View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Bulgarian Digital Mathematics Library at IMI-BAS



# DATA INDEPENDENCE IN THE MULTI-DIMENSIONAL NUMBERED INFORMATION SPACES

Krassimir Markov

*Abstract: The concept of data independence designates the techniques that allow data to be changed without affecting the applications that process it. The different structures of the information bases require corresponded tools for supporting data independence. A kind of information bases (the Multi-dimensional Numbered Information Spaces) are pointed in the paper. The data independence in such information bases is discussed.*

*Keywords: Data independence; Multi-dimensional Numbered Information Spaces*

*ACM Classification Keywords: D.2 Software Engineering; H.2 Database Management*

## Introduction

It is well-known, the concept of the data independence denotes that a database is designed and maintained independently of applications that retrieve and manipulate the data [CODASYL, 1971], [Martin, 1975], [Date, 1977], [Gabrovsky and Markov, 1977], [Connolly and Begg, 2005]. Dr. E.F. Codd had published a list of rules that concisely define an ideal relational database, which have provided a guideline for the design of all relational database systems ever since. The rules from 8 till 11 concern the data independence [Codd, 1985]. There are several kinds of data independence.

*Physical data independence* means that physical details of data organization and access are transparent for the application programmer. The physical details are determined by the database designers or even earlier, by the designers of a database management system. Some physical details (e.g. indices supporting the access to data, special file organization, special methods of performing operations, etc.) are under control of a database administrator (DBA). DBA uses special administration module to tune the database operation according to the actual demands of applications, but still, nothing in applications has to be changed due to the tuning [Subieta, 2005]. Physical data independence is the rule 8 of Codd's list - the user is isolated from the physical method of storing and retrieving information from the database. Application programs and terminal activities remain logically unimpaired whenever any changes are made in either storage representations or access methods.

*Logical data independence* means that DBA is able to perform some operations on the database structure, for instance, add new data kinds, add or remove some object or table attributes, change user privileges, add and remove views, database procedures, triggers etc. without unconscious influencing the applications [Subieta, 2005]. Logical data independence is the rule 9 of Codd's list - how a user views data should not change when the logical structure (tables structure) of the database changes. Application programs and terminal activities remain logically unimpaired when information-preserving changes of any kind that theoretically permit unimpairment are made to the base tables. (This rule permits logical database design to be changed dynamically, e.g. by splitting or joining base tables in ways which do not entail loss of information.) Continuing this principle we may talk about *Conceptual data independence* that means that DBA is able to change the structure of the database conceptually without changing existing (legacy) applications, for instance, through special wrappers, mediators, views, updatable views, do instead of triggers, and other means that allow to change significantly the database schema and its organization, perhaps with minor changes of applications. This kind of data independence is referred to as schema evolution and conceptually is close to software change management methods and the aspect-orientation in databases. [Subieta, 2005]

*Integrity Independence* is the rule 10 of Codd's list - the database language (like SQL) should support constraints on user input that maintain database integrity. Integrity constraints must be definable in the relational data sub-language and storable in the catalogue, not in the applications program. Certain integrity constraints hold for every relational database, further application-specific rules may be added.

*Distribution Independence* is the rule 11 of Codd's list – the user should be totally unaware of whether or not the database is distributed (whether parts of the database exist in multiple locations). A relational DBMS has

distributional independence - i.e. if a distributed database is used it must be possible to execute all relational operations upon it without knowing or being constrained by the physical locations of data. This must apply both when distribution is originally introduced, and when data is redistributed [Codd, 1985].

It is clear; the importance of data independence in a database management system is well recognized in the database community. To ensure the data independence is the reason for developing the three levels database architecture. Its goal is to separate the user applications and the physical database. Basically, “Three-schema Architecture” has an “Internal” level, a “Conceptual” and an “External” level. The advantages of the three tiered architecture are that this division into levels allows both developers and users to work on their own levels. They do not need to know the details of the other levels AND they do not have to know anything about changes in the other levels.

