Data for GIS

Vasily POPOVICH, Andrei PANKIN, Yan IVAKIN

(Prof. Dr. Vasily POPOVICH, Andrei PANKIN, Yan IVAKIN
SPIIRAS, OOGIS Laboratory, 14 Liniya VO, 39, St. Petersburg, Russia, Email: Popovich@mail.iias.spb.su; http://www.oogis.ru)

Abstract – Various purpose and scale geoinformation systems (GIS) are first and foremost interfaces to arrange for data access and user-friendly control of their transformation. As a rule GIS data specifics consists in their geographic reference (bind) i.e. reference to geographic coordinates.

The fact that GIS have quite a heterogeneous nature makes their use somewhat complex. At that various applications require an access to different data groups; many GIS applications are expected to work in real time or close to real time modes, thus, causing complications while arranging for data access and way of data representation.

Conventionally GIS data problems were solved through development of various data formats or specialized data bases; however, since recently the number of applied tasks requiring GIS capabilities keeps rapidly growing, the use of currently existing standards or development of new ones cannot contribute to the above problem solving.

The here presented research proposes to consider the complete GIS data transformation cycle to further implement these GIS as interfaces supporting decision making systems. Judging from the so far accumulated experience of GIS implementation and based on the available literature study it could be said that the class of these systems is extremely large. However, a special interest is now drawn to the GIS based decision making systems for managing the municipal aggregations, transportations (maritime, ground); controlling various purposes and scales monitoring systems.

In general case the process of data access and transformation is assumed to be considered in the light of three abstracting levels: harmonized, integrated and fused data. The above mentioned levels possess a physical basis and really reflect the information flows propagating from measuring devices up the hierarchy. The paper is only focused on main statements and does not go into details specifying each level.

1 INTRODUCTION

First of all let us cite a well known definition for data notion: Data - distinct pieces of information usually formatted in a special way. Data can exist in a variety of forms -- as numbers or text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind [6]. Metadata describes how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications [6].

This paper is not aimed at identifying the difference between data and information notions, though by many studies as well as in common practice the above notions are perceived as identical; our research will mainly discuss data.

Significant interest was drawn to information and data notions upon Internet emergence and before that upon GIS emergence. Perhaps, the GIS developers were the first ones who faced the problems of using versatile and bulky information in real-time or close to real-time scales.

The given problem becomes even more acute when GIS are used as interfaces in decision support and decision making systems for subject domains of municipal aggregations management, maritime navigation and ground transportation control, ecological monitoring and other; actually, today, it easier to list the areas where GIS are not used than to mention all of those where GIS are intensively used or are under way.

GIS diversity stipulated an emergence of great number of information resources aimed at providing GIS for necessary data. Each of above resources, as a rule, is based on a definite data model, or on a so called data representation format. However, the major problem here is that the currently existing data formats and, consequently, based on them resources, with the exception of special cases, do not cover the information needs of advanced decision support and decision making systems (DSMS). Thus, a problem of various data sources grouping at a concept level arises.

This paper proposes to segregate three data groups or three data types: harmonized, integrated and fused data. The above grouping is important to better understand the following issues in data use and transformation processes:

- data type definition (measured data, preprocessed data, extrapolated and/or interpolated data and other);
- data source definition along with data quality and their credible degree.
- a possibility to use data in certain problems' solving;
- a possibility to perform further data transformations (as a rule, isomorphic transforms of integrated data are impossible).

The above given list does not claim to be exhaustive though allows explaining an idea or objective of data conceptual decomposition.

2 INFORMATION HARMONIZATION

The given process assumes a definition of main notions and their interrelations (ontologies) based on matching subject domains and/or responsibility areas. For instance, certain division into existing knowledge domains can be made: hydro acoustic, hydro meteorology, radiolocation,, theory of search, etc. Information harmonization can solve the following main problems:

