute of Geodesy and Cartography
(00-950 Warszawa, ul. Jasna 2/4)

Data integration in a seamless geodatabase. Selected problems

The paper is organized into two sections. The first sets the background for data integration, and identifies some of the key issues that need to be considered. The second describes possible solutions of problems connected with the data integration. The steps of integrating data coming from different databases, datasets and data files are described in order to create a seamless geodatabase.

The necessity for data integration

Nowadays, there is a rapid growth of the availability of digital spatial data and a growing need to use it for all kinds of GIS applications and to support the decision-making process. The development of communication technology makes it possible to collect datasets from a variety of sources and different types of application. Data providers make data available to users via the Internet. It seems to be a lot of databases, datasets, and other geographical information like satellite imagery, aerial photographs, and maps in digital and analogue forms. It also becomes possible for every user to share some spatial data, and not to collect them from the very beginning. Sharing data requires, first of all wide information about the scope of data, and the place where they are stored, furthermore translation from the original source of data into the users system and adaptation to specific GIS applications. The data adaptation process could be called data integration. Data integration is the most valuable function of GIS, and the data that is integrated meets user needs more precisely.

Data integration means combining of data files, datasets and databases originating from different sources into a one common database. Hence unification of codes, defining models of objects and data definitions is of the utmost importance. Integration of spatial data consists also in creating relations among various categories of descriptive and geometric data, as well as joining them.

Integration is not data transmition and conversion from various systems to one homogeneous database environment. Its meaning is much more wider. The data transfer (transmition among systems or conversion from bases and lower generation software to higher) is only a technical part of the data integration process. Data transfer is characterized

as formal and it is based on uniform and constant structures. However, data transfer is the first step to overcome in data integration.

Data formats that refer to the logical structure for most GIS applications contain only enough information for the originating GIS application to be able to use it properly. The data formats usually carry the features, units and some basic projection information. Almost all GIS software has its own internal file format, usually proprietary. They are not designed for the outside native system, that is why most systems also support transfer file formats.

Data exchange formats are usually more robust. They typically carry information that would allow the use of the data in a variety of systems so they are generally standardized and well documented. Exchange formats usually carry some minimum metadata to describe the dataset as well as data quality statements. Producers of data typically use data exchange formats. Unfortunately, format transfer systems do little to support translation of semantics. The real problem is lack of common semantics.

Integration starts from transferring data from its native applications to a destination database management system, next merging them and performing all activities needed to create a database. Two significant conflicts: semantic and spatial have to be solved at the very beginning.

Semantic conflicts

Data integration requires an intimate knowledge of data semantic in order to preserve it. A critical point for the integration of the thematic datasets is semantic heterogeneity caused by different interpretation of data. Diversified understanding of meaning is also connected with classified objects especially in the environment (e.g. soil, vegetation, land cover). During the classification process, each entity type must be uniquely defined to preclude ambiguity. Semantic integration is additionally complicated by different applications schema used in GIS software. The most frequently occurring semantic conflicts are: conflict of names (homonyms and synonyms), meanings and schemas.

Homonyms arise when the same meaning concerns different objects. A typical example of homonym is the item code. Almost every dataset includes the item code that describes entities by the code number. The value of the code sometimes maybe the same but it does not mean that the meaning is also identical; on the contrary the meaning differs. Synonyms occur when different words describe the same feature in the database, for example object road in one dataset could be characterized by attributes of on express road but in the other – highway. After joining data one should recognize that they are synonyms. Misunderstanding of meanings (terms) is caused by different definitions and interpretation of feature. A good example is a street, which is usually defined as an area within two edges in land information systems, or only by an axis – in GIS. A process of schema integration should overcome discrepancy in database schemas.

Recently, problems of semantic conflicts in database integration have been under research. One way is using ontology for resolution of semantic heterogeneity in GIS (F., Hhakimpour, S., Timpf, 2001).

