11 INFORMATION AND INFORMATION SYSTEMS IN WATER MANAGEMENT

The present information problem, i.e., generation, acquisition, storage, effective processing and retrieval of information, cannot be solved by conventional facilities and procedures. *Automation of information systems* using system approach, computers, modern methods of message transmission, graphic storage devices, data communication equipment, and in addition, various activities of information systems, are necessary to cope with the information explosion.

The information systems include elements from conventional branches (libraries, bibliography, document processing, etc.) and the elements inherent in automation such as programming of computers, cybernetics, logics, linguistics, semionics, statistics, mathematics, and system sciences.

The automation of information processes requires new classification and processing of knowledge in new adequate forms. In many cases an algorithmization of procedures is necessary for computer coding.

The systems approach and application of systems sciences seems to be the most effective method for the identification and analysis of information and for the automation of information processes.

11.1 INFORMATION AND ENTROPY

Information can be defined as a measure of freedom of choice in selecting a message; it is a negenergetic value proportional to the decrease of entropy (disorganization) of a system. This definition is rather broad and further definitions using different aspects of information can be used (syntactic, semantic and pragmatical information).

As the computer is the basic tool in the automation of information systems, the term information will be used to mean syntactic information with the exception of section 11.1 wherein the definition of information is used in relation to entropy. As far as the general meaning is concerned, the term information will be used as a *message about reality* that was accepted as information and can influence the behaviour of the accepting body. The *information system* is a set of subsystems, facilities and persons involved in the acquisition, storage, processing, transmission, retrieval and dissemination of information. In the automation of information systems certain functions of the live part of information systems are allocated to the "non-live" part.

The flow of information is a continuous process of information processing in sequences, determined by information inputs that are often the information outputs of previous information processing.

The set of facilities over which the signal is sent is called the communication *channel*. This notion of a channel includes (1) all the technical facilities which transform the signal before transmitting, (2) transmitting and receiving, (3) the transformation of the received signal, and (4) all the space used for transmitting the signal from transmitter to receiver.

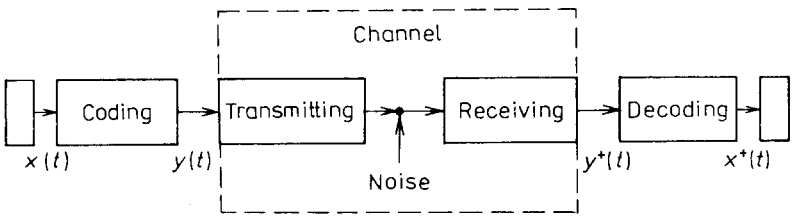

Fig. 11.1 Scheme of a communication system

In Fig. 11.1 there is a schema of a communication system and its elements. The symbols are: $x(t)$ = information source, $y(t)$ = the signal on the channel input (coded message), $y^+(t)$ = the signal on the channel output, and $x^+(t)$ = the decoded (reconstructed from signal) message in the receiving end of the communication system.

The procedure of coding and decoding, i.e., the transformation from message $x(t)$ to signal $y(t)$ and from signal $y^+(t)$ to $x^+(t)$ is expressed mathematically in the form of a functional relationship $y = f(x)$ and $x^+ = \varphi(y^+)$. A general case where the relationships between x and y and between y^+ and x^+ are stochastic was investigated by Kolmogorov (1956).

If the same output signal corresponds to a given input signal, the channel is called noiseless. In such a communication process the transmitted signal is identical with the received one and the input and output messages are two forms of realization of the same message.

The presence of *noise* disturbs this relationship between the transmitted and the received signal. Then both messages and noise form stochastic processes, and in this situation it is possible for different output signals to correspond to one message, or the received signal can be identified with different message realizations. In principle, two basic tasks are performed – the determination and separation (filtration) of signals hidden in noise. Therefore the theory of information can use the methodology of the probability theory and the theory of stochastic processes.

The theory of *entropy* is one of the important parts of the probability theory.

