NEW TRENDS IN THE DEVELOPMENT OF COMPUTERIZED STATISTICAL INFORMATION SYSTEMS

Pavol Dujnič*

August 1982
CP-82-46

*Computer Research Center
Bratislava, Czechoslovakia

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

PREFACE

This paper is one of a set of background and research papers for the study of information systems for regional planning. The spatial dimension of information systems as a decision aid in regional development planning has too often been neglected. Therefore, much more attention should be paid to the design and development of information systems reflecting socio-economic processes, so as to arrive at a better representation of regional systems and a better adaptation to the needs of regional planners. The major aim of the study is to provide in a systematic way a set of guidelines to be taken into account in the design and use of information systems for regional planning.

The present paper by Pavol Dujnič (Computer Research Center, Bratislava, Czechoslovakia) describes a set of informational models to be used as a base for modern computerized statistical information systems.

Laxenburg, July 1982

Boris Issaev
Leader
Regional Development Group

CONTENTS

NEW TRENDS IN THE DEVELOPMENT OF COMPUTERIZED STATISTICAL INFORMATION SYSTEMS

Pavol Dujnič

1. INTRODUCTION

Information systems are an essential aid to successful decision-making, hence the increasing attention they have received in recent years. There are many important avenues for discussion with respect to their development. In this paper, we concentrate mainly on technological problems associated with improving the accessibility, comprehensiveness, accuracy, and timely availability of the output of computerized statistical information systems (SIS). SISs have a direct link with and influence on regional information systems, being their principal source of data. Moreover, they are hierarchically organized with a technical base that is adequate for performing the complete cycle of data processing and evaluation. For these reasons, SISs are frequently utilized by regional planners.

The paper is organized along the following lines. Sections 2 and 3 present a model of SIS together with its data bank system. Section 4 outlines approaches to the interactive evaluation of economic data, and the final section describes the metainformation system—a regulatory tool for SIS.

2. GENERAL MODEL OF THE STATISTICAL INFORMATION SYSTEM

SISs process statistical data and present the resulting information in a form suitable for the user, for example, from the statistical services. The organization of such services varies considerably from country to country. Nevertheless, regardless of the organizational form, it has to perform certain basic functions. The procedures related to these functions are similar in all statistical services.

Thus, we are able to construct a general model of a computerized information system that is more technologically oriented. The individual blocks included in this model, together with their interrelations are indicated in Figure 1. There are three main structural elements: an input-oriented subsystem, an output-oriented subsystem, and a metainformation system. The emphasis of the input subsystem is on data gathering, micro-processing, and primary data storing. The major macro-processing and presentation work is done within the specialized output subsystem after the "clean" micro data and selected aggregate macro data have been placed in the data bases. The basic information about the components of the information system is organized by the metainformation system. It should be emphasized that the user plays an active role in operating the system. He triggers the information production cycle by his request for information.

When Klas (1979) and Sundgren (1981) formulated their models of a statistical information system (which are similar to ours), they emphasized the dynamic attributes of the individual elements and of the system as a whole. They indicated that the construction and implementation of a statistical information system would be a long-term process because of the complexity of the models used. In general, an architecture such as we describe here should in itself be of help in understanding and designing the individual statistical information system. It provides the designers with a basic structure with which to work. Of course, not all design problems can be solved simply by adopting a uniform architecture. The architectural pattern outlined in this paper is certainly not very detailed and the structure is crude. Indeed, one should

