Formation of canonical structure of territory database regarding requirements for reliability of obtaining spatial information

G I Moskvitin¹, N V Khvatysh², A B Pismennaya¹, T A Sokolova²

¹ Russian University of Transport, 9, bldg. 9, Obraztsova st., Moscow, 127994, Russia
² State University of Land Use Planning, 15, Kazakova street, Moscow, 105064, Russia

E-mail: KhvatishNV@guz.ru

Abstract. The article discusses the problem of forming the canonical structure of the territory database regarding the requirements for the reliability of obtaining spatial information. The canonical database structure is a minimal conceptual scheme. It can be obtained using a step-by-step procedure for combining users’ views of the data. A possible application of a formalized model and methods of pre-project analysis of structure of information flows and data processing technology is proposed. The methods application enables to give a formal description of the input, output and all intermediate arrays both in the analysis of the existing and the designed system; determine the procedures for transforming arrays in the process of forming the required output; determine the structure and select the technical means of the system; assess the quality characteristics of the system and choose the best option. In addition, an approach to the analysis of the structure of the projected data processing system is proposed. The approach enables to determine the necessary sequence of obtaining data items, which is simplified if the elements of the constructed reachability matrix are ordered by the levels (stages) of their processing. For this purpose, a reordered reachability matrix, a structured adjacency matrix and the corresponding graph of information relationships are constructed on the basis of the adjacency and reachability matrices. The graph of information relationships is used to refine input, intermediate and output data items, feedback loops, levels and sequence of data processing, which will allow creating a canonical structure of the territory database. In the future, it will be possible to obtain spatial information with high reliability on the basis of the proposed formalized model and methods of pre-project analysis of the structure of information flows and the technology of processing model data, and taking into account the peculiarities of the blockchain functioning.

1. Introduction
There are federal, regional and special purposes geographic information systems (GIS). The special purposes geographic information systems are understood as systems used to serve the information needs of specific sectors of the national economy. GIS, as systems, are designed, created and operated in a complex of their constituent components (blocks, subsystems). Integration is the most important feature of the interaction of geoinformatics with its environment. One of its consequences is the emergence and development of borderline disciplines. Integration processes affect not only the
relationship of the classical triad “remote sensing – geoinformatics – cartography” or their pairwise relationships. Modern practice provides many examples of integrated solutions based on the existence of a single technological digital environment. The information basis of GIS is formed by digital representations (models) of reality. With the advent of the computer, the entire set of data was divided into two types: digital and analog data [1].

Recently, much attention has been devoted to another type of DBMS being object-oriented system (here this term refers only to the structure of the database and the programming language, and not to the object as reality). The use of such DBMS aims to reduce the amount of stored information and the time of sequential search in the database. In GIS, such structures are used to manage complex real objects in a more reasonable way than simple points, lines and polygons, as well as modify the database when overlaying polygons. Object-oriented databases require geographic data to be presented as a collection of elements. At the same time, they are characterized by a series of attributes and behavior parameters that determine their spatial, graphic, temporal, text/numerical dimensions. Examples of such elements are as follows: a section of a railway and an associated station building; pipeline assembly with branches from pipes of different diameters, etc. This structure unifies the storage of geometry and attributes when displaying interconnected objects [2].

Object-oriented databases are determined primarily by the needs of practice, specifically, by the need to develop complex information application systems. The technology of previous database systems was not entirely satisfactory for them. Digital terrain modeling as one of the important modeling functions of geographic information systems includes two groups of operations, the first of which serves to solve the problems of creating a relief model, the second – its use. A digital terrain model (DTM) is usually understood as a means of digital representation of three-dimensional spatial objects (surfaces or reliefs) in the form of three-dimensional data constructing a set of elevation marks (depth marks) and other applicate values (Z coordinates) at the nodes of a regular or irregular network or a set of contours records (isohypsum, isobaths) or other isolines.