*The Project INSPIRE*

A particularly interesting example of multi level implementation in practice is the European Union’s INSPIRE (Infrastructure for SPatial Information in Europe) initiative [INSPIRE, 2007]. This was launched in 2001 with the objective of making available relevant, harmonised and quality geographic information to support the formulation, implementation, monitoring and evaluation of Community policies with a territorial dimension or impact' ( http://inspire.jrc.it ). INSPIRE is seen as the first step towards a broad multi sectoral initiative which focuses on the spatial information that is required for environmental policies. It is a legal initiative that addresses “technical standards and protocols, organisation and coordination issues, data policy issues including data access and the creation and maintenance of spatial information” [Masser, 2005].

The Figure 1. provides a simplified overview of key elements in the technical architecture of INSPIRE.

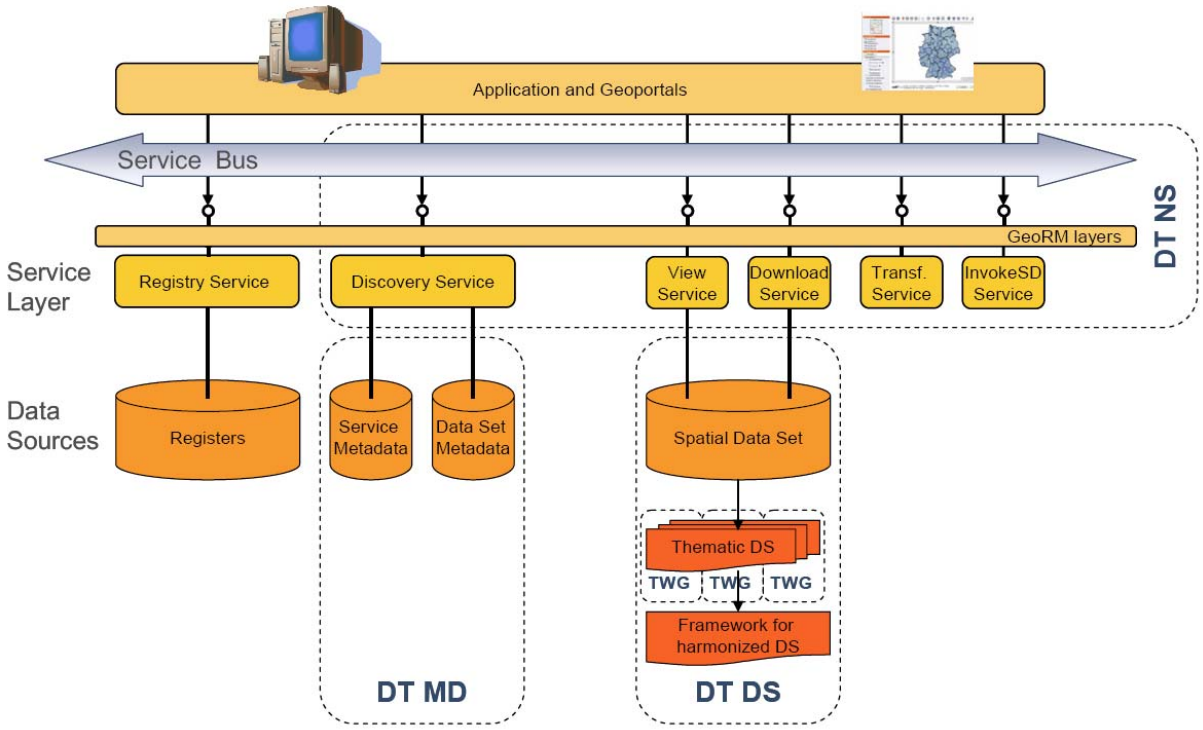


Figure 1: INSPIRE technical architecture overview presented in [INSPIRE TR, 2007].

The core resource in the diagram is the actual content, i.e. the spatial data in spatial data sets. “Spatial data” denotes any data with a direct or indirect reference to a specific location or geographic area The use of the word “spatial” in INSPIRE is unfortunate as the meaning goes beyond the meaning of “geographic” – which is understood to be the intended scope. Therefore, “spatial data” is understood as a synonym for the term

"geographic information" as used in the ISO 19100 series of International Standards. "Spatial dataset" denotes identifiable collection of spatial data.