Spatial aspect of integration

The assurance of data continuity and topology is a very important aspect of spatial data integration. Data mismatch can stem from many factors including incompatible projections, inconsistent map units, and different plotting scales. Differences in the relative age of data sets may mean differences in data collection methods and accuracy. The improper application of a datum to a dataset is an increasingly common and very important cause of data alignment problems. All these discrepancies and others should be removed during the integration process.

Data integration means also the implementation of vector, raster, TIN (triangulated irregular network) and other data models into one seamless geodatabase, and using them for analytical purposes and spatial modeling. Usually data integration is time consuming and expensive. As a result we have data well structured from an analytical perspective.

Steps towards an integrated database

1. Data transfer to the internal file format used in GIS software.
2. Examining the data (entities), solving semantic conflicts.
3. Transforming data to the fixed projection and the co-ordinate system; unifying map units.
4. Spatial data merging (within one thematic layer):
   - generalization to provide similar data details
   - edge matching and map joining (including rubber sheeting transformation)
   - error correction and entering missing data
   - forming topology
   - verification of data consistency and error correction
   - attaching attributes.
5. Vertical data matching (among different thematic layers covering the same area).
6. Converting data to the appropriate data model.
7. Indexing.

The afore mentioned steps describe the general problem of data integration. Some activities may be neglected according to the data diversification and existing discrepancies. However examination for solving semantic and spatial conflicts is always required.

Geographical Information System for delimitation of Less Favoured Farming Areas in Poland – case study of data integration

The GIS for the delimitation of Less Favoured Farming Areas (LFA) and the database collecting essential data has been implemented. The database that stored data in the structural and imposed manner can best serve the purposes of LFA quantification. The implemented data model was a hybrid one also called georelational (ESRI, 1997). The

hybrid model used a set of files that contained coordinate and topological data and stored attributes elsewhere in DBMS tables. Each feature has a unique identifier that links it to a row in a DBMS table. The drawback of this model is the impossibility of simultaneously optimizing the data store mechanism for spatial and tabular data. The benefits are: quick spatial analysis, good display performance and reasonable access to attribute data.

All indispensable data is already in digital form. It has been stored in heterogenous databases or files; with heterogeneity relating both to DBMS and GIS software as well as to diversity of data structure. Diversification was, however, not the main problem when data was integrated. The main problem we had was almost a complete lack of metadata and a data dictionary. Although data documentation should have been the concern of data providers, and therefore a matter of trust, the documentation that we received from data suppliers was rather modest and insufficient. The data description attached to the files usually consisted of data contents and sometimes of data accuracy and information pertaining to the co-ordinate system. Our knowledge of data models and data structure was hampered due to a lack of pertinent documentations. We should complement the information about data by examining the data in detail and studying software requirements.

A particular effort was given to the design of the LFA database followed by the integration and harmonization of heterogeneous datasets and data files. The source datasets had to be re-engineered due to the conceptual and logical schema of the LFA database.

The created LFA GIS database maintained by ARC/INFO comprises both geometrical and descriptive data. The following information is indispensable to define LFA (Bielecka E., 2001):

- the administrative units of Poland,
- land use/land cover,
- site characteristics (altitudes and slopes),
- protected areas,
- soil characteristics,
- Agriculture Quality Index (AQI) values,
- demographic data (in terms of a commune): population, population density, inhabitants effecting farming activity, education of inhabitants effecting farming activity,
- the data derived from the farm census (in terms of a commune): number of farms, farms by occupation, farms by growth prospects, farms by area groups, farms by production branches.

The main guidelines concerning data quality are: up-to-date information, positional accuracy 1:100 000, the descriptive statistical information referenced to the level of communes, completeness of the data, and derived from reliable sources that guarantees the credibility of the data.

The data models

Geodata represents only a picture of reality from a certain perspective. Different LFA perspectives lead to different data models because the LFA GI Systems database employs the following data models:

1) a topological vector model for representation of discrete spatial data;
2) a TIN model for representation of the DTM;
3) a raster model for continuous-type data;
4) a relational data model to serve the descriptive information.