2. Formalized models and methods of pre-project analysis of structure of information flows and data processing technologies

Currently, methods of diagnostic study have been developed only for certain issues of analysis. In this regard, the issues of creating a general diagnostic analysis methodology and options for decomposing the control system, analysis and presentation of the characteristics of the available standard modules in order to determine the possibility of their use in the developed data processing system are of great importance. It should be noted that the stage of studying and analyzing existing control systems is 30% of the entire period of data processing system development. As a rule, this stage is completed manually [3]. Automation of control systems study and analysis requires formal models that reflect the process of data processing.

The methodology for the development of technical specifications is a methodology for the development of complete, consistent and unambiguous requirements which provide background for the further development of data processing system and determine what the system under development should do, what basic tasks and modules it should contain. The extreme importance of this stage is quite obvious, since the incompleteness and incorrect definition of what the developed data processing system should do require very large redesign costs. At the same time, the number of errors contained in technical assignments for data processing system is quite large. On average, even relatively good technical specifications contain 3 to 5 errors per page.

Currently, the requirements for data processing system are usually formulated in natural language. However, formal languages are also used to describe the requirements for the characteristics of individual modules. A number of formalized non-automated methods for presenting the results of the analysis of existing systems and requirements for modular data processing system have been developed and used. The use of these methods enables the following: to give a formal description of the input, output and all intermediate arrays both in the analysis of the existing and the designed system; determine the procedures for transforming arrays when forming the required output, as well as
determine the structure and select the technical means of the system; assess the quality characteristics of the system and choose the best option. The set of document forms provides an accurate and complete description of the entire system.

Much attention was also paid to the development of languages and standard operators, which were supposed to formalize and to some extent automate the stages of analysis and synthesis of the system. In particular, the development of information algebra was focused on the creation of a theoretical basis for automating the development of specifications, which was one of the most important achievements of the CODASYL committee [4,5]. The work experience analyzed at the stage of drawing up a technical task for the design of data processing system shows that one of the most important tasks at this stage is the decomposition of the system into subsystems (modules), which provides the extremum of a given structuring criterion regarding the convenience of subsequent detailed analysis, development and implementation. The task of allocating subsystems or modules of data processing systems having the minimum number of information links subject to the constraints on the total number of allocated subsystems will be considered below. This task arises at the stage of producing technical specifications and technical design when general requirements for the information and software system of data processing system are formed, and the functions or procedures performed by the system for processing input records obtaining intermediate and output results are determined.

Sets of different types of input, intermediate and output data and necessary procedures for their transformation are basic for the task under consideration. Informational links between data processing procedures are formalized using a multigraph, whose vertices are procedures, and the arcs connecting them are marked with numbers of data items common to these procedures or with different colors.

In the graph interpretation, the task is to define multigraph structuring with colored or labeled arcs as subgraphs. The multigraph provides a minimum of the total number of arcs of different colors connecting subgraphs subject to constraints on the total number of selected subgraphs, the number of arcs and vertices of each subgraph, and the number of links between individual subgraphs.

The analysis of information flows for each subsystem and the system as a whole is of great importance at this stage of data processing system design. To study the information flows, a number of models of data processing systems have been developed in the form of block diagrams, arrow diagrams, graphs, tables, matrices of a special type, displaying various components and characteristics of processing [6]. Based on the best practice of using various models of data processing system, a set of interrelated matrix and graph models has been proposed. The set provides a formal analysis and definition of the characteristics of the studied data processing systems at the stage preceding the technical design of data processing system, as well as formalized methods for presenting the results of studying and analyzing control systems [7].

The proposed approach used when developing a system is based on the principle of sequential equivalent transformation of matrix models of data processing systems based on the stage of system analysis and the need to obtain the required characteristics. The proposed set of models is focused on automated analysis of data processing systems and correction of initial information in a dialogue with a designer and provides information preparation for the subsequent synthesis of a modular data processing system optimal for a given criterion at the technical design stage. The initial data for the analysis, systematization and formation of requirements for the block diagram of data processing developed by data processing system is the information obtained in a dialogue with the designer about the pairwise relations between sets of data items of the data processing system. The data items are formalized as an adjacency matrix and obtained as a result of preliminary studying information about the restrictions on the structure of the block diagram of data processing system [8].