All other resources shown in the diagram, e.g. data set metadata, are only needed to find, access, interpret or use the spatial objects in the spatial data sets that form part of the infrastructure. "Spatial object" denotes an abstract representation of a real-world phenomenon related to a specific location or geographical area. Note: This term is understood as a synonym for geographic feature as used in the ISO 19100 series of International Standards [INSPIRE, 2007].

It is important to note that in INSPIRE all access to spatial data and metadata occurs via spatial data services. The implementation platform for these services is expected to be web services. All services are described by service metadata (service descriptions), allowing humans and software applications to discover specific service instances in the infrastructure [INSPIRE TR, 2007].

The INSPIRE infrastructure will be built upon existing or emerging infrastructures in the member states and international organisations. In particular it has to be emphasised that changes to existing data capturing, updating and management processes within the member states and international organisations are in general not foreseen by the implementation of INSPIRE. Instead, the INSPIRE Implementing Rules aim at providing access to existing spatial data in a harmonised way.

In INSPIRE there are 34 different spatial data themes. For some of them already there exist specialized information bases - cadastral parcels, geographical grid systems, meteorological geographical features, etc. Only for a part of them are developed web access and good end user support. A large group of themes are under investigation and under current or near future development of the appropriate information bases. For instance, such themes are:

- Administrative units - Units of administration, dividing areas where Member States have and/or exercise jurisdictional rights, for local, regional and national governance, separated by administrative boundaries.
- Elevation - Digital elevation models for land, ice and ocean surface. Includes terrestrial elevation, bathymetry and shoreline.
- Human health and safety - Geographical distribution of dominance of pathologies (allergies, cancers, respiratory diseases, etc.), information indicating the effect on health (biomarkers, decline of fertility, epidemics) or well-being of humans (fatigue, stress, etc.) linked directly (air pollution, chemicals, depletion of the ozone layer, noise, etc.) or indirectly (food, genetically modified organisms, etc.) to the quality of the environment.
- Utility and governmental services - Includes utility facilities such as sewage, waste management, energy supply and water supply, administrative and social governmental services such as public administrations, civil protection sites, schools and hospitals.
- Environmental monitoring facilities - Location and operation of environmental monitoring facilities includes observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc.) by or on behalf of public authorities.
- Natural risk zones - Vulnerable areas characterised according to natural hazards (all atmospheric, hydrologic, seismic, volcanic and wildfire phenomena that, because of their location, severity, and frequency, have the potential to seriously affect society), e.g. floods, landslides and subsidence, avalanches, forest fires, earthquakes, volcanic eruptions.
- Atmospheric conditions - Physical conditions in the atmosphere. Includes spatial data based on measurements, on models or on a combination thereof and includes measurement locations.
- Habitats and biotopes - Geographical areas characterised by specific ecological conditions, processes, structure, and (life support) functions that physically support the organisms that live there. Includes terrestrial and aquatic areas distinguished by geographical, abiotic and biotic features, whether entirely natural or semi-natural.

  etc.

These themes are characterized by the need of collecting, storing, processing and distributing the space information with more than usual three geographical dimensions. In some cases the number of dimensions

exceeds one hundred. In addition, the collected information is not homogeneous and there exist the problem of mapping the different parts of information and easy connecting them to geographical coordinates without of permanently duplicated triples of values. For this type of practical necessity it is important to propose new tools with corresponding possibilities.

One of the main problems to be solved is representing the corresponded digitalized information in the appropriate data bases which are aimed to support time depended, multidimensional, multimodal and multimedia, individualized and confidential access to searched digitalized information objects. In the same time, the digitalized information objects need appropriate tools for storing, retrieving, processing and multimodal time depended representing for a great number of types of users. This problem could not be solved using popular in the practice (as a rule – relational) data base management systems (DBMS). The current multidimensional time depended extensions, such as these in the newest versions of the Oracle DBMS, are aimed to operate with not so complex and complicated information objects. The main area of implementing of such systems is the business information service. The pointed themes of INSPIRE pose the question about developing of principally different approach for building the information bases.