A sample space R partitioned into a set of mutually exclusive and exhaustive random events $A_1, A_2, \ldots, A_n$ will be the starting point. In each experiment only one event can take place. For $n = 2$ a simple alternative exists – a couple of contradictory events. If the events $A_1, A_2, \ldots, A_n$ together with their probabilities $p_1, p_2, \ldots, p_n$ ($p_i \geqq 0$, $\sum_{i=1}^{n} p_i = 1$) are given, a finite schema is thus defined (in the case of $n = 2$ it is known as the Bernoulli scheme):

$$A = \left\| \begin{matrix} A_1, A_2, \ldots, A_n \\ p_1, p_2, \ldots, p_n \end{matrix} \right\| \tag{11.1}$$

Every finite scheme describes some state of uncertainty: before the performance of an experiment, the results of which the events have to be $A_1, A_2, \ldots, A_n$ only the probabilities of the possible results are known. The degree of uncertainty is different in different schemes. In two simple schemes:

$$\left\| \begin{matrix} A_1, & A_2 \\ 0.5, & 0.5 \end{matrix} \right\| \quad \left\| \begin{matrix} A_1, & A_2 \\ 0.99, & 0.01 \end{matrix} \right\|$$

the first definitely has more uncertainty than the second one, where the result will "almost certainly" be A_1. In the first example no prediction is possible.

In applications it is desirable to introduce a degree of uncertainty of a finite scheme. This measure was defined by Shannon (1948) as

$$H(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{n} p_k \lg (p_k) \tag{11.2}$$

All logarithms in (11.2) have an arbitrary, but common base. For $p_k = 0$ by definition $p_k \lg (p_k) = 0$ is used. The quantity $H(p_1, p_2, \ldots, p_n)$ is called entropy of a finite scheme (11.1).

The properties of entropy:

1. The necessary and sufficient condition for $H(p_1, p_2, \ldots, p_n) = 0$ is that one of probabilities $p_1, p_2, \ldots, p_n$ is unity, and all the others are zero. It is the case when the result of the experiment can be predicted with a total certainty – i.e., no uncertainty is present. In all other cases entropy is positive.

2. The scheme when all the results are equally probable, i.e., $p_k = 1/n (k = 1, 2, \ldots, n)$ has the highest uncertainty.

3. Suppose A, B are two finite schemes

$$A = \left\| \begin{matrix} A_1, A_2, \ldots, A_n \\ p_1, p_2, \ldots, p_n \end{matrix} \right\| \quad B = \left\| \begin{matrix} B_1, B_2, \ldots, B_m \\ q_1, q_2, \ldots, q_m \end{matrix} \right\|$$

and suppose these schemes are independent, i.e., the probability that both events A_k and B_i occur simultaneously is $p_k q_i$. The set of events $A_k B_i (1 \leqq k \leqq n; 1 \leqq i \leqq m)$

with probabilities π_{ki} forms a new finite scheme that is called the fusion of schemes A and B and denoted AB. Suppose $H(A)$, $H(B)$, $H(AB)$ are entropies of schemes A, B, and AB, respectively, then:

$$H(AB) = H(A) + H(B) \tag{11.3}$$

It follows from:

$$-H(AB) = \sum_k \sum_i \pi_{ki} \lg(\pi_{ki}) = \sum_k \sum_i p_k q_i (\lg(p_k) + \lg(q_i)) =$$

$$= \sum_k p_k \lg(p_k) \sum_i q_i + \sum_i q_i \lg(q_i) \sum_k p_k = -H(A) - H(B)$$