Let $D = (d_1, d_2, ... , d_s)$ be the set of data items of the developed data processing system or their sets, where $5$ is their number. Based on the information from developers and a priori information about the structure of data processing system, by an adjacency matrix we mean a square binary matrix indexed along both axes by a set of data items and containing record 1 in position $(i, j)$ if and only if, based on information from developers and a priori information about the structure of data processing system, there is such a relation $R_{ij}$ between data items $d_i$ and $d_j$ that to obtain a data item
$d_j$, it is directly necessary to refer to the data item $d_j$. For the convenience of formal consideration, we will also assume that each element is reachable from itself, i.e.

$$d_i R_0 d_j , i = 1, S. \quad (1)$$

Let us denote such a relationship between $d_i$ and $d_j$ as $d_i R_0 d_j$, and its absence – as $d_i \overline{R_0} d_j$, which corresponds to the entry 1 or 0 in the position of the adjacency matrix $B$.

The adjacency matrix $B$ is associated with a graph of information relationships $G(D, R_0)$, whose set of vertices is a set of data items, and the arc $d_i, d_j$ corresponds to record 1 in the position $(i,j)$ in the adjacency matrix $B$, i.e., it corresponds to the satisfaction of the condition $d_i R_0 d_j$.

By the reachability matrix $M$ we mean a square binary matrix indexed in the same way along both axes by a set of data items $D = (d_1, d_2, ..., d_n)$. The record 1(0) in each position $(i,j)$ $i,j = 1, S$, of the reachability matrix corresponds to the presence or absence of a reachability relation $R$ with the transitivity property for all ordered pairs of data items $d_i, d_j$ of the data processing system. A data item $d_j$ is reachable from the data item $d_i (d_i R_0 d_j)$ if a directed path from top $d_i$ to top $d_j$ can be indicated on the graph of information link $G(D, R_0)$, i.e. if data item $d_i$ is used to obtain data item $d_j$.

Matrix $M$ analysis enables to define input, intermediate output, and update data.

Data items whose rows in the matrix $M - E$ do not contain ones (zero rows) are the output elements of the data processing system, and the data items corresponding to the zero columns of the matrix $M - E$ are the input elements of the data processing system, where $E$ is the identity matrix of $E = \| e_{ij} \|$ type, where

$$e_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \quad i = 1, 2, ..., S. \end{cases} \quad (2)$$

Other data items are intermediate. Analysis of subsets of reachability $M$ relations enables to define a set of data items to be updated. Let $A(d_j)$ be the set of precedence of the data item $d_j$, $x$ is the input data item of the data processing system, and $x'$ is the updated $x$ version. Then, if for the data item

$$y \in A(d_j) x R y \land y R x', \quad (3)$$

then $x' R y$ i.e. if $x$ precedes $y$ and $y$ does not precede $x'$, then to obtain $y$, $x'$ being an updated $x$ version is to be used. By analyzing the records of the reachability matrix, it is possible to determine all sequences of data items of the $(x, d, y)$ type in the following way

$$R(x) = \{d, y\} \text{ and } A(y) = \{x, d\}. \quad (4)$$

where $x$ belongs to the set of input data items of the data processing system. Having additional a priori information from the developer about the structure of data items of the selected triplets, it is possible to indicate a set $\{d\}$ of data items being updated versions of the corresponding data items of the set $\{x\}$, i.e. $d = x'$.

The reachability matrix $M$ is determined based on the adjacency matrix $B$. In this case, the property of transitivity of the reachability relation is used. A single reachability matrix $M$ corresponds to different square binary adjacency matrices. Moreover, they are related by the Boolean equation

$$B^{n-1} \neq B^n = B^{n+1} = M, \quad (5)$$

where the exponent of $n$ degree is a positive integer which is less than the number of data processing system data items ($n \leq S - 1$). Thus, a set of adjacency matrices, any of which, as follows from equation (5), has the same reachability matrix can correspond to a given reachability matrix $M$. Accordingly, there are many graphs of information interaction, any of which contains the necessary information to build a reachability matrix.