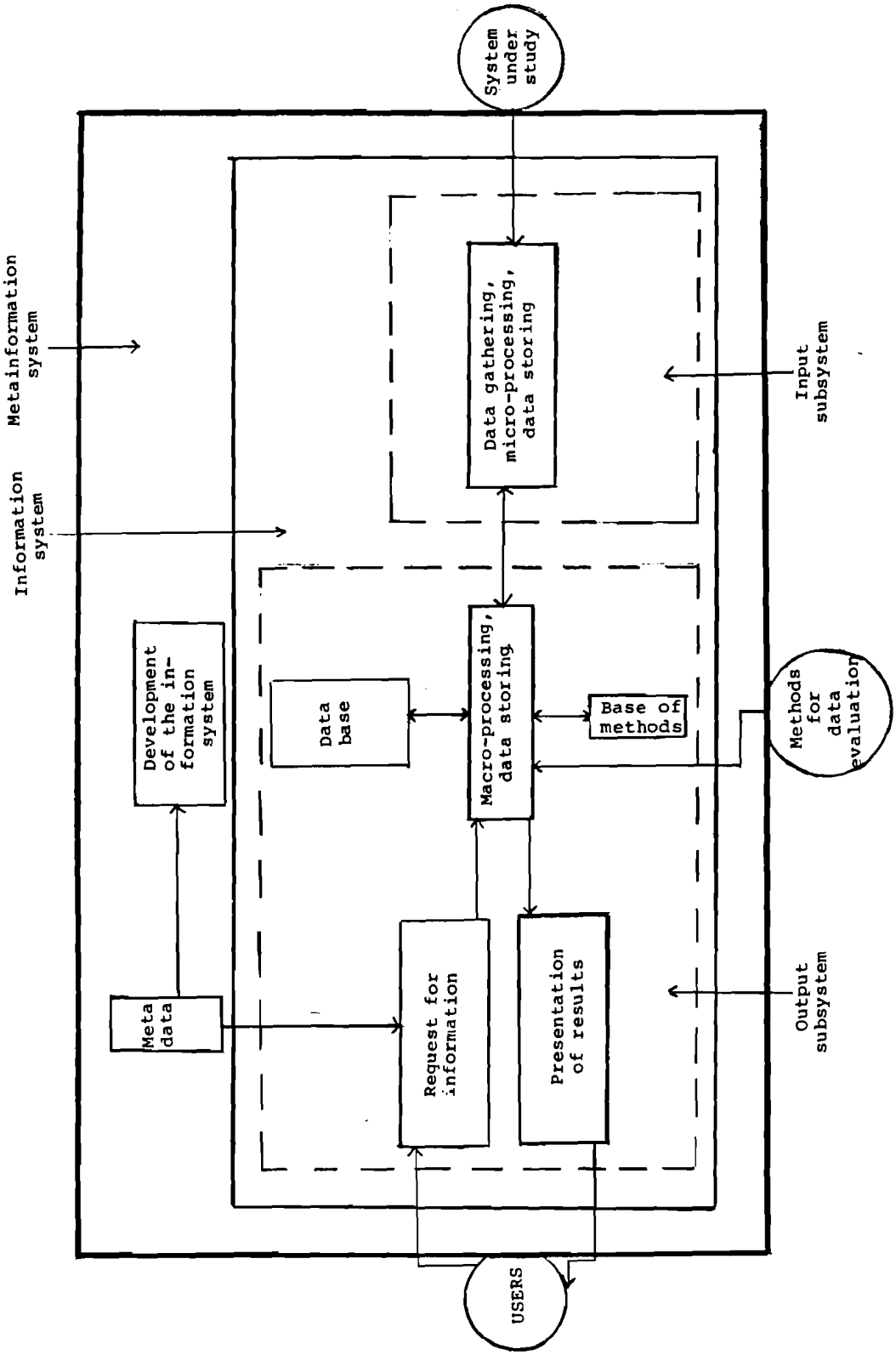


Figure 1. Model of the computerized information system.

expect to have problems when adding the detail. For example, it could be problematic on a technological level to select one among several alternative software products for solving a specific subproblem, or to choose between standardized software and tailor-made programming.

It is impossible in a short paper to present a more complete overview of all existing or emerging technological problems. However, in order to give an idea about recent progress in the area, we shall discuss some (sub)problems associated with:

- developing data banks for a statistical information system (this can easily be viewed as the central structural problem);
- applying more efficient methods for interactive communication between user and computer system (the construction of interactive systems of work based on direct communication is one way of providing the untrained user with access to the computer);
- developing the system such that an experienced user can work without a programmer's intervention (adequate results can be achieved in this mode by improving the system's analytical and forecasting activities);
- creating a qualitatively new tool for using, managing, and improving the information system effectively. The metainformation system can perform such a function.

3. THE DATA BANK WITHIN THE STATISTICAL INFORMATION SYSTEM

The general model of the statistical information system includes a data bank system, as shown in Figure 1. The advantages currently offered by the data bank are:

- it provides several users with the retrieved information;
- it generates specific responses to specific requests;
- it reduces the volume of redundant data in the data bases, thus providing a greater degree of integration and protection;

— it allows for a more dynamic approach to working with data.

The first phase of establishing data banks within statistical information systems has already been accomplished. This is evident from studying the progress of data bank projects in a number of developed countries. In recent years several new trends have become evident:

— statistical services have designed and constructed hierarchical data bank systems;

— there has been a move towards the use of non-procedural languages intended for non-programmer users in order to promote interactive forms of communication;

— more sophisticated types of data base model have been implemented (network models of the CODASYL type and the relational data model);

— economic and mathematical-statistical methods are being used to evaluate the data extracted from the data base;

— more complex methodological tools are being employed in data base design.

We shall now describe these trends in the context of the research being conducted at the Computer Research Center (CRC), Bratislava.

3.1. The System of Statistical Data Banks

The main goal in the area of statistical information systems is to create an integrated hierarchical system of data banks. The system of data banks will include a federal data bank (FDB), federal sectoral data banks, and regional data banks.

The function of FDB is to store, to update, and to evaluate the indicators characterizing the main directions of national economic development. These indicators create an integrated, stabilized (with regard to content and methodology), open system, which is able to respond to the management and planning requirements of the national economy. The information is produced primarily for top management.

Federal sectoral data banks comprise selected data for a cross-section of national economic problems (e.g., scientific and technical development, prices, capital assets) and include detailed information on individual sectors of the national economy (e.g., industry, agriculture). Federal and federal sectoral data banks have a design similar to that of typical statistical data banks. The data they contain are collected and processed by the statistical services.