Under a topological vector representation all data regarding the course of administrative structure boundaries of Poland was saved as well as the information on the situation of protected areas and soils. As a vector data model requires it, the topological structure is saved within ARC/INFO in the form of Arc-Node. The Arc-Node structure enables to group polygons into regions and assign attributes to them. This is important that one real world object (a commune, a protected area) is composed of several objects within the database. Such a topological representation of vector data permits for more effective management of the database and improves the Systems analytical capacity (Zeiler M., 1999).

A TIN model was chosen to represent a digital terrain model as it features considerably high representation efficiency when a diversified mountain relief is of concern. TIN algorithms are capable of representing lines and discontinuity surfaces as well as isosurfaces. This considerably increases the quality of the results (Carrara et all, 1997). Interpolation employs a QUIMTIC (five order polynomial) method implemented in ARC/INFO that is capable of generating a continuous and smooth surface.

A raster data model was chosen for representation continuous data such as: land use, altitudes and slopes for which time-consuming, multi-variant spatial analyses have been made. ARC/INFO saves a raster as a GRID using a Run-Length-Encoding technique for data compression, thus enabling a very reliable reproduction of the original information (ESRI, 1990).

Taking advantage of possibilities of GRID consisting in an integrated management of many attributes assigned to a raster, raster layers with associated parameters containing information on the commune assignment, land cover categories, altitudes and slopes were created. The targeted approach i.e. saving all such data within a single raster and a relational tables feature increases the LFA Systems flexibility and analytical power, diminishing the database capacity at the same time. Also, the raster model features a lack of constraints related to the raster size and the number of columns and lines and attributes assigned to the raster cells. A raster of 100 m in size (a 1 ha area) was adopted in GRID coverages.

A relational data model, developed by E.F. Code in the 1970st80s became a basis of the architecture of popular RDM systems. Because of its syntactic uniformity, interrelations with the algebra and a smart representation, a relational approach based onto a single basic data structure has seen its formal and careful synthesis. By way of simplification, a relational database may be considered as a set of tables, the lines of which describe the entities or interrelations occurring within the modelled world as well as a set of semantic relations describing general principles and rules to be observed within a database. The objects and their interrelations are represented by the syntax always in the same manner that is considered to be one of the most essential advantages of the relational approach. A relational structure governed by INFO houses descriptive data.

The adopted presumptions concerning the data models employed here are optimal because of their required accuracy, analytical capabilities and economical reasons, meant in terms of optimisation of the processing time.

Diversified sources data

As the process of gathering detailed data is extremely time-consuming, costly and complex – from both technical and organisational point of views – only indispensable data was gathered within the dedicated information system that permits to obtain the expected results possibly within a short time and at minimal expenditure required. The database that serves to define LFA takes into consideration information on land cover, the administrative structure of Poland, attitudes, slopes, protected areas, and statistical demographic data concerning population density, education of inhabitants, population by occupation, as well as farm census data (number of farms, growth prospective for farms, systematic of farms). Over 10 different data sources were used to build a seamless geodatabase and an equal number of data integrating techniques were necessary.

The administrative structure

The dataset administrative structure was prepared in-house, at the Institute of Geodesy and Cartography, in 1998. Since then it is regularly up-dated. It used the Gauss-Krüger projection, and the national co-ordinate system ''1942''; because of the co-ordinate system administrative structure is stored in two separate ARC/INFO coverages (one for zone 3 the other one for zone 4). The geometric part of the database stores vector information about the course of administrative boundaries and its accuracy corresponds to a 1:100 000 map. The descriptive part of the database is comprised of the statistical code of communes, their names and the former statistical codes (from before the reform).

Land cover

Information on land cover has been obtained as a result of visual interpretation of images taken by Landsat TM, in accordance with the Europe-commune methodology CORINE Land Cover. The methodology features a hierarchic structure and has 44 categories of land uses (at 33 forms existing in Poland) at the national level corresponding to a 1:100 000 map. The original CORINE Land Cover data is stored in ARC/INFO coverages corresponding to map sheets, in the ''1942'' co-ordinate system. The information on land cover has been coded. The code system is arranged hierarchically to enable further classification of the information.