### 3. Approach to analyzing structure of designed data processing system

The analysis of the structure of the designed data processing system and the determination of the necessary sequence for obtaining data items is simplified if the elements of the constructed reachability matrix are ordered by levels (stages) of their processing. For this purpose, on the basis of the adjacency and reachability matrices, a reordered reachability matrix, a structured adjacency matrix and the corresponding graph of information relationships are built. Input, intermediate and output data
items, feedback loops, levels and sequence of data processing are refined based on the graph. This also defines persistent data, data on updating, and current data. The procedure for separating the levels of data processing is as follows [9].

A data item \(d_i \in D\) belongs to a plurality of top-level items, reachability matrix, if \(R(d_i) \cap A(d_i) = R(d_i)\). Based on this definition, for any two elements \(d_i\) and \(d_j\) of the upper level the relations \(d_i \neq R d_j\) and \(d_j \neq R d_i\) or \(d_i \neq R d_i\) and \(d_j \neq R d_j\) are valid.

Using the reachability matrix this definition allows structuring the set of data items as subsets in accordance with the levels \(L, j = 1, 2, ..., l\) of their processing, and \(L_j\) are determined iteratively, starting from \(L_1\) on the basis of the relation

\[
L_j = \{d_i \in D - L_{j-1} = \cdots = L_j - R_j^{-1}(d_i) = R_j^{-1}(d_i) \cap A_j^{-1}(d_i)\},
\]

where \(D = \{d_i\}, i = 1, 2, ..., S\) is a set of data items, while \(R_{j-1}(d_i)\) and \(A_j^{-1}(d_i)\) are, correspondingly, a reachability set and an element precedence set \(d_i\) on subset \(D - L_j = \cdots = L_{j-1}\). The described structuring of data items corresponds to a structured graph of information relationships whose data items (points) are divided into different levels. Structuring the graph of information relationships by levels enables to highlight the sequence and main stages of data processing and processing cycles at each level.

Distributing the set of data items of the data processing system into levels leads to reordering the rows and columns of the reachability matrix. Thus, it takes a block-triangular shape: block-diagonal square submatrices are allocated along the main diagonal of the reachability matrix. They are formed by data items of the same level, to the right of which all records are equal to zero.

By the reachability matrix we mean the matrix transformed into a block-triangular form.

A large number of procedures and data items, the presence of data processing cycles at various levels can complicate the analysis and synthesis of the structure of the developed block diagram presenting data processing system. Therefore, in some cases, it is necessary to simplify the reachability matrix by reducing the cycles and thereby, reducing, the dimension of the reachability matrix [10]. If a digraph contains a path from some point to itself, then this path is a cycle. Let the \(L_j - j^{th}\) be level of the reachability matrix \(M\), \(N_j\) – the block-diagonal submatrix of the matrix \(M\), indexed by the data items of the level \(L_j\) and which itself is the reachability matrix. In this case, the matrix \(N_j\) generates the division of the set of data items of the level \(L_j\) into two subsets \(F_j\) and \(U_j^*\), whereby the submatrix indexed by the items \(F_j\) is the identity matrix, and the submatrix indexed by the items \(U_j^*\) is a block diagonal matrix, whose submatrices on the main block diagonal are completely filled with the ones, and all other submatrices are equal to zero. Let \(F_j\) be the acyclic component of the level \(L_j\). A data item \(d_i \in L_j\) is included in \(F_j\) if it is its own reachability set at the level \(L_j\), i.e. \(R_j(d_i) = d_i\). Otherwise, the data item \(d_i\) is included in the subset \(U_j^*\) that will be the cyclical component of the layer \(L_j\). In a structured graph of informational relationships, all points \(d_i \in F_j\) are isolated at their own level. The cyclical component \(U_j^*\) generates structuring the set of data items belonging to \(U_j^*\) as the subsets that form the set of cycles.