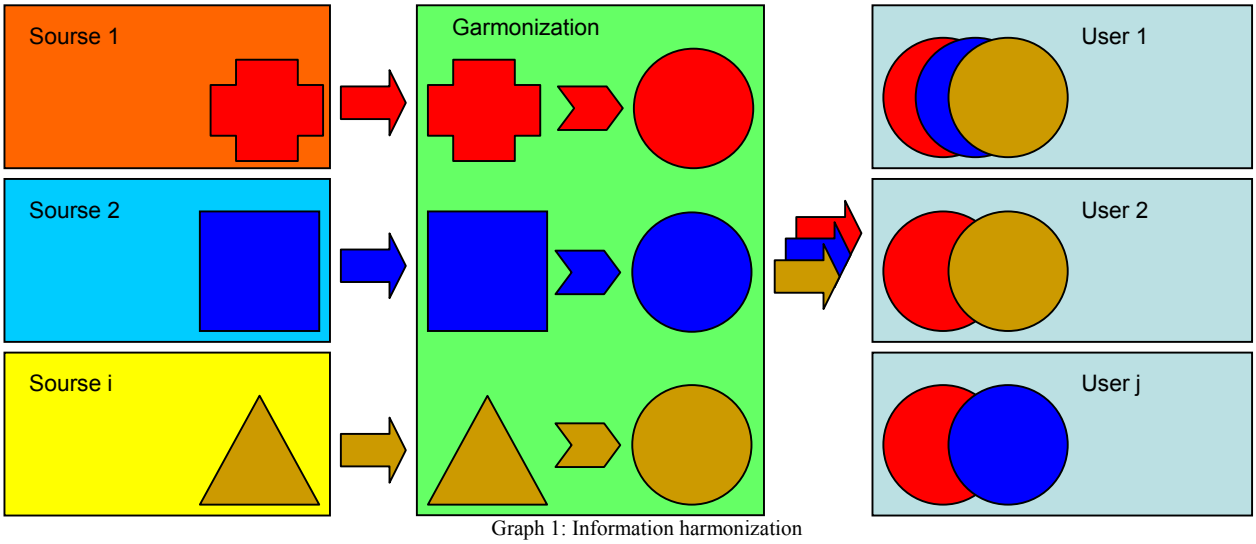
- arranging for an access to possibly great number of information primary sources;
- allowing an information transformation into user-friendly representation (decoding, recognition, translation, …);

- arranging for an access to existing information resources.
Harmonization in a broad sense can be interpreted as data standardization.

Arranging for an access to primary information sources can be done on two levels: hardware and software.
Generally, for GIS three following information sources' types can be proposed:

- not formalized information (regular text, raster graphics, photos, etc);
- formalized information (e.g., in XML format);
- formalized measurements' results (in textual and digital form);
- various formats of data bases;
- cartographic information in specialized formats;
- medium information in various specialized formats.

Graph.1 illustrates the graphic harmonization.



Graph 1: Information harmonization

As seen in Graph. 1 an access to every information source, as a rule, is arranged through different protocols, methods and/or mechanisms. As an example an access to Internet resources, data bases, GPS, GSM data, books, papers, etc can be considered. The harmonization idea consists in realization of clear principles and mechanisms of an access to information, their unification and types number contraction. Codes of the World Meteorological Organization (WMO) can serve as a good example. Currently the data main flow is being propagated though cables, transmitted as facsimile telegrams, and that significantly hampers their processing and further using, and this why WMO works on data reducing to uniform XML format to simplify operations with data of the kind.

Information harmonization process distinguishing feature is that harmonization result is oriented to a great number of users (customers).

As mentioned in [1], environmental information harmonization is more than essential in regional, national, European and global contexts.

First and foremost the above is stipulated by:

- global monitoring of the Earth surface, natural resources, and other data to be processed and realized in accordance with the Kyoto Protocol;
- European environmental policy including environment protection, cities development, protection against natural cataclysms;
- danger of adverse emissions, geophysical hazards, and technological risks;
- international cooperation, security policy pursued through development of appropriate maps and decision support and decision making systems.

Development of regional, national, European or global infrastructure of spatial data puts forward a requirement of information availability and mutual exchange. The above generates a requirement to introduce standardization and matching technologies. Status quo raises a question about information availability and exchange between different communities, thus, stimulating the efforts being put in data harmonization through a development of a common data geo model.

Development of the common data geo model will allow users to have an access to various data sources and software as the users need.

European Geoinformation Community stated a task to develop an open organization coordinating efforts in information harmonization.

By the initiative of British Geological Survey (BGS) and the Geological Survey of Canada (GSC) special meeting was arranged for in Edinburgh in November of 2003. Representatives of fifteen geological agencies from different countries and continents (Europe,

America, Asia, Australia) attended this meeting. A work group for a development of data models was organized then to work under the auspices of the Commission for the Management and Application of Geoscience Information (CGI), that is a new commission of the International Union of Geological Sciences (IUGS). The work group establishe three subgroups: "Conceptual Model/Interchange", "Testbed" and "Classification Requirements".