Regional data banks contain data on the economic indicators of specific territorial units. This information is intended for regional management bodies. Regional data banks should preferably store data on population, manpower, natural resources, production sectors, nonproduction sectors, technical infrastructure, living standards, and so on. We wish to stress that the data entering the regional data banks is obtained from a variety of sources. Figure 2 illustrates their relations and structure.

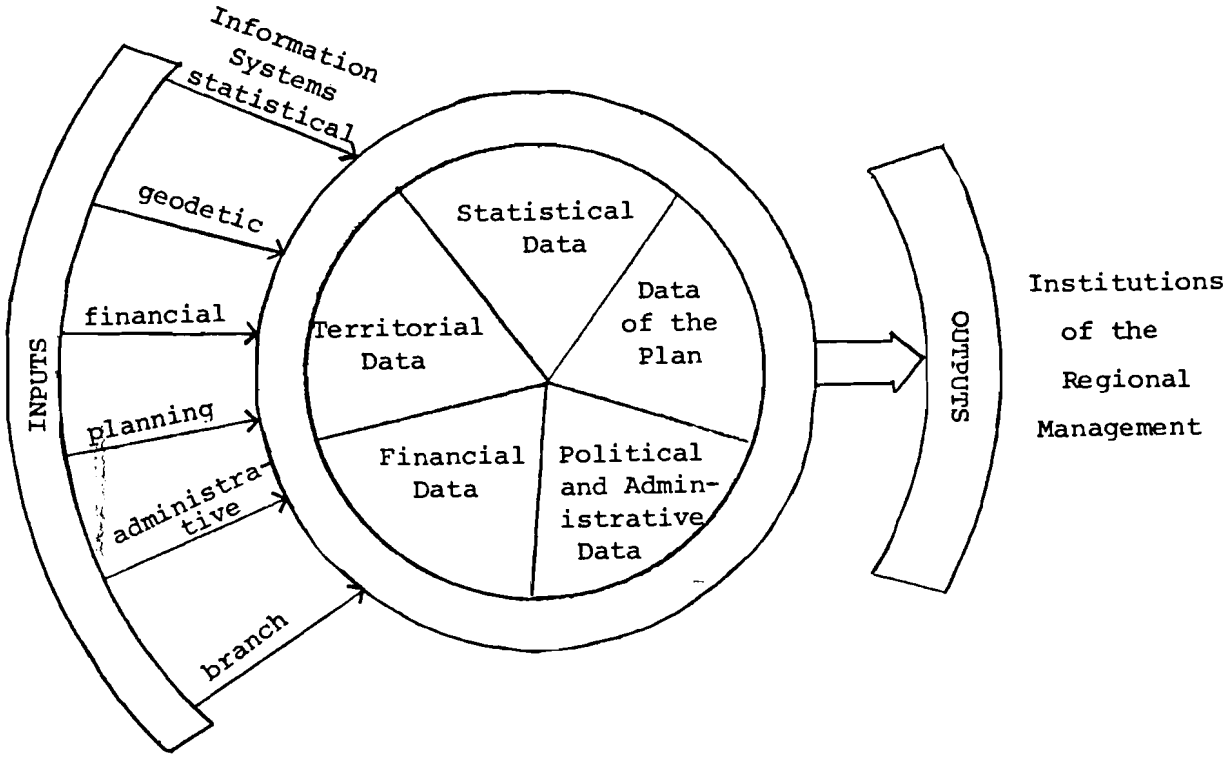


Figure 2. The relations and structure of the regional data bank.

3.2. The Structure of the Data Bank System

The three types of data bank mentioned in Section 3.1 have a uniform structure consisting of four main elements (Figure 3):

- metasystem tools,
- data base,
- bank of methods,
- data bank and bank of methods software.

Each element is discussed below.

The metasystem tools of the data bank are a set of purpose-designed catalogues and registers of indicators. They have three functions:

- an information function allowing the user to identify the content of the data and the way and form in which they are stored;
- a communication function allowing the user to operate the data bank in an interactive fashion using a non-procedural statistical language;
- a linkage function connecting the indicators and the operating data, thus effecting the search, retrieval, and output of the required data.

The purpose-designed catalogues are intended to:

- give a unified formal description of the data base;
- aid in the design of uniform user language;
- protect and integrate the data stored in the data base;
- consider the links with the metainformation system's current catalogues describing the entire statistical information system (see Section 5).

The data base contains the files of data segments, in which there are data on the system under study organized in a certain format, i.e., as statistical data characterizing the socioeconomic entity. There are two types of segment--macrodata segments and microdata segments (on individual organizations or reporting units).