4. Let us investigate the case where schemes A and B are dependent. Denoting q_{ki} as the probability that in scheme B event B_i will occur on the condition that event A_k has occurred in scheme A, we get:

$$\pi_{ki} = p_k q_{ki} \; (1 \leqq k \leqq n, \; 1 \leqq i \leqq m)$$

$$\text{Then } -H(AB) = \sum_k \sum_i p_k q_{ki} (\lg(p_k) + \lg(q_{ki})) =$$

$$= \sum_k p_k \lg(p_k) \sum_i q_{ki} + \sum_k p_k \sum_i q_{ki} \lg(q_{ki})$$

For each k, $\sum_i q_{ki} = 1$ and the sum $-\sum_i q_{ki} \lg(q_{ki})$ is the conditional entropy $H_k(B)$ of scheme B computed on the assumption that event A_k has occurred in scheme A. We get

$$H(AB) = H(A) + \sum_k p_k H_k(B)$$

The conditional entropy $H_k(B)$ is a random variable in scheme A: its values are determined by the occurrence of events A_k in scheme A. Therefore, the last right-hand-side term is a mean value of the random variable $H(B)$ in scheme A that is denoted as $H_A(B)$. We then get

$$H(AB) = H(A) + H_A(B) \tag{11.4}$$

In a special case where schemes A and B are independent, equation (11.4) will be identical with equation (11.3). In all cases, the inequality $H_A(B) \leqq H(B)$ holds true. This inequality can be interpreted so that knowledge of the results in scheme A can, on average, decrease the uncertainty in scheme B.

5. The required property of entropy is

$$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n)$$

i.e., addition of an impossible event to the given scheme or addition of an arbitrary number of such events does not change its entropy.

If an experiment is carried out, the results of which are shown in the given scheme, some information is obtained (we know then which of the events A_k has occurred)

and the uncertainty of the given scheme is totally cancelled. Information given by the results of an experiment rely on the abolishment of some uncertainty which existed before the experiment took place. The higher the uncertainty, the higher the value of the information gained by abolishing its uncertainty. As for the measure of uncertainty of a finite scheme A, its entropy $H(A)$ was chosen, and the amount of information gained by removing the uncertainty can be measured by an increasing function of the variable $H(A)$. The choice of this function means the choice of a scale for the amount of information, and, in principle, it is arbitrary.

However, the properties of entropy show that it is usually adavantageous to consider an amount of information proportional to the entropy.

One unit of information or entropy has the form:

$$\begin{Vmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{Vmatrix}$$

For the basis of logarithms equal to 2 it is the amount of information of choice of one from two equally probable possibilities. The entropy of this scheme is equal to

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

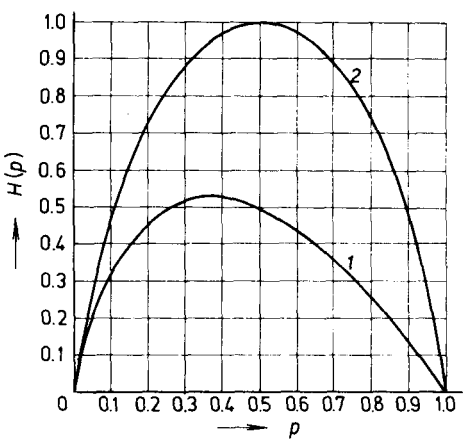

Fig. 11.2 Graph of entropy values

This unit is called *bit* (abbrev. binary digit). The entropy functions $H(p)$ in bits for some basic examples are shown in Fig. 11.2.

For example, entropy of a source that is transmitting a message composed of letters from an alphabet with 32 symbols with an equally probable occurrence is equal

$$H = -\sum_{1}^{32} \frac{1}{2^5}\log_2\left(\frac{1}{2^5}\right) = -\log_2\left(\frac{1}{2^5}\right) = 5$$

Entropy of the source is then 5 bits per 1 symbol. If the source is transmitting, say 40 symbols per minute, then the entropy of the source is 200 bits per minute.

The letters of texts in any one language are not, in fact, independent, their frequencies depend on previous letters of the text. Therefore, the conditional probabilities of each letter can be investigated in a set of experiments, i.e., sample texts; the mean values can be used for the estimation of conditional information (Dupač and Hájek, 1970).