In 1998 in Germany was established a governmental commission IMAGI (Interministerial Committee for Geo Information) aimed at developing and implementing the German National Spatial Data Base (Geodateninfrastruktur Deuschland: GDI-DE) having its major objective to introduce the harmonization and availability of needed geodata data in response to the query put through Internet.

Information harmonization assumes solving of certain tasks, whose totality can be divided into the following groups:

Organizational tasks, supposing definition of data sources and users, data acquisition system and users' informational level;

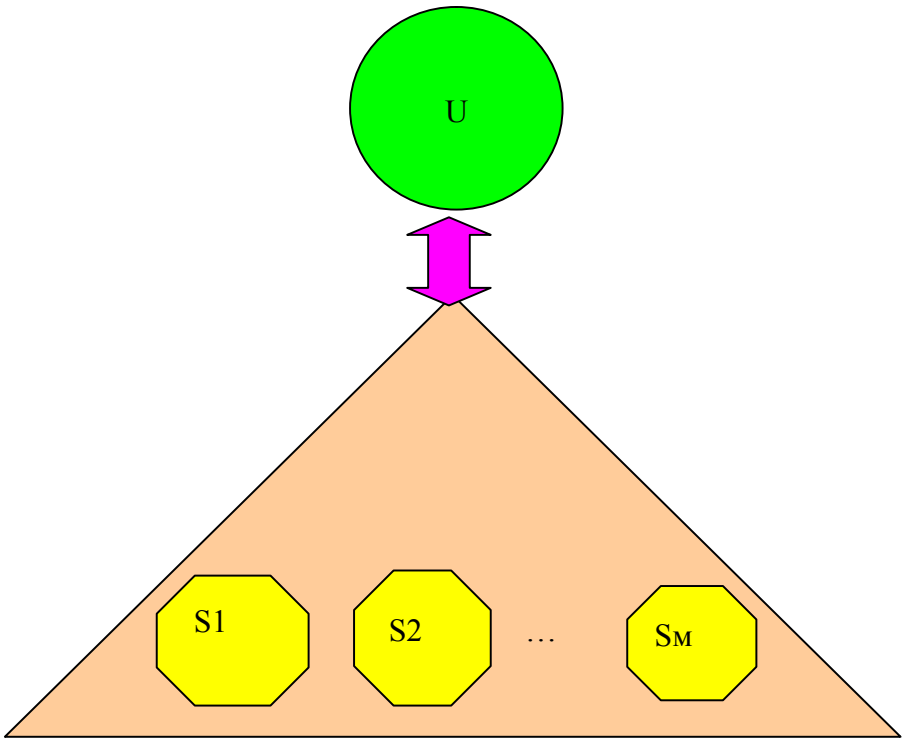
Technical tasks include protocols and standards realization by software and technical means as well as data access realization.

Legal issues. Development of license agreements, data copy rights and statuses, splitting of general information, arranging for security, copying, along with copy rights assurance.

Economic and social issues. Arranging for funding to maintain various works as well as estimating costs of information and services provided. Specifying the information market and costs, as well as expected profit and its distribution.

3 INFORMATION INTEGRATION

Information integration (access to information resources) for current tasks solving (modeling), see Graph. 1 Integration inevitably leads to data volumes enhancement, and, as a rule, is stipulated by a necessity to operate with huge data arrays in real-time or close to real-time mode. Integration is carried out to arrange for solving quite a narrow tasks' scope. Information integration for GIS can be illustrated by certain data formats like S 57, VPF and other specialized ones. Via these formats information is represented in a definite way, say, as structured data arrays. The purpose of such data arrays is solving of a certain tasks' set, e.g., data in S 57 format are aimed at arranging for navigation security within a given sea area; data in SXF format provide for topographic tasks solving on the Russian Federation territory. Currently a tendency of developing XML technologies based complex distributed data arrays can be observed. The above technology uses as a core the OWL language (Web Ontology Language). Specialized extension set GML (Geographic Markup Language) aimed at GIS data description is developed.



Graph 2: Information Integration

Data access is realized through different mechanisms, and it depends on the following factors:

- required speed of data processing (real-time or with an allowable delay);
- need of parallel processing and/or visualization of great data's number.