Digital terrain model

Data indispensable for the creation of DTM has been prepared by the Institute of Communication. A set of points with geographical coordinates ϕ, λ, H has been derived

from 1:50 000 topographic maps. Points are stored for each 1:100 000 map sheet separately, in ASCII files. The spatial resolution of DTM is 250 m, whilst the altitude accuracy is not below 20 m.

Protected areas

The class of protected areas covers national parks, natural beauty parks and protected landscape areas as well. The database contains information on the situation of a given site in the year 1998, its name, and area and identification number. The protected area information is stored in a shapefile – non-topological vector format, in the "1942" co-ordinate system, separately for each type of protected area. The owner of the database is the Institute of Environmental Protection. The descriptive data relating to the protected areas is on the Internet page of The Ministry of Environment.

Soils

The database of soils in Poland as a part of the European Soil Database was created following the methodology developed by a EU working team, at the Warsaw Technical University, by the Faculty of Geodesy and Cartography. The methodology distinguishes two units: an SMU (Soil Mapping Unit) and a STU (Soil Typological Unit) including information about a soil sub-type. Data is stored in vector format, in Albers projection.

The Agricultural Quality Index

This is descriptive information covering a synthetic evaluation of four environmental components namely: soils, climate, relief, and water system as well as the summary value of the Agricultural Quality Index. Data should be attached to communes using the statistical code as the joining item. The digital form of the Agricultural Quality Index was prepared in EXCEL spreadsheet by the Institute of Soil Science and Plant Cultivation, in 1990.

Statistical data

General statistical data, demographical data and farm census information in terms of communes is derived from the Statistical Office in DBF format. The statistical code of communes originated from a register of identifiers of terrestrial units of Poland TERYT. The statistical code of a commune is unique and the rule of its creation enables us to group communes into districts and provinces.

Bottom-up approach

As all indispensable data has already been changed into digital form, the bottom-up approach was chosen to build up a geodatabase. Data was fragmented into both zonal (geographical or spatial partitioning) and thematic (layer partitioning) sets. Different data files and datasets should be then integrated, for the delimitation of LFA. Data integration in the meaning of the LFA application denotes consolidation-fragmented data into the centralized seamless spatial database maintained by ArcInfo. This reveals inconsistencies in data accuracy and quality that had to be overcome. In the bottom-up approach during geographic (spatial) integration, due to inexact matching at the zones or map sheet boundaries, some difficulties occur in order to ensure geometric and topological continuities between the dispersed databases. During thematic data matching the main difficulty is the existence of some discrepancies between the positioning of objects. After merging the problem is similar to sliver polygons, which should be deleted.

Difficulties should be overcome

During data integration both semantic and spatial conflicts were solved. The following discrepancies were detected and removed:

The diversity of spatial representation of geographic information (homonyms and synonyms)

A serious problem connected with homonyms was fitting the different meanings of the item code into a common identification system. The item code exists in the following sources coverages: Administration, CORINE, and Protected Areas as well as in The Agricultural Quality Index table and Demographic data table, but its meaning is different. In the Administration data set code means the statistical code of communes, in the CORINE data set – the code of land cover category, in the Protected Areas data set – the identification statistical number of protected areas. The code attached to communes in the Agricultural Quality Index table was the previous one (from before the reform). The item *code* was renamed and an abbreviation describing the thematic scope was added after the word code e.g. *code_gm* means the statistical code for communes (gminy), *code_CLC* – the category of land cover classification, *code_park* the statistical identifier for national parks, etc.

The diversity of global projection and discrepancies in co-ordinate systems

Four co-ordinate systems based on different projections were used to store source datasets: geographic, Albers with datum Bassel, ''1942'' zone 33 and 34 (Gauss-Krüger projection using datum Krassovsky as well as datum WGS-84) and ''1992'' destination co-ordinate system (Gauss-Krüger projection, datum WGS-84). Finally all datasets were projected to Gauss-Krüger and datum WGS-84 using the Bursa-Wolf seven-parameter method.