Let us formulate a new set of information elements \(D'\) as follows. Let the set \(D'\) include all data items of the set \(D = \{d_i\}, i = 1, 2, ..., S\) that are not contained in any of the cycles, and replace each set of cycles in the level with one of the elements of this set. Obviously, if there are no cycles in \(D\), then the sets \(D\) and \(D'\) coincide. Otherwise, the set \(D'\) contains fewer data items than the set \(D\). Now let us delete all rows and columns corresponding to data items not included in the set \(D'\) in the reachability matrix \(M\) for each cyclic component \(U_j^*\). Thus, we obtain a new reachability matrix \(M'\) indexed by the data items of the set \(D'\). Let us call matrix \(M'\) a condensation matrix, in which the left data items of the cycles are in the same semantic relation of reachability to other data items as the matrix \(M\) cycles which they replace. Simplification of the model of the developed data processing system using a matrix and a condensation graph in comparison with the reachability matrix and the information reachability graph facilitates the analysis and systematization of the data processing structure in the
data processing system mainly by reducing the number of relationships between data items and excluding data processing cycles. Note that in the enlarged description of the data processing structure, all data on the reachability of data items of the original data processing system structure represented by the reachability matrix $M$ are fully preserved.

A given condensation matrix $M'$ of a reachability matrix $M$ with $l$ levels ($l > 2$) corresponds to a set of adjacency matrices. Let $A$ be one of them, and therefore, it has the property

$$A^{l-2} \neq A^{l-1} = A^l = M',$$

where raising to the power is performed using Boolean multiplication and addition. Let also any matrix $B \neq A$ satisfying property (7) have a large dimension. Then the adjacency matrix $A$, which can be considered the exact root of the $(l-1)^{th}$ degree of the condensation matrix $M'$, is called skeletal.

The skeletal matrix $A$ is a matrix with the minimum number of ones that stores all the information about the reachability of data items, which is reflected in the condensation matrix $M'$, and therefore in the reachability matrix $M$. Removing any unit from the matrix $A$ violates the reachability of data items, which is reflected by the matrices $M'$ and $M$. The given condensation matrix $M'$ corresponds to a single skeletal matrix $A$. The skeletal graph corresponding to a skeleton matrix retains all information about the reachability of the condensation graph and contains the minimum possible number of arcs [10].

Let us consider the main algorithm and procedures for obtaining a skeletal matrix $A$ for a known condensation matrix based on the concept of a block-diagonal expansion of a matrix $M'$. Let $l$ be the number of levels in each of the matrices $M$, $M'$, $A$ and, in accordance with this, each matrix has $2l - 1$ block diagonals, whose number $k$ takes on value $1 - l, 2 - l, \ldots, 0, \ldots, 1 - l$. For the main block diagonal $k=0$, the block diagonals below it are given negative numbers, and above them are given positive numbers. Let us denote matrix $C(k)$ of the same dimension as the condensation matrix $M'$ and containing the block diagonal of the matrix $M'$ with number $k$, while all other entries of the matrix $C(k)$ are equal to zero. Any submatrix $G_{ij}$ of the matrix $C(k)$ indexed by a pair of levels $i$ and $j$ of matrix $M'$ is given as $C_{ij} = \delta_{j-i}^k M_{ij}$, where $M_{ij}$ is a submatrix of the matrix $M'$ indexed by levels $i$ and $j$ and a logical variable $\delta_{j-i}^k$ is defined as

$$\delta_{j-i}^k = \begin{cases} l, & \text{where } k = j - 1 \\ 0, & \text{if } k \neq j - 1, \end{cases}$$

where $i, j = 1, 2, \ldots; l; k = 1 - l, 2 - l, \ldots, 0, \ldots, 1 - l$.