Depending on the above factors a direct access is provided for in the same format as the data storing format. Pretty often an intermediate data transformation is needed, as a rule, in the systems of GIS data visualization. This is stipulated by technical constraints of graphical stations and network performance and/or processors. For instance, data in VPF format can comprise information exceeding hundreds of GB depending on the scale. None of intelligent navigations can arrange for a real-time operation with such a data array. Nevertheless, given format integrated data are much more applicable to be further processed than just harmonized ones, located in different sources with different speed and access discipline.

Information integration distinguishing feature is that its result is aimed at solving tasks of a definite class.

Information integration assumes a definition of a certain data model. An example of information integration is a theoretic model of data in S 57 format. This model would be rather complicated if the further details of data vector representation were accounted for. The complexity is caused by the factor that navigation purposes require interrelated data of different kind, like isobaths, navigation signs, navigation channels, etc. At that certain data changing or updating may affect other interrelated data.

Along with the theoretic model a format that assumes an incorporation of one or several interrelated files was developed for data storing at computer medium.

It should be noted that data integration does not mean just a physical information amalgamation in one place, say, at a local user. The approach depends on the stated tasks and on the existing conditions. For instance, with regard to any format, providing for navigation security, a definite water-craft should have complete information about its current and planned navigation areas. Thus, the arranging for a timely correction of already available information is an absolutely different task. While, for example, solving the research tasks the information can be physically distributed within a certain network either local or global.

Information integration assumes number of tasks solving:

Organizational tasks. Definition of data sources and users, data acquisition system, their interrelation and updating system.

Technical tasks. Data formats realization and extension at transportation level as well as at interface level. Development and implementation of a system for data, in a given format, production, distribution, protection, and correction.

Legal issues. Development of license agreements, data copy rights and statuses, splitting of general information, arranging for security, copying, along with copy rights assurance.

Economic and social issues. Arranging for funding to maintain various works as well as estimating costs of information and services provided. Specifying the information market and costs, as well as expected profit and its distribution.

4 INFORMATION FUSION

Receiving a new information quality (reducing information volume) is the most complicated stage in data transformations. The given notion is associated with a known research area, whose history numbers several dozens of years. The development of data fusion model (DF) made a qualitative leap in this area. The above model is known as Joint Directors of Laboratories (JDL) Data Fusion Model [3]. It is noted in [4] that "fusion" can be considered in different contexts:

- software (Cold Fusion, e Business);
- physics (medicine, nuclear fusion);
- combining (integration-combination of various elements into a certain formation, integration-composition of components into a certain whole;
- knowledge (data fusion, susceptors' data fusion and information fusion).

"Data" and "information" notations are separated in [4]. Data fusion is an organized combining in the interests of analysis and decision-making, while information fusion is data combining aimed at receiving knowledge. DF is defined in [2] as a process of data from different sources composition. DF objective is specified as receiving information of higher quality. At that high quality notion depend on the application area. It might be noted, that majority of DF research perceives the information (data) quality improvement as DF main objective.

In most of advanced GIS applications the data high quality problem turned into a sequence of correctly formulated and stated tasks, having various solution versions and providing data high quality for definitely stated tasks.

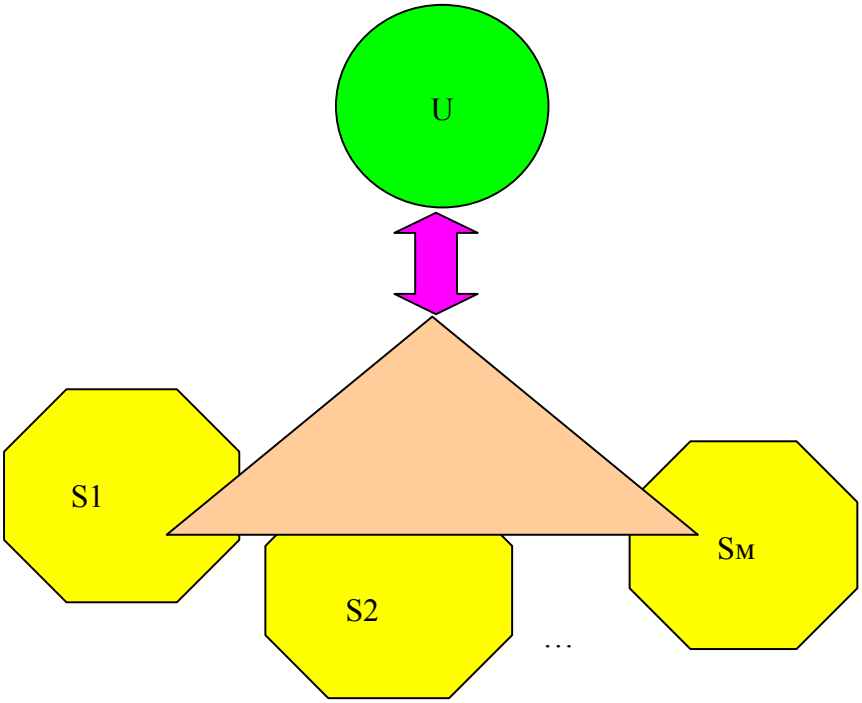
Currently the problem as a rule consists rather in data quality change than in data high quality.