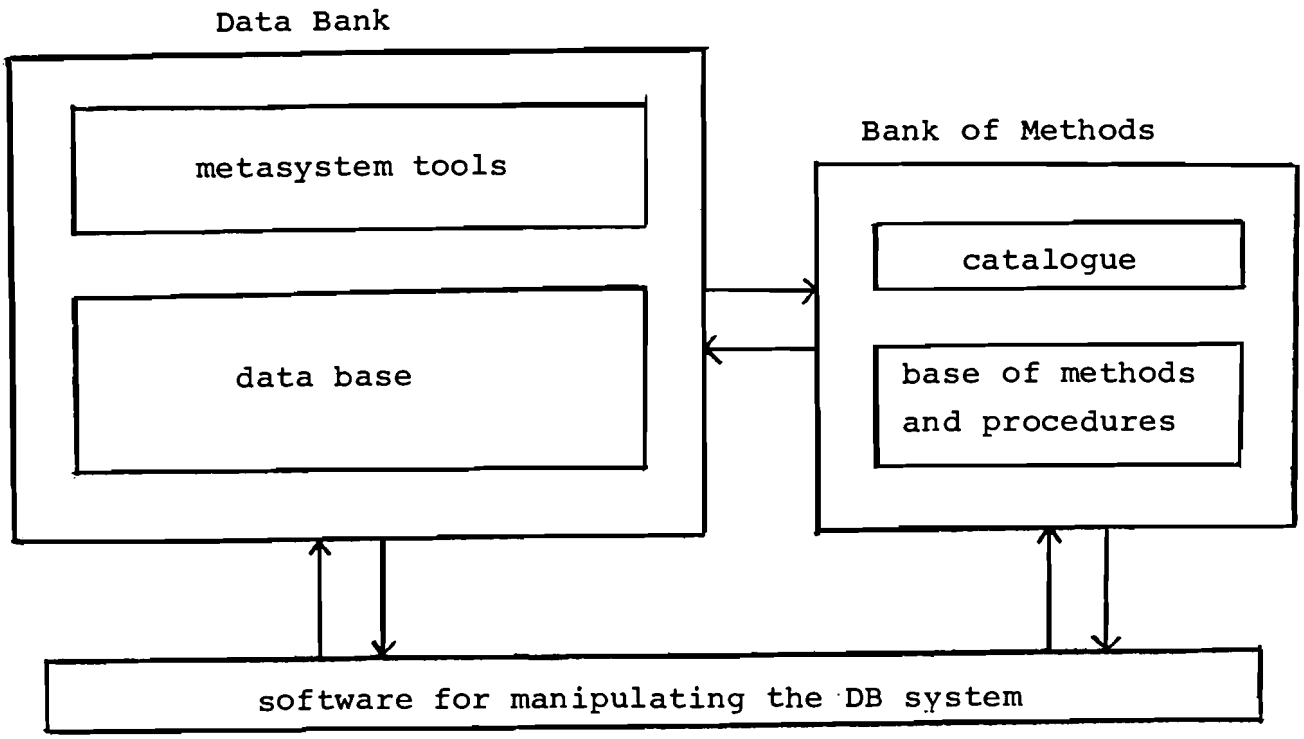


Figure 3. Structure of the data bank system.

The core of the bank of methods is the base of methods and procedures,in which the computer programs for implementing the mathematical and statistical procedures and methods are stored. These procedures and methods are employed by the user for computing certain information from data in the data base. The information characterizing and documenting the procedures and methods of these programs is stored in the catalogue (as in the case of the metasystem tools).

The software for manipulating the data bank system includes a set of software products for:

— manipulating the system (a typical function of data bank management systems);

— interactive operation of the data bank by the user;

— generating, maintaining, and modifying the procedures and methods;

— user communication with the base of methods in order to combine the necessary parameters and, in this way, to activate the computing procedure.

The concept of the bank of methods is based on an interactive approach that allows the user to evaluate data in the data bank. We define the term "method" as a sequence of steps in which the economic-statistical, mathematical-statistical manipulations specified by the user are performed on data retrieved from the data base. We define the term "procedure" as a program module for executing certain complete algorithms. The user can thus employ the prepared method or construct his own based on the prepared version. In such a case, he should be supplied with a special language for integrating the procedures in different ways. This will allow him to examine the method flexibly and to evaluate the alternatives.

4. APPROACHES TO THE INTERACTIVE EVALUATION OF ECONOMIC DATA

4.1. General Remarks

The construction of interactive systems of work, based on direct communication between user and computer system, is currently underway. This can be viewed as an extension of the effort to provide the user with maximum support for improving analytical and forecasting activities, which are becoming an integral part of statistical practice.

The user participates in formulating the processing requirements during operation of the current processing cycle. The main advantage of the interactive method over batch mode processing is that it allows the user to intervene in the processing cycle. In addition, interactive computing allows him to adapt his requests for automated processing with a reduced volume of effort. As he gradually becomes familiar with the possibilities provided by the computer, he is able to select and process the information more effectively, thus improving the efficiency of the operation.

Non-procedural languages applied in the interactive mode have certain special features. The user formulates the task in non-procedural language by stating the required result, regardless of the format of the data and without specifying the computing procedure to be used. In other words, if a non-procedural

language is used, he does not need to know how the data are stored, their format, or the particular procedure for computing the required result.