The diversity of types and values for the same item located in a different site

Diversity of the items type occurred in the case of the item *code,* describing the statistical identifier of communes. It has ''character'' type in statistical data and ''integer'' in Administration coverage. It was assumed that the code should have ''character'' type so ''integer'' was changed into ''character''. Diversity of value occurred in the item pertaining to the acreage of agricultural area between CORINE and statistical datasets. This discrepancy was left because of different source data. Only an explanation was added to the corrected documentation.

Discrepancies in data timeliness

Some of the data had to be up-dated. The geometry and attributes of existing objects have been altered, as well as the new objects, which have appeared into Protected Area coverage. New entities have been added to the Agricultural Quality Index tabular data.

The diversity of positioning accuracy

The diversity of positioning accuracy varied from 1:50 000 for the protected area boundaries to 1:250 000 for soil type boundaries. Boundaries of protected areas were generalized up to 1:200 000. The generalization process used ARC/INFO commands to erase, amalgamate and simplify the 1:50 000 – scale data for output at the smaller scale. Soil type boundaries were edited more precisely due to information about land cover and elevation. Figure 1 shows the example of generalization of administrative boundaries.

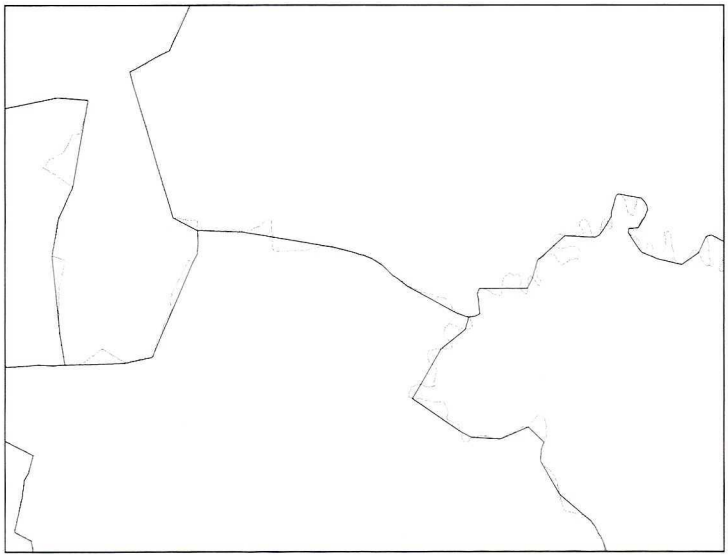

Fig. 1. Generalization of administrative boundaries

The boundary between commune Trzebowisko and commune Krasne goes on river current, hence we can observe the intricate line before generalization (dotted line) and simplified one after generalization.

Discrepancies in boundary alignment

A serious problem occurred while matching more than three hundred CLC datasets in order to create a seamless CORINE Land Cover coverage (Fig. 2). Spatial combining was also necessary to create Protected area and Administration coverages, because they consist of two zones 33 and 34. Spatial data merging was time consuming and sometimes required additional materials such as thematic and topographic maps as well as satellite images.