The data quality change requires deep analytic study of the subject areas, so in a given context data (information) fusion will assume GIS technologies. Fusion (synthesis) main point is depicted in Graph.3. A diagram of information fusion for various purposes monitoring systems is given in Graph.4 as an example. The above diagram shows the information quality change upward the hierarchy. The idea of data fusion becomes obvious if to account for monitoring systems' complexity and their spatial distribution. The systems of the kind will hardly operate without the given mechanism.

Information fusion process distinguishing feature is a receiving of new information quality along with information volute reducing.

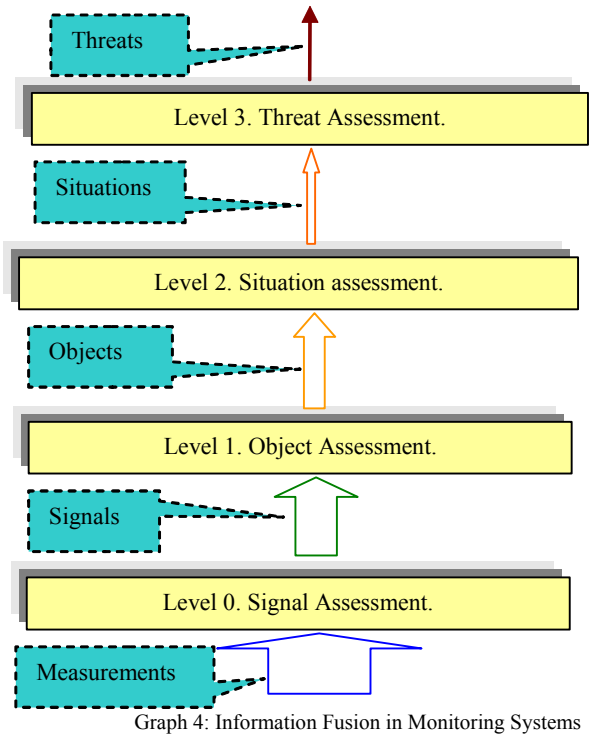
The levels given in Graph.3 demonstrate quality leaps in information representation. The mentioned diagram can be practically applicable to any monitoring system, and possibly not only to this class of systems. The considered case illustrates the Hegel's principle of quantity into quality transformation. However, there exists a nuance that is an absence of universal mechanisms able to perform such qualitative transformations. This rather is a system of special research incorporating a whole series of scientific leads.

Let us make further comments to Graph.3. At the zero level a signal is enhanced by a system or a physical fields measuring device against a background noises and interferences.



Graph 3: Information Fusion

At the first level a decision is made about definite class signal detection. The given levels differ in primary information as well as in mathematic methods. At the zero level a classic theory of signal detection is dealt with, and at the first level methods of classification and recognition are engaged. In some cases one more subdivision is added to the first level that is a track analysis having sufficient independence both in research methods and in application areas.