To obtain the required result with problem-oriented languages, the user with a minimum knowledge of how to operate the computer is able to formulate his request directly onto the computer terminal without the aid of a programmer. The computer is used in this way as an everyday working tool for an increasing number of professions.

We now consider the two interactive modes of operation: the simple interactive mode and the conversational mode. In the simple interactive mode, communication is controlled by the user, who places his requests to start operating the programming system. The computer responds by signalling the end of the operation and its preparedness to receive further statements. At the request of the user, the results are selectively displayed. In the conversational mode, communication takes place in "question-answer" form. The computer displays a list of available operations and the user selects the one he wishes. The computer successively demands a more detailed specification of, for example, essential parameters and data, and the user gives his answer, and so on. The operation is performed by the programming system only when all of the computer's questions related to this particular operation have been adequately specified. The conversational mode minimizes the volume of information that the user has to remember in order to communicate with the computer. This explains its popularity among lay users. However, for an experienced user or for a programmer it is too slow and clumsy, since it obliges him to answer many questions.

When considering statistical analyses performed in the interactive mode, one should bear in mind that for certain types of statistical analysis the sequence of individual steps may be specified a priori whereas for others the sequence of steps is stochastically dependent on the results achieved at each preceding step. Thus, the core of the work consists in evaluating the results and in determining how the analysis should be continued.

The ideal state is for the automated programming system to perform the evaluation and to recommend the subsequent step in the analysis. Methods for automated evaluation of the results have not yet been worked out. The chief difficulty here lies in formulating the often heuristic decision-making procedures of the user; this is a complex and often impossible task. Up until now automated evaluation has not received sufficient attention. Therefore, at present the responsibility for evaluating the results and for deciding how to proceed with the analysis lies entirely with the user. If the communication is carried out in the conversational mode, a mere display of functions available at further steps is of no practical value to him. This list is usually large and disturbs the fluency of communication unnecessarily. In such cases, the simple interactive mode of operation may be more suitable.

4.2. Components of Interactive Economic Data Evaluation

We shall now consider the three components of the data evaluation process: data base, evaluation methods, and software environment.

4.2.1. Data Base

The data base used for economic analysis on a macroeconomic level may be referred to as a working data base, since its characteristics are determined by the economic problem under investigation. The retrieval of data loaded in the base is also influenced by these problems. Although the data may be obtained from a variety of sources, for the sake of simplicity we assume that the source is a centrally controlled more complex data base. A limited quantity of relevant data, which usually have a homogeneous structure and are considered to be independent (i.e., there are no observed relations between them), are included in the working data base. These data may be further transformed and edited in different modes to fit the requirements of the analysis, as a result of which the data base acquires a "personal" character particular to a given user-analyst, or group of users.

4.2.2. Mathematical and Statistical Methods

The method(s) chosen for the analysis depend on the economic problem to be solved. Since the details pertaining to the use of these methods in the interactive mode is described in Section 4.1, here we need only add that the user should not access the computer until he is familiar with the sequence of steps for performing the analysis. This sequence of steps includes the selection of analytical methods and suitable data, analysis of the interim results, multicriteria decision-making, and so on.

4.2.3. Software Environment

The software should allow the methods to be applied to analysis of the data in a flexible way. However, it does not offer the user any support in evaluating the results (e.g., from the point-of-view of statistical acceptability) or in multicriteria decision-making. The success of an interactive evaluation of economic data depends on the form of communication between software and user. The software is specialized in the sense that it fits the conditions affecting the components described above. These conditions are summarized below.

1. An inexperienced user will operate the computer.
2. An interactive mode of operation with the computer is employed.
3. Given the problems to be solved, a simple non-procedural language is the most suitable means of communication.
4. The user is entirely in command of the analytical process and selection of relevant data.
5. The computer provides only limited support in guiding the user along the analytical course.
6. Strictly modular access is maintained during the application of mathematical and statistical methods.
7. The user needs interim and final results to be presented immediately after processing on the terminal, and/or on another device, primarily the printer (in tabular, graphic, and other forms).

8. During user-computer communication, attention should be paid to diagnosing and announcing errors that occur.

4.3. Features of the CRC Software for Interactive Evaluation

Operation of the basic version of the interactive evaluation programming system is illustrated in Figure 4.

The syntax of the software communication language consists of key words that define the activated system functions and of parameters containing a more detailed description of items on which and with which the activated functions operate. The system functions are divided in accordance with the complexity of the record as follows:

— single-statement

KEY - WORD < parameter > *

— multi-statement

KEY - WORD - 1
{KEY - WORD - 2 }
{< parameter > * }
END

Each statement is analyzed and executed independently. Some statements have global validity, i.e., for some functions the parameters included are preserved as predefined values. An "interface" module aids the user in familiarizing himself with the system. It also gives him a rapid orientation during interactive work. The most important information stored in the module includes:

— rules for operating the system,

— a description of the communication language syntax,

— a catalogue of available methods,

— a catalogue of diagnostic messages.