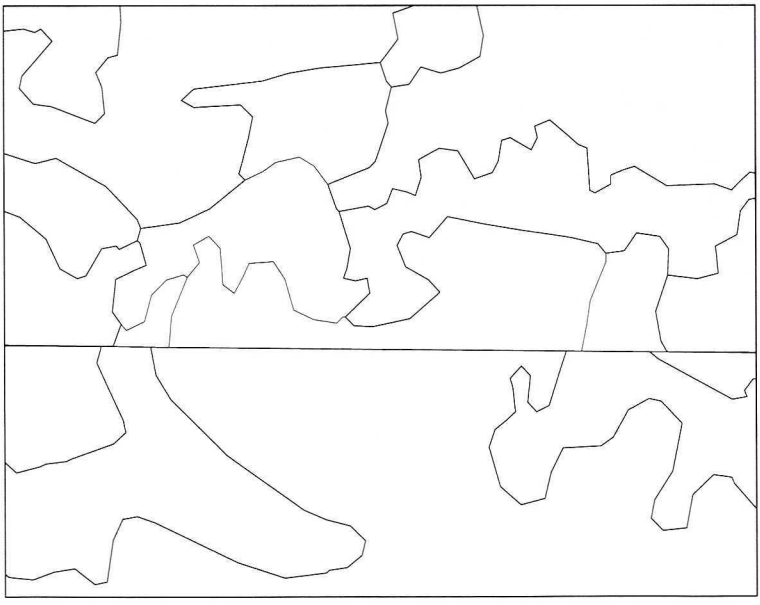

Fig. 2. Discrepancies in boundary alignment

Discrepancies in vertical boundary alignment

Vertical alignment was required when boundaries of the CORINE Land Cover category were compared with Protected area and Administration coverages. This stage was also time consuming. Additional supplementary materials were available for checking errors. Figure 3 shows discrepancy between boundaries of CORINE Land Cover categories (forest) and boundary of Roztoczański Landscape Park.

Fig. 3. Discrepancy between forest boundary (CORINE Land Cover categories) and boundary of Roztoczański Landscape Park

Creating relations between geometric and descriptive data

Farm census and statistical data was attached to the administrative units of Poland.

Finally topology operations were performed. Duplicate line segments from input datasets were removed, this occurred when CORINE Land Cover files were joining. Next, line segments (in the vector coverages) were connected to reduce the number of features by combining their geometry. The labeling was checked, and some other typical activities for ARC/INFO software were completed. Because forming topology also reveals many errors, so error correction should be repeated.

Database indexing was the next step. Two types of indexes were used: spatial indexes and item indexes to improve the analytical capabilities of the database. Spatial indexes increase the selection speed of graphical queries on spatial features and improve the function of any operation that retrieves coverage features by location. After indexing spatial selections, executions of IDENTITY improve 2 to 25 times. Improvements are due to a reduction in I/O as well as CPU. Item indexes increase several times the query speed of logical expressions against an INFO file item.

The creating of the data documentation was the final step in the building of the database. Data documentation is critical both from the perspective of users and data administrators (managers). Database documentation is accurate and up to date for coverages (vector and grid) and for tabular data. It is composed of two parts, the first one – Facesheet – provides

the basic information of the coverage including the history of data use; the second – Attributes – comprises documentation of feature attributes. This documentation is a part of the metadata that is currently under implementation at the Institute of Geodesy and Cartography.

CONCLUSIONS

Data integration is the most valuable function of GIS. Users should realize that the proper data integration usually requires a settlement of two conflicts: semantic and spatial. Resolution of semantic heterogeneity in GIS still requires more study in order to offer more efficient methodology. Spatial data integration requires extensive knowledge in the field of geomatics as well as technical infrastructure. Merging of different databases, datasets and data files is very complex, time consuming and expensive task. It should be solved in terms of geometry and topology.

The goal of this paper was to give an overview of the problems arising from dealing with dispersed data sources and to show some possible solutions. The database created for the purpose of delimitation of the Less-Favoured Areas in Poland was set as an example. As this database covers the entire country, and over 10 different data sources were used, almost all problems connected with data integration occurred. The established resolutions were introduced they are, however, not unique.

REFERENCES

[1] E. Bielecka, *Delimitation of Less-Favoured Farming Area in Poland – Methodology with the Use of ESRI Software.* The 16th EAME ESRI Conference, October 17–19, 2001, Lisbon, Portugal.

[2] E. Bielecka, R. Jankowski, *Delimitation of Less-Favoured Farming Areas in Poland using CLC Database.* PTL, Land Cover, EEA 2000.