The words in a sentence or in the whole text are not independent. There is some a priori probability of the occurrence of words an word associations. The deviations of observed frequency from this probability are important for the system of automatic indexation (see section 11.6).

Redundancy of information: a scheme in which all elements are equally probable has maximum entropy. Generally, however, the elements have a different probability of occurrence and the entropy of their scheme will be lower. The ratio of the actual entropy H of a given source to the maximum possible entropy $H_{\max}$

$$h = \frac{H}{H_{\max}} \tag{11.5}$$

is called the *relative entropy*. The difference $H_{\max} - H$ is called *inner* (or redundant) *information* and the ratio

$$R = \frac{H_{\max} - H}{H_{\max}} = 1 - h \tag{11.6}$$

is a *coefficient of redundacy* or redundancy of source.

The redundancy of information in languages has been much investigated. It facilitates the abbreviation of texts (in a cable, by omitting some words), or it makes possible the increase of reliability of correct acceptance of a message (to estimate and correct the errors in the text). It is important in coding where an optimal abbreviation of the original message is desirable. Khinchin (1954) has shown that for entropy of a message equal to H the minimum value of the coefficient of abbreviation $1 - R$ is

$$1 - R = \frac{H}{H_{\max}} \tag{11.7}$$

Using this relationship, the lower boundary of message abbreviation can be found and used for optimal coding. The methods of coding used in practice can be a evaluated by comparing them with this optimal coding.

The notion of entropy can be generalized for continuous random variables in relation to their probability distribution types for one-dimensional and k-dimensional distributions (Lapa, 1971).

11.2 BASIC ACTIVITIES AND FUNCTIONS OF INFORMATION SYSTEMS

The basic activities of information systems are derived from their representation as *communication systems* of scientific information starting from the document and ending at its destination: the user of the information. These activities can be classified into three main groups: generation, processing and use of information.

The generation of information involves activities that are necessary if the document is to be published. This phase, comprising contacts with authors, publishers, reviewers, editors, consultants, etc., is not formalized and in the automation of information systems it forms the environment of the system. The processing of information comprises the acquisition, classification, indexing, organisation of documents and their indices into document files and index files; secondary information is treated in processing, i.e., information of information.

Communication with users requires communication links between the source of information and its user via the information channels and information centres that act as switching centres in the process of information retrieval.

The main activity in information processing is information retrieval, based on the content of the document. Such activity is very complex and deals with a large number of items. Therefore, this is the first stage where automation could be used, e.g., automation of selective dissemination of information.

11.3 THE ORGANIZATION AND DEVELOPMENT OF INFORMATION SYSTEMS

The following aspects are taken into account in the organization and development of information systems: a comprehensive plan for automating management information systems, the functions of computers in information systems, data communication and data transmission systems, modern graphic devices, bibliographical document processing, organization of scientific document files, communication problems of relevant information retrieval, etc.

The production of primary information is influenced by the information systems. In a creative process the question arises, as to whether the problem has already been investigated and, if it has, then the necessary information has to be delivered to the user. In view of the information explosion it is becoming progressively more tedious and complicated to find a solution to this basic information problem.

Automated information systems enable information to be processed by *synthetic intelligence*, which use secondary information stored in computer memories and on the selection and retrieval of information by algorithms and sub-systems based on synthetic intelligence.

Information files are based on national reference services, which are interlinked

and form international and world-wide information systems with libraries and their services.

Information centres are a modern branch of information systems, which serve selected groups of users and cater to their requirements.

The development of information systems requires a higher type of centre: analytical information centres which carry out the active acquisition of documents, and produce the criteria for a selective dissemination of information. If these centres are decentralized, then they can promptly react to users' requirements, and provide information feedback.

As water resources systems and water management have an inter-disciplinary character, the selection of documents is a complex problem, and an analytical approach is recommended. Analytical information centres in water management are therefore fully justified.