The main executive components of the programming system are two relatively autonomous program sets: the system for manipulating the data base and the system for interactive data evaluation. The data manipulation system has three functions:

— it defines the working data bases;

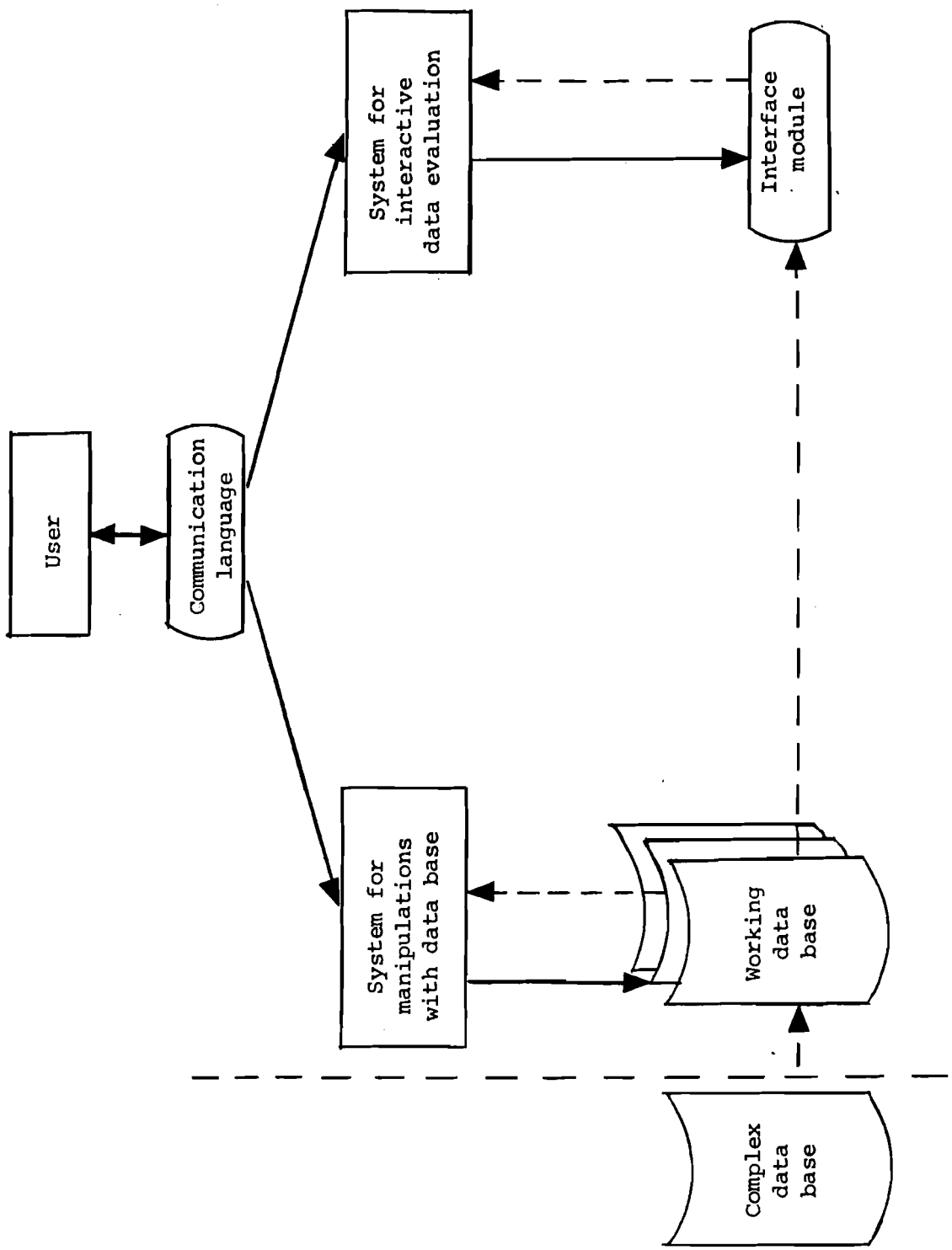


Figure 4. Operation of the programming system.

— it creates the working data bases;

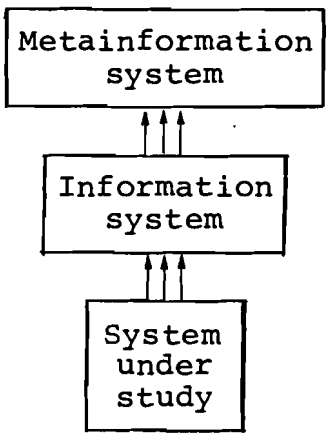
— it maintains the working data bases (adding and deleting records, generating new records, transferring selected records between bases, changing record values, and so on).

The interactive data evaluation system selects the required analytical methods from those available in the system and applies them (in the open system they may be complemented by additional methods). The system also includes a function that presents the output in the form of tables or simple graphs on the computer and terminal devices available.