One such approach, which has been developed and experimented more than for twenty years, is using the multi-dimensional numbered information spaces [Markov, 2004]. The main its advantage is the possibility to build space hierarchies of information objects and the great power for building interconnections between information elements of the stored in the information base objects. Practically unlimited number of dimensions and the opportunity of representing and storing the information only about the existing parts of the real objects make possible the creating effective and useful tools for working with information. This approach allows possibility for building the very large information bases and supporting the time depended, multidimensional, multimodal and multimedia, individualized and confidential access to searched digitalized information objects.

## Data Independence in the Multi-dimensional Numbered Information Spaces

The main peculiarity of the multi-dimensional information bases is the need of multi-dimensional schema at every level and very important feature is supporting the corresponding data independence, i.e. the capacity to change the schema at one level of the information base without having to change the schema at the next higher level. In other words, if the schema at one level is changed, the mapping to the next higher level needs to be changed to ensure the schema at the next level to remain unchanged.

*The Internal Level* is a description of the physical storage structure of the information base. The operations performed here are translated into modifications of the contents and structure of the files (archives).

The only tool which allows physical organization of multi-dimensional numbered information spaces is the FOI Archive Manager (ArM)®. ArM is based on the "Multi-Domain Information Model" (MDIM) [Markov, 2004]. One of the first goals of the developing of ArM was representing the digitalized military defense situation which is characterized with variety and complexity of objects and events which occur in the space and time and have long period of variable existence. The great number of layers, aspects and interconnections of the real situation may be represented only by multi-dimensional information spaces.

The ArM main information structures are:

- Basic information elements - arbitrary long strings of machine codes (bytes) which may represent information structures of any kind. When it is necessary the strings may be parceled out by lines. The length of the lines may be variable.
- Numbered information spaces of different ranges - the basic elements are organized in hierarchy of the numbered information spaces with variable ranges.

Every element as well as every space has unique number in the space it belongs. This way the element may be accessed by correspond "space address" (coordinates) given via coordinate array of numbers of the spaces that contain it in the hierarchy. The quantity of levels of the hierarchy is unlimited.

So, we have only two constructs for the physical organization of the information base – basic information elements and numbered information spaces, which may be accessed exclusively via coordinates and the Archive Manager provides 100% physical data independence.

*The Logical level* is a description of the structure of the entire information base. It hides the details of physical storage and concentrates on describing entities, data types, relationships, user operations, and constraints. The logical organization of multi-dimensional numbered information spaces is too complicated and it is very important to have special multi-dimensional logical schema and corresponding technology for operating with it. Such special technology called Cell Oriented Programming (CellPro) was presented in [Markov et al, 1995]. The mathematical foundations of CellPro may be find in [Lisper, 1989]. The meta-information for describing the logical organization and for its mapping over the physical one is represented by the Cells and Cell Structures. It is stored in special mapping ArM-archives, i.e. again in multi-dimensional numbered information spaces. The FOI System Builder (SyB) is a tool for multiagent cell oriented programming. The SyB main features include defining the cell structures and using the different types of cells; sets of cells; cell scripts as the specific cells' configurations with fixed or variable structure and activities; sets of cell scripts etc. SyB is constructed by System Building Service Module (SyB_SM) and Real-time Management System (SyB MS).

The SyB_SM is a system for service the defining the cells, sets of cells and cell scripts. The SyB_SM can translate the descriptions of the scripts in the internal format for the SyB_MS interpretation. The SyB_MS is a system for real-time concurrent control using the cell scripts. It may be linked with a user program and may be called using a special procedure type of interconnection.

The Cells are capsulated items that have their own description and activity; i.e. the cells are agents. The description of a cell may contain: name of the cell; type of the cell; definition of the cell activity; interconnection between the cell and other cells, etc. The cell may be:

- data cell which may contain: single data (strings of any size; integer or real numbers; dates; graphic images, search patterns, scripts); sets of single data, or sets of more complex structures which may contain both sets and single data
- functional cell which may contain built-in standard functions, constructively integrated user defined functions, external standard or user defined programs.

The activity of the cell depends of its type. For instance it may be a simple operation with the standard data (such as creating, deleting, editing and copying) or more complex information processing. The cell may execute the build in functions or scripts as well as the external standard programs such as search processing or hierarchical hypertext service. It may perform compress and decompress of information subspaces, transfer the information between points of physical information space, etc. At the end the activity of the cell may be executing the user defined scripts or programs of any kind, which may be written using different programming languages.