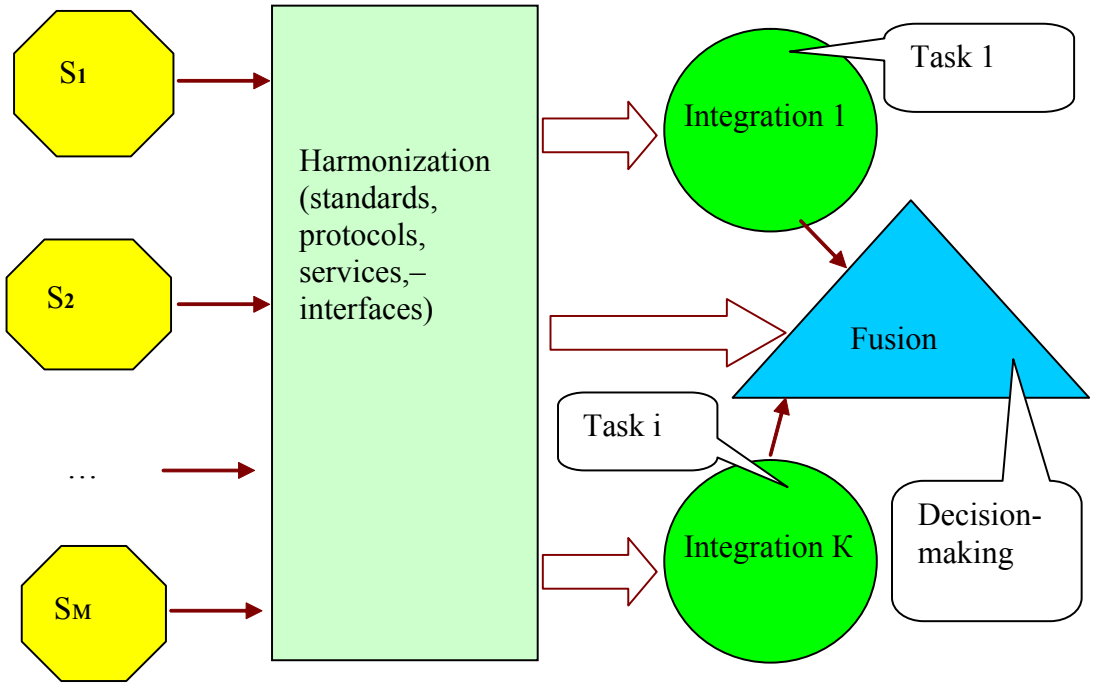


Graph 4: Information Fusion in Monitoring Systems

At the second level the possible situations generated by definite objects actions or inactions. This is one more qualitative leap in a sense of a subject being studied and in a sense of research methods being implemented.

As a rule, analysis of possible situations is not end in itself, and the following stage is a evaluation of potential threats, that might be derived by the current situation.

The above specified information processing levels are interrelated and interdependent as shown in Graph.5



Graph 5: Interrelation between information processing levels

Information fusion supposes some sequence of actions:

Organizational tasks: specifying data sources and users, data acquisition systems, their interrelation and updating system.

Technical tasks. Data formats realization and extension at transportation level as well as at interface level. Development and implementation of a system for data, in a given format, production, distribution, protection, and correction.

Legal issues. Development of license agreements, data copy rights and statuses, splitting of general information, arranging for security, copying, along with copy rights assurance.

Economic and social issues. Arranging for funding to maintain various works as well as estimating costs of information and services provided. Specifying the information market and costs, as well as expected profit and its distribution.

5 CONCLUSIONS

The information processing issues in GIS and GIS applications considered by this paper reflect that currently the problem goes far beyond conventional GIS research as well as their practical applications. Any endeavor of GIS introduction into real systems gives a rise to the above considered tasks of data harmonization, integration and fusion.

Here emerges a urgency of theoretical and practical investigations of different information processing levels for GIS and GIS applications. Some problems of the kind are being now solved directly or indirectly; technologies like web-services and service-oriented architectures concept are directly applicable to realization of information harmonization. Various extensions of laying-out languages, like GML, can and have to be used to solve data integration and harmonization tasks.

Historically data integration problem used to be solved through development of specialized data formats or data sets; the best known ones are S57, VPF, SXF, and various Shape formats. However, GIS implementation experience shows that there always arises a need to use data in at least two or more formats (i.e. global sources).

Data fusion in GIS is a field requiring the most complicated research and technological solutions. The peculiarity of this level is the orientation much more focused to a definite user than in case of integration; it is also complicated by a need of using rather sophisticated mathematical approaches and models.

6 REFERENCES

1. Kristine Asch, Boyan Brodaric, and etc. An International Initiative for Data Harmonization in Geology. //10th EC-GI&GIS Workshop, Abstracts,
2. Warsaw, 23-25 June,2004, p.9.
3. L. Valet, G. Mauris, and Ph Bolon, “A statistical overview of Recent Literature in Information Fusion,” Fusion 2000. IEEE AES March 2001.
4. White, F.E., “A Model for Data Fusion”, Proc. 1st National Symposium on Sensor Fusion, 1988
5. E., Blasch, “Fundamentals of Information Fusion and Applications”, Tutorial, TD2, Fusion 2002.
6. James Llinas, et al. “Revising the GDL Data Fusion Model II” 2005.
7. www.webopedia.com