5. METAINFORMATION SYSTEM

5.1. General Remarks

Present-day large information systems—computerized statistical information systems belong in this category—have two distinctive features. Their structure is complex and comprehensive and they require a technologically advanced level of equipment. At the current stage of their development, further improvement and more efficient use cannot be guaranteed if currently available control and administrative tools are applied exclusively. A qualitatively new tool is therefore needed. We refer to this new instrument, which serves as an information source for the main system, as a "metainformation" system (METIS). The relations between the information and metainformation systems are indicated below:



METIS is focused on the entire information system of the system under study with all its structural classes of elements (including data, personnel, equipment, methods, and information). It can also serve as a common tool for several related information systems. We should emphasize that our interpretation differs from that of most other researchers, who focus the metainformation system only on a specific part of the information system represented by two structural classes of elements—data and information.

The core of the metainformation system consists of a set of catalogues, each of which contains a description of certain elements. The general architecture of the set of catalogues is outlined in Figure 5.

We will now describe the user-oriented functions of the metainformation system.

5.2. Functions of the Metainformation System

METIS has five main user-oriented functions: documentation, information, administration, rationalization, and integration.

5.2.1. Documentation Function

The documentation function results directly from the formalized description of the given information system, its elements, and activities. In addition, to ensuring that the metadata are coordinated and consistent, this function continually updates the metadata base. It is also comprehensive, i.e., it covers the entire information system, making full use of the advantages of centralized documentation. This does not mean that a subsequent decentralization is not possible, if necessary, in a directly processable form on suitable storage media.

The principal users of this function of METIS are the personnel responsible for improving and operating the information system (in our case the statistical information system), i.e., designers and administrators.

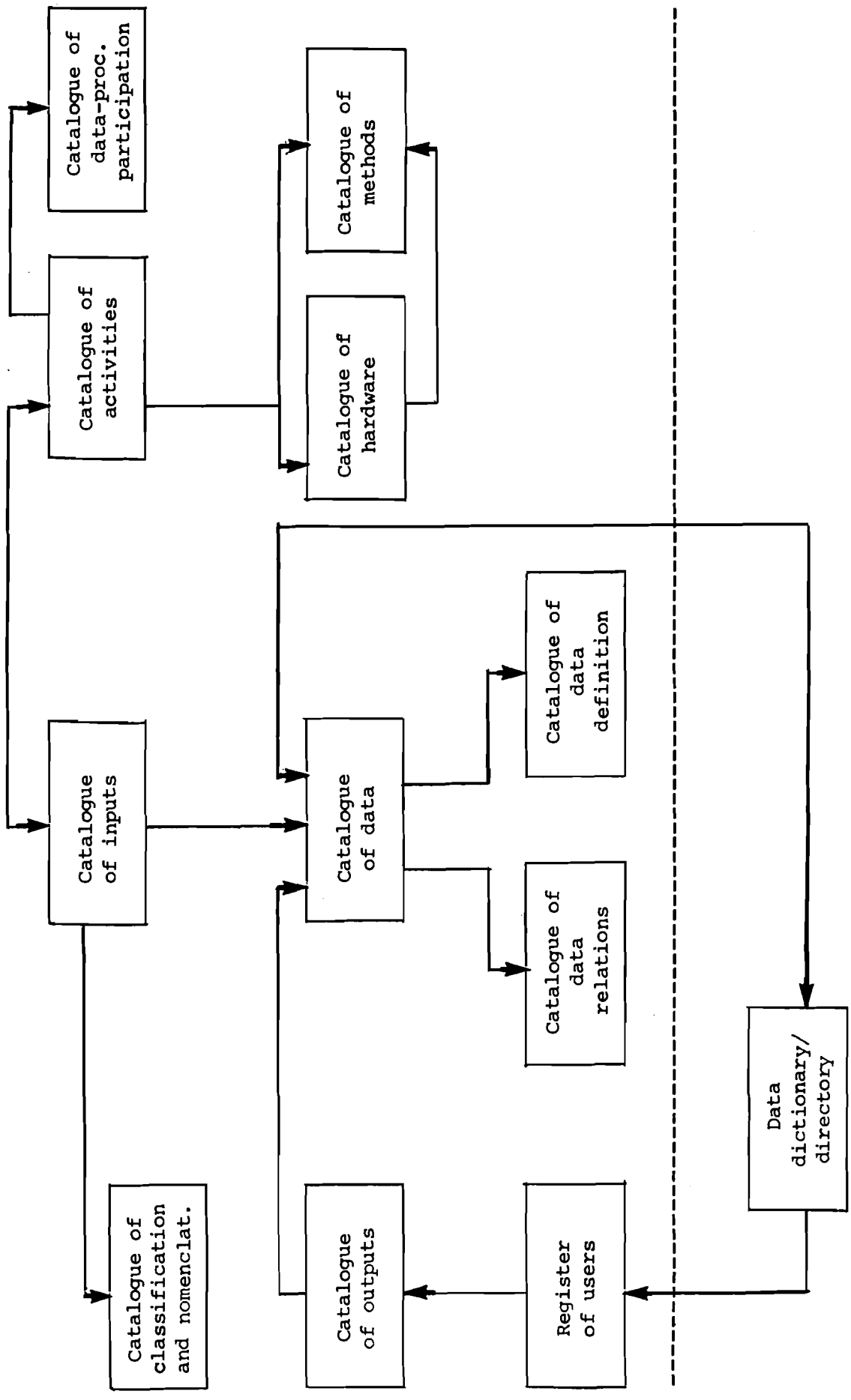


Figure 5. The general architecture of the set of catalogues.