There are two general categories of cells: user accessible and user not accessible. The first one may be used for building the user interface with given concrete application and the second - for organizing the information base and internal information processing.

In addition to standard characteristics, the user visible and/or accessible cells have theirs screen representations and relations with representations of the other cells of such type.

The cells may have two possible types of connection with the data – absolute or relative. The absolute connection is given directly by set of coordinates, organized in the coordinate array. The relative connection is given by a combination of base, offset and index coordinate arrays.

*The External Level* is formed by the end users' special views that are tailored to their specific needs. Some of these views may be forms to fill out, others for interactive retrieval of information, etc.

Each user group refers only to its own external schema so the DBMS must transform a request on an external schema into a request against the conceptual schema, then into a request on the internal schema for processing. The process of transforming requests and results between levels are called mappings. These mappings may be time consuming, so some many DBMS do not separate the three levels completely.

In multi-dimensional information bases the mappings could not be simple functions. In addition the dynamic of user activities needs special tools for mapping the user views to the schema. In this work such mappings are realized using the SyB's scripts [Markov et al, 1999].

The SyB's scripts are specific cells' configurations with fixed or variable structure. It is possible to describe different configurations (not only the spreadsheets type) of the cells.

The fixed script's structure may be created and used as it was defined without any change during the usage. The variable one may be changed during the usage in accordance with any conditions. The variable structure may be connected to a special index of co-ordinates for building the final configuration of the cells in a given script.

There are several cell types used only for fixed cell scripts: the text string <S>, the integer <I> and real <R> number, the script <T>, the user program <P>, the standard function <F>, the read only cell <L>, etc.

The cell types used both for fixed and variable cell scripts are string text cell <Z>, number cell <X> or <Y>, date cell <D>, and functional cell <W>, etc.

The difference between cell's types is in the possibility for iterative access to the information space, which is available for variable cell scripts. This means that the cell may take part in the script many times and every time it cans access different zone of the archives. As usual special kind of indexes is tool for such adaptive processing.

It is clear, the <S> and <Z> cells may contain only text strings and the cells <I>, <X>, <R> and <Y> - only numbers. The cell <D> services using the dates. The activation of <T> cell will start execution of the subscript, which the cell can contain. True the <P> cell we may start execution of the user-defined program, which is pointed in the cell as well as <F>, and <W> cells will start execution of the user-defined function, which is connected, to the cell. The possibility of the <L> cell is only to visualize its content.

The description of the script may be done using the SyB language. It contains three main parts: interconnection between cells, individual functional type of the cell and activities of the cell.

The interconnection between the cells may be given by the exactly, default or functionally described cell names or coordinates i.e. addresses. Any function for description of co-ordinates may be the built-in or the user given. The built-in function of the SyB is iterative and may be used for description of repetitively use of the cells of the CIS. For instance when one wants to describe a table where every row contains the same set of cells but with other values and activities he may use this type of the description. The user given functions may access the whole information space.

*Integrity Independence* in the multi-dimensional information bases needs to be cleared. In the relational DB there are primary and secondary keys but when the dimensions are more than two it is not clear what will play the role of keys and what should support constraints on user input that maintain database integrity. Integrity constraints must be definable in any sub-language and storable in the catalogue. In the frame of Cell oriented approach, the integrity is supported by implementing the constraints in the cells, in the structures of cells as well as in the cell scripts.

*Distribution Independence* plays a very useful role in the development of Geospatial Web Service architectures [Dadi and Di, 2007]. It is important to note that for INSPIRE it is assumed that all kind of data and metadata access and processing are performed using web services. All services are described by service descriptions (service metadata, as part of the INSPIRE metadata), allowing humans and software applications to discover specific service instances in the infrastructure and invoke them automatically [INSPIRE TR, 2007].