11.4 THE SYNTHESIS AND ANALYSIS OF INFORMATION SYSTEMS

In the application of the systems approach to the information problem, the starting point is systems analysis, i.e., the analysis of the structure and behaviour of information systems. In this process the whole system is investigated, or this investigation

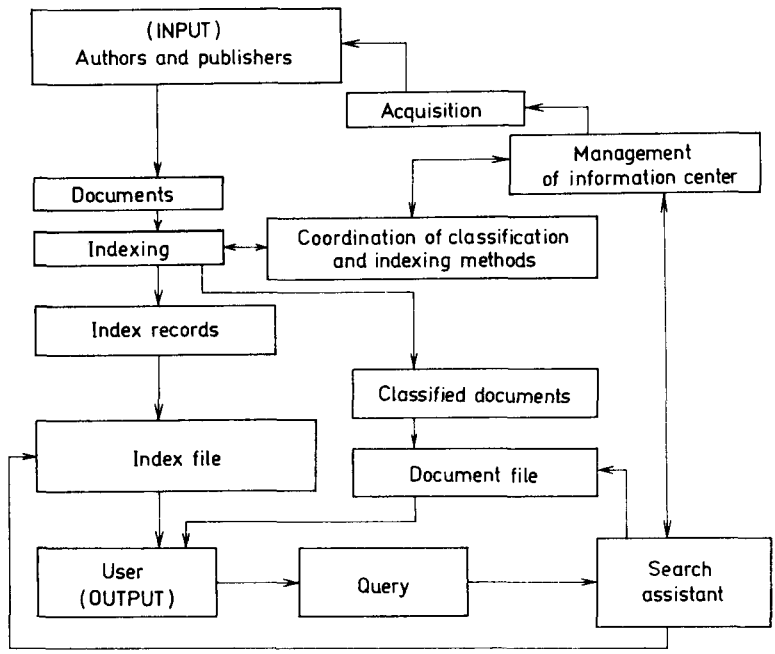

Fig. 11.3 Example of elements of an information system and their interrelations

is focused on the structure or behaviour of the system. If the information system does not yet exist, systems analysis is performed on a simulation model of an information system with the known properties of certain components. Simulation helps to determine the properties of other components and thus the whole structure can be modelled.