5.2.2. Information Function

Another qualitatively new feature of METIS, in addition to its comprehensiveness, speed, and straightforward organization, is its capability of providing information, in the form of questions, directly according to the user's demands. It can be operated in the conversational mode by which it answers precisely specified enquiries, thus enabling the user to state his questions in increasingly precise terms. These features are especially important with respect to the relation between METIS and the end-user. METIS provides basic information on its own capabilities and main functional divisions. It also gives detailed information on the data and other components of the information system and, through its retrieval capability, provides access to data.

It is an appropriate tool for two groups of end-user: the inexperienced untrained user and the user-analyst. The former can use METIS to gain an idea of the contents of the information system and its potential for information provision. In contrast, the user-analyst can use METIS to gain information on the structure of the information system and on the degree to which relatively detailed statistical data cover specific socioeconomic phenomena over long periods. This type of user also requires a survey of analytical methods.

5.2.3. Administration Function

The administration function basically consists in manipulating the data of the information system using the data dictionary, which has a link to information on physical data storage on media, their organization, and so on. The main user of this function is the manager of the information system's data base.

5.2.4. Rationalization Function

Although METIS as described in the above-mentioned functions has a primarily representational purpose, it has greater potential as an active tool, i.e., for rationalizing the information system.

Two aspects of the rationalization function should be noted. First, as a rationalizing element METIS can be inserted directly into the structure of the information system. For example, the METIS catalogue system can be used to rationalize communication between information system and user. Second, METIS can be used as a source of supporting documents for improving the information system. Here it does not have a direct rationalizing function but is employed as a tool for improving the efficiency of the system. For example, it uses formalized descriptions of statistical indicators for analyzing potential redundancies.

5.2.5. Integration Function

Since the process of improving not only statistical information systems but information systems in general has been oriented primarily towards integration, we should define the integration function of METIS. It is evident that this function can be extended, since the METIS catalogue enables us to investigate the structure of the information system including all its elements and relations. We can therefore determine how the internal consistency of the information system may be improved, so that eventually a fully integrated SIS (integrated in the sense of coordination of all groups of elements) can be constructed.

In the future we will be able to employ METIS not only as a tool for integrating the information system internally (among the individual classes of elements) but also for integrating the system with other information systems.

The principal users of the integration function are the designers of the information system.

REFERENCES

Condík, S., and F. Hajnovic (1982) Terminal oriented software for the construction and application of models in macroeconomic statistical analysis and forecasting. In: Report for ISIS '82 Seminar No. CES/SEM.15/R.10, organized by the Economic Commission for Europe, 10-14 May 1982, in Bratislava, Czechoslovakia (in English and Russian).

Computer Research Center (1981) The Central Data Bank, the System of Data Banks. Research Paper 163. Bratislava: Computer Research Center (in Slovak).

Dujnič, P., and J. Frankova (1979) Organization of data in computerized information systems. In: ALFA. Bratislava (in Slovak).

Klas, A. (1978) Integrated Statistical Information System. Research Paper 146/1. Bratislava: Computer Research Center (in Slovak).

Lieskovsky, P. (1982) The data bank model in the automated statistical information system. In: Report for ISIS '82 Seminar No. CES/SEM.15/R.14, organized by the Economic Commission for Europe, 10-14 May 1982, in Bratislava, Czechoslovakia.

Nijkamp, P. (1982) Information Systems for Multiregional Planning. CP-82-27. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Soltes, D. (1979) Metadata bases—Data dictionaries/directories. In: Report for ISIS '79 Seminar No. CES/SEM.11/R.8, organized by the Economic Commission for Europe, 10-14 September 1979, in Bratislava, Czechoslovakia.

Stibic, V. (1980) A few practical remarks on the user-friendliness of online systems. Journal of Information Science 2.

Sundgren, B. (1981) Statistical Data Processing Systems—Architectures and Design Methodologies. Paper presented at the Golden Jubilee Celebration Conference on "Statistics: Applications and New Directions", 16-19 December 1981, held at the Indian Statistical Institute, Calcutta.

Swatek, S.D. (1979) The design of time-series data banks. In: Report for ISIS '79 Seminar No. CES/SEM.11/R.9, organized by the Economic Commission for Europe, 10-14 September 1979, in Bratislava, Czechoslovakia.