[3] A. Carrara, G. Bitelli, R. Carla, *Comparison of techniques for generating digital terrain models from countour lines.* Int. J. Geographical Information Science, vol.11, no.5, 1997, 451–473.

[4] ESRI, 1997, *GIS Data Storage Trends. Implications for Utilities.* An ESRI White Paper, March 1997.

[5] ESRI, *Understanding GIS: the ARC/INFO Method.* Redlands, California, ESRI Inc. 1990, 2–27.

[6] F. Hhakimpour, S. Timpf, 2001, *Using Ontologies for Resolution of Semantic Heterogenity in GIS.* Proceedings of 4th AGILE Conference on Geographic Information Science, April 19–24, 2001, Brno Czech Republic, 385–395.

[7] R. Laurini, *Spatial multi-database topological continuity and indexing: Step towards seamless GIS data interoperability.* Int. J. Geographical Information Science, vol. 12 No. 4, 1998, 373–402.

[8] M. Zeiler, *Modeling Our Word. The ESRI Guide to Geodatabase Design.* Redlands, California, ESRI Inc. 1999, 46–60.

Received January 16, 2002
Accepted April 17, 2002

Elżbieta Bielecka

Integracja danych w jednorodnym środowisku bazy danych przestrzennych. Wybrane problemy

Streszczenie

Rozwój technologii informatycznych i wzrastające zapotrzebowanie na dane przestrzenne w wielu dziedzinach gospodarki spowodowały, że raz pozyskane dane są wielokrotnie wykorzystywane w różnych systemach. Powtórne wykorzystanie danych wymaga ich adaptacji do potrzeb danej bazy danych i aplikacji GIS oraz zintegrowania z innymi danymi. Integracja danych przestrzennych polega na łączeniu danych pochodzących z różnych źródeł oraz na tworzeniu relacji pomiędzy zbiorami danych geometrycznych i opisowych. Integracji nie należy utożsamiać z transmisją i konwersją danych. Transfer danych stanowi jedynie techniczną część procesu integracji i ma wyłącznie formalny charakter oparty na jednorodnych i stałych strukturach, abstrahujących od aspektów znaczeniowych.

Integracja danych pochodzących z różnych systemów wymaga zachowania semantyki tych danych, czyli przeanalizowania powszechnie występujących konfliktów nazewnictwa (homonimy, synonimy), znaczenia i schematów. Homonimy powstają wówczas, gdy ta sama nazwa przypisana jest różnym danym (obiektom lub pojęciom), synonimy – kiedy różne nazwy opisują dane o tym samym znaczeniu. Konflikt znaczenia jest wynikiem odmiennych definicji lub interpretacji tego samego pojęcia, zaś konflikt schematów – różnic w zastosowanych schematach aplikacyjnych.

Ważnym aspektem integracji danych w systemach informacji przestrzennej jest zapewnienie zgodności w przebiegu odpowiadających sobie elementów geometrycznych i uzgodnienie styków pomiędzy danymi pochodzącymi od różnych dystrybutorów, a także zapewnienie zgodności topologicznej wewnątrz warstw i pomiędzy warstwami tematycznymi. Doprowadzenie do zgodności i poprawności topologicznej danych pozyskiwanych różnymi metodami jest zwykle procesem długotrwałym, a co za tym idzie i kosztownym. Rozważając przestrzenny aspekt integracji należy pamiętać o różnych systemach odniesień przestrzennych, odwzorowaniach i układach współrzędnych oraz związanych z nimi zniekształceniach i poprawkach odwzorowawczych. Integracja danych przestrzennych to również łączenie i zapewnienie wspólnych możliwości analitycznych danych geometrycznych zapisanych w postaci różnych modeli oraz towarzyszących im danych opisowych.