By the block-diagonal extension of the condensation matrix we mean the Boolean sum

$$M' = \sum_{k=1}^{l-1} C(k)$$

or $M' = M_L + M_0 + M_R$, where $M_L = \sum_{k=1}^{l-1} C(-k) = \sum_{k=1}^{l-1} C(k)$ is called the left block-triangular component $M'$, $M_0 = \sum_{k=0}^{l-2} C(k)$ - main; $M_R = \sum_{k=l-1}^{l} C(k)$ - right block-triangular component $M'$. Since the condensation matrix is written in block-triangular form, then $M'_R = 0$ and

$$M' = \sum_{k=1}^{l-1} C(-k) = M_L + M'_0$$

The general principle of constructing a skeletal matrix $A$ requires the entry 1 in the $(-k)^{th}$ block diagonal of the matrix $A$ to be contained only in those positions in which the matrix $C(-k)$ contains the entry 1, and, moreover, no corresponding path that can be represented on the condensation graph by means of paths, already represented in the block diagonals $(-l, -2, \ldots, -k+1)$. According to this principle, the matrix $A$ can be constructed in an iterative manner using the following equations:
\[ \beta_i = \sum_{k=0} B(-k), \quad (10) \]
\[ \varepsilon_i = M' - \beta_i^{-1} = \sum_{m=1} E_i(-m), \quad (11) \]
\[ B(-k) = E_{i-k}(-k), \quad (12) \]

where equations (10) and (11) are block-diagonal extensions similar to equation (9); all operations except for subtraction are Boolean. The set of matrices \( \{ \beta_i \} \) is constructed as a sequence of best approximations to \( A \), until finally \( \beta_n = A \), when \( \varepsilon_n = 0 \), where \( n \leq 1 - 1 \). The initial conditions are \( B(0) = C(0) \) and \( B(-1) = C(-1) \). The block diagram of the algorithm for obtaining the skeletal matrix is shown in Fig. 1.

**Figure 1.** Block diagram of algorithm for obtaining skeletal matrix.

4. **Conclusion**

Thus, the construction of the canonical structure of the territory database regarding the requirements for the reliability of obtaining spatial information based on a formal analysis of reachability matrices, provides the selection of sets of cycles at each level of data processing. When developing data processing system, a more detailed analysis of cycles is required, since they are those processing areas that are repeatedly used in the process of solving a problem and require large expenditures of computer time for their implementation. In addition, their presence impedes programs development.
and debugging. The data processing cycles of the data processing system correspond to submatrices located along the main diagonal and filled with single entries of the reachability matrix presented in block-triangular form with a set of data items ordered by processing levels. The proposed models enable to carry out a formal analysis of cycles using the concepts of threshold and geodesic matrices.

References

[1] Kashtaeva S V 2020 Mathematical modeling: textbook (Ministry of Agriculture of the Russian Federation, Perm Agrarian and Technological University named after academician D.N. Pryanishnikov ", Perm: IPC "Prokrost") 112 p
[2] Lebedev S V, Nesterov E M 2018 Spatial GIS modeling of geoeological objects in ArcGIS (St. Petersburg: Russian State Pedagogical University named after A. I. Herzen) 280 p
[3] Baranova I V, Batova M M, Zhao K 2020 Information tools for digital transformation of high-tech enterprises (Moscow: Creative Economy) 222 p
[4] Barsegyan A A, Kupriyanov M S, Stepanenko V V, Kholod I I 2004 Data analysis methods and models: OLAP and Data Mining (SPb: BHV-Petersburg) 336 p
[5] 1986 Understanding GIS: The ARC/INFO Way, ESRI, Redlands, CA. 4. Fleet. H. «SAGIS: a full-function public-domain GIS for micro and m inicomputers»
[6] Chrisman N R 1990 Deficiencies of sheets and tiles: building sheet less databases International Journal of Geographical Information Systems 4 157-67
[7] Berry B J L, Baker A M 1968 Geographic sampling (NJ: Prentice Hall, Englewood Cliffs)
[8] Zvonarev S V 2019 Fundamentals of mathematical modeling: study guide (Yekaterinburg: Ural Publishing House) 112 p
[9] Krutskikh V V 2021 Modeling in labview: a textbook for universities (Moscow: Yurayt Publishing House) 171 p
[10] Prigarin S M 2020 Statistical modeling of multidimensional Gaussian distributions: a textbook for universities (Moscow: Yurayt Publishing House) 83 p