The INSPIRE Network Services can be seen as the protocol being used to realise a pan-European geo spatial service bus (see fig. 2):

- Different *SDI providers* who contribute INSPIRE-conforming services (access only)
- INSPIRE Network Services expose services for machine-to-machine communication. At least a workflow that follows the “publish – find – bind” design pattern should be possible. However, users do not necessarily have to follow this pattern; they can also invoke services directly.
- INSPIRE Applications solving specific tasks by involving INSPIRE services.
- INSPIRE geo-portal at Community level and further Member States access points offering INSPIRE functions to the different user groups (usage of INSPIRE services). A user can access services on an EU level via the INSPIRE geo-portal but also on a MS level – usage on the EU level offers the advantage to access data that integrates seamlessly data from different member states.

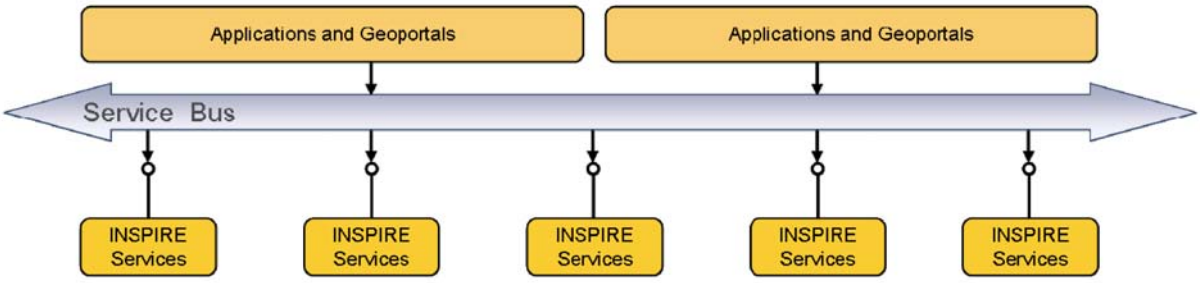


Figure 2. INSPIRE Network Service bus presented in [INSPIRE TR, 2007].

Following this understanding we may point that in our case two types of distribution independence we may have – at internal (physical) level and at the external (user) level. In the first case, the information base is assumed to be distributed and all its part are accessible automatically via INSPIRE Service bus. In the second case, the scripts may contain information request to different part of information base distributed at different servers and the integration of information will be provided by the used INSPIRE portal.

## Conclusion

It is clear; the importance of data independence is well recognized in the database community. The data independence is the main direction for investigation especially in the multi-dimensional spatial information bases.

The main conclusion is that we are on the threshold of the new style of organization of information which needs special attention. The generalization and implementing of the techniques for data independence in a special kind of information bases (the Multi-dimensional Numbered Information Spaces) ware discussed in the paper.

To ensure the data independence is the reason for developing the three levels database architecture. The advantages of the three tiered architecture are that this division into levels allows both developers and users to work on their own levels. The main levels of data independence were remembered. The investigation has been based on the ideology of the European Union's INSPIRE (Infrastructure for SPatial Information in Europe) initiative [INSPIRE, 2007]. The main 34 themes of INSPIRE were discussed and it was drawn attention to a large group of themes which are under investigation and under current or near future development of the appropriate information bases. The pointed themes of INSPIRE pose the question about developing of principally different approach for building the information bases.

One such approach is using the multi-dimensional numbered information spaces [Markov, 2004]. The main its advantage is the possibility to build space hierarchies of information objects and the great power for building interconnections between information elements of the stored in the information base objects. Practically unlimited number of dimensions and the opportunity of representing and storing the information only about the existing parts of the real objects make possible the creating effective and useful tools for working with information. This approach allows possibility for building the very large information bases and supporting the time depended, multidimensional, multimodal and multimedia, individualized and confidential access to searched digitalized information objects.

The main levels of data independence in the multi-dimensional numbered information spaces were described. The main peculiarity of the multi-dimensional information bases is the need of multi-dimensional schema at every level and very important feature is supporting the corresponding data independence, i.e. the capacity to change the schema at one level of the information base without having to change the schema at the next higher level.

The logical organization of multi-dimensional numbered information spaces is too complicated and it is very important to have special multi-dimensional logical schema and corresponding technology for operating with it. Such special technology is Cell Oriented Programming (CellPro). The meta-information for describing the logical organization and for its mapping over the physical one is represented by the Cells and Cell Structures.