Rozważania dotyczące integracji oparte zostały na doświadczeniu zdobytym przy realizacji projektu dotyczącego wyznaczania obszarów o niekorzystnych warunkach dla gospodarki rolnej. W bazie danych, zaprojektowanej i założonej na potrzeby projektu, zgromadzone zostały dane o: podziale administracyjnym kraju, użytkowaniu ziemi, rzeźbie terenu (spadki, wysokości), glebach, obszarach chronionych oraz dane statystyczne dotyczące waloryzacji rolniczej przestrzeni produkcyjnej, demografii oraz charakterystyki gospodarstw rolnych. Dane pochodzące z różnych źródeł przechowywane są w formacie wektorowym (dane dyskretne), rastrowym (dane ciągłe), TIN (dane o rzeźbie terenu) oraz relacyjnym (dane opisowe). Wybór modelu danych podyktowany był względami pragmatycznymi związanymi z zapewnieniem optymalnych warunków zarządzania danymi i możliwości analitycznych systemu.

Zaprezentowane rozważania dotyczące integracji danych w jednolitym środowisku bazy lub hurtowni danych przestrzennych są kontynuacją dyskusji nad budową infrastruktury danych przestrzennych i zapewnienia dostępu do danych szerokiemu gronu użytkowników.

Ельжбета Белецка

Интеграция данных в однородной среде пространственных данных. Избранные проблемы

Р е з ю м е

Развитие информационных технологий и повышающийся спрос на пространственные данные в многих областях хосяйства вызывают, что уже полученные данные используются многократно в разных системах. Повторное использование данных требует их приспособления под потребности определённой базы данных и аппликации географической информационной системы, а также интеграции с другими данными. Интеграция пространственных данных заключается в содениению данных происходящих из разных источников, а также на создании отношений между множествами геометрических и описательных данных. Интеграции не следует идентифицировать с трансмиссию и конверсию данных. Передача данных это только техническая часть процесса интеграции и имеет исключительно формальный характер, базирующий на однородных и постоянных структурах, не касающихся смысловых аспектов.

Интеграция данных, происходящих из разных систем, требует сохранения семантики этих данных, т. е. проведения анализа обще присутствующих конфликтов номенклатуры (омонимы, синонимы), значения и схем. Омонимы возникают тогда, когда это самое название касается разных данных (объектов или значений), синонимы – когда разные названия описывают данные с таким же значением. Конфликт значения является результатом разных дефиниций или интерпретаций того же понятия, а конфликт схем – разниц в принятых схемах применений.

Важным аспектом интеграции данных в системах пространственной информации является обеспечение совпадения проведения соответствующих друг другу геометрических элементов и сводка стыковок данных происходящих от разных распределителей, а также обеспечение топологического совпадения внутри слоев и между тематическими слоями. Приведение к совпадению и топологической правильности данных получаемых разными методами является обычно долгим, а в результате тоже дорогим процессом. Рассматривая пространственный аспект интеграции над помнить о разных системах пространственных отнесиений, отображениях и системах координат, а также связанных с ними деформациях и коррекциях отображений. Интеграция пространственных данных это тоже соединение геометрических данных, записанных в виде разных моделей, и обеспечение совместных аналитических возможности, также сопутствующим описательным данным.

Рассуждения, касающиеся интеграции, были основаны на опыте, приобретённым в ходе реализации проекта, касающегося определения областей с ниеблагоприятными условиями для сельского хозяйства. В базе данных, запроектированной и основанной для потребностей проекта, были собраны данные о: административном делении страны, землепользовании, рельефе местности (уклоны, высоты), почвах, защищаемых районах, а также статистические данные по валоризации сельскохозяйственного производственного пространства, демографии и характеристике сельских хозяйств. Данные, происходящие из разных источников, хранятся в векторном (дискретные данные), растровом (сплошные данные), TIN (данные о рельефе местности) и реляционном (описательные данные) форматах. Выбор модели данных был определён прагматическими отношениями, связанными с обеспечением оптимальных условий управления данными и аналитическими возможностями системы.

Представленные обсуждения, касающиеся интеграции данных в однородной среде базы или склада пространственных данных являются продолжением обсуждений по созданию инфраструктуры пространственных данных и обеспечению доступа к данным широкому кругу пользователей.