The external level of data independence is based on cell scripts. In multi-dimensional information bases the mappings could not be simple functions. In addition the dynamic of user activities needs special tools for mapping the user views to the schema.

Integrity Independence in the multi-dimensional information bases needs to be cleared. In the frame of Cell oriented approach, the integrity is supported by implementing the constraints in the cells, in the structures of cells as well as in the cell scripts.

Distribution Independence plays a very useful role in the development of Geospatial Web Service architectures. From point of view of the project INSPIRE were commented the web based possibilities for data independence.

## Acknowledgments

This work is a part of the project "ITHEA XXI", partially financed by the Consortium FOI Bulgaria. Author is indebted to Krassimira Ivanova and Ilia Mitov for the collaboration in this project.

All registered and trade marks in the paper are property of their owners.

## Bibliography

[CODASYL, 1971] Codasyl Systems Committee. Feature Analysis of Generalized Data Base Management Systems. Technical Report, May, 1971 / Информационные системы общего предназначения (Аналитический обзор систем управления базами данных). Москва, Статистика, 1975.

[Codd, 1985] E.F. Codd. "Is Your DBMS Really Relational?" and "Does Your DBMS Run By the Rules?". ComputerWorld, 14 and 21. October 1985

[Connolly and Begg, 2005] T. Connolly, C. Begg, Database Systems: A Practical Approach to Design, Implementation, and Management, 4th ed., Pearson Education Ltd., 2005

[Dadi and Di, 2007] U.Dadi, L.Di. Data Independence and Geospatial WEB Services. Geoinformatics 2007 Conference (17–18 May 2007). http://gsa.confex.com/gsa/2007GE/finalprogram/abstract_122248.htm

[Date, 1977] C.J. Date. An Introduction to Database Systems. Addison-Wesley Inc. 1975. / К.Дейт. Введение в системы баз данных. Москва, Наука, 1980.

[Gabrovsky and Markov, 1977] I.Gabrovsky, Kr.Markov. About Constructing Data Independent Programs using the Package BOMP. Proceedings of the Conference of professionally connected users, Varna, 1977. pp.240-245. (in russian)

[INSPIRE TR, 2007] Infrastructure for Spatial Information in Europe Reference: INSPIRE Technical Architecture Overview, 05-11-2007 Page 5 of 12

[INSPIRE, 2007] DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007, establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union 25.4.2007 L 108/1 http://www.epsiplus.net/content/download/3477/38314/file/l_10820070425en00010014.pdf

[Lisper, 1989] B.Lisper. Syntesizing Synchronous Systems by Static Scheduling in Space-Time. Lecture Notes in Computer Science, No.: 362. Springer-Verlag, Berlin, 1989.

[Markov 2004] K. Markov. Multi-Domain Information Model. Int. Journal "Information Theories and Applications", 2004, Vol. 11, No. 4, pp. 303-308

[Markov et al, 1995] K.Markov, K.Ivanova, I.Mitov. Cell Oriented Programming. International Journal "Information Theories & Applications" (IJ ITA), 1995, Vol. 3, No. 1.

[Markov et al, 1999] K.Markov, K.Ivanova, I.Mitov. *Multiagent Information Service Based on Scripts*. Научно-теоретический журнал „Искусственный интелект", №2, 1999, ISSN 1561-5359, ИПИИ НАНУ, стр.129-135.

[Martin, 1975] J.Martin. Computer Data-Base Organization. Prentice-Hall, Inc., Englewood Cliffs, New Jersey / Дж. Мартин. Организация баз данных в вычислительных системах. Москва, Мир, 1978.

[Masser, 2005] Ian Masser. THE FUTURE OF SPATIAL DATA INFRASTRUCTURES. ISPRS Workshop on Service and Application of Spatial Data Infrastructure, XXXVI (4/W6), Oct.14-16, 2005 Hangzhou, China http://www.commission4.isprs.org/workshop_hangzhou/papers/7-16%20Ian%20Masser-A001.pdf

[Subieta, 2005] Kazimierz Subieta, Principles of modern database query and programming languages. (December 2005) http://www.sbql.pl/Topics/Principles%20of%20query%20programming%20lang.html

## Authors' Information

*Krassimir Markov – Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: markov@foibg.com*