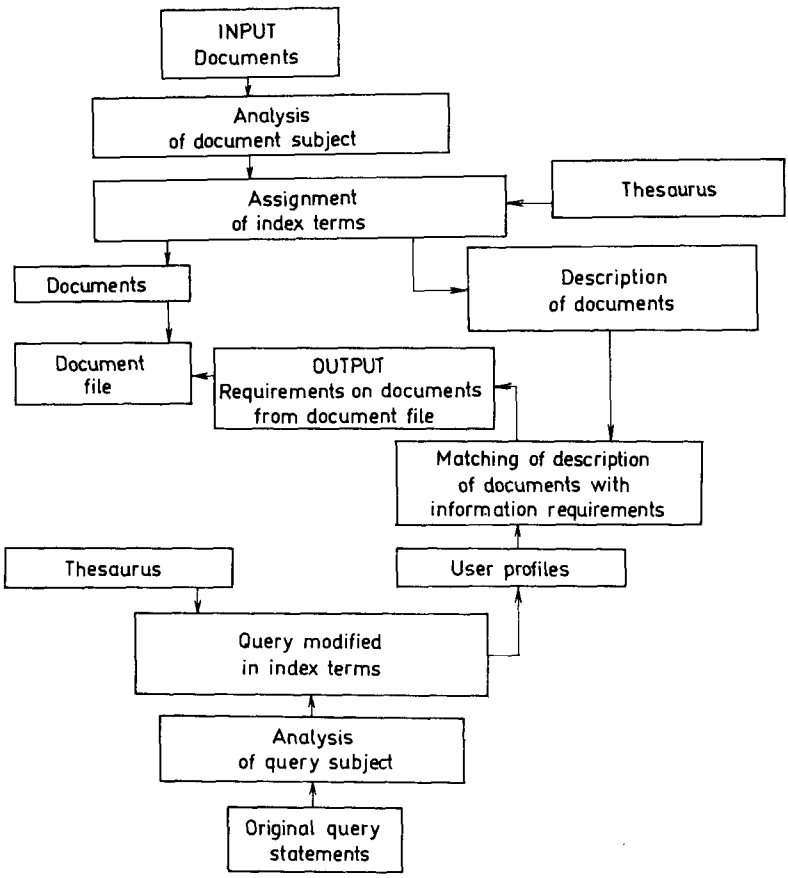

Fig. 11.4 Example of an alternative definition of the system, elements and interrelations for the same information problem as in Fig. 11.3

Using the results of systems analysis, an information system is synthesized with a unified form of the information records, without undesirable duplication. The services should transmit information at the proper time to its correct destination with maximum efficiency, easy interpretation of the information delivered and an information system must be adaptable to changes.

Systems analysis comprises three main steps: the definition of the system, the analysis of the properties of the system and the correction of the system. The definition of an information system includes the choice of elements of the system and corresponding relationships, and the acquisition, organisation and storing of data in a proper form.

Models of recording and data processing algorithms are determined for the elements chosen as shown schematically in Fig. 11.3 for an information retrieval system.

In Figs. 11.3 and 11.4, the systems are defined on a rough discriminating level; they serve as examples of information retrieval and communication of information.

In a more detailed description, the names of the elements are given together with all the important input and output operations, coding, transformations of information in the elements, which are listed together with activities performed in the element, activities for input errors detection and processing, with rules for activities on auxiliary conditions (e.g., if input errors hinder further processing of information), space and time capacities of elements, etc.

Once they have been described, the elements and relationships can be represented graphically in *flow charts*. In information systems, the flow of information plays a decisive role in this mapping of the system. In the automation of information systems, the structure of the system can be represented in matrix form. Systems analysis often reveals that some elements are connected to the system by a very weak link or that the amount of the flow is not important for the objective of the investigation. Such an element can be omitted or, be coupled to another element, or the whole system can be divided into several subsystems which can be investigated separately in further analysis. This disintegration of the system is the main method of investigation of large and complex systems.

Having defined the elements of the system, their interrelationships and relationships to the environment of the system, researchers then attempt to obtain a better definition of the interior of the system, especially of the transformations that take place in flows in and out of the elements. The systems variables and their relationships are investigated. At the same time the consequences of the designed changes are considered in the direction from input to output, and back if feedback is present. The description of the transformations is corrected on the basis of knowledge gained from other systems in operation. In new systems, this is followed by an experimental stage.

In information systems the transformation for each element can hardly formally be described, analytically. Therefore, the description of decisions and operations that take place in each element of the system is an acceptable result, suitable for the simulation of relationships between input and output.

In the *analysis of information systems* the goal is often considered as known that the elements of the system have been determined and so have the decisions and

transformation processes which are to take place in these elements. Then task of the systems analysis is thus the efficient transportation of information to the proper place, by adequate means and the processing of data by suitable system software.

The indications of a malfunction in an information system are: an insufficient quantity of information (e.g., in the retrieval of documents), undesired multiple processing of the same information, contradictions in data in information files, delayed and incorrect information without necessary details, and/or a large amount of information that requires too much time for users' evaluation, etc.

The analysis of existing or simulated systems helps to reveal these drawbacks. There are inner and outer methods of applied systems analysis. In outer tests the method of the "black box" approach is used, i.e., the reactions on input are investigated without analysis of the structure of the system. In the inner test, on the other hand, the structure of the system is investigated from the viewpoint of correctness, complexity, and dynamic character of the transformations in each element of the system. The most effective seems to be a combination of these two methods.

In outer tests, past records describing the inputs and their corresponding outputs are analysed first. Sometimes the isolation of various impacts and the determination of components is difficult or impossible. The *experiments* are performed when only one input variable (or a few input variables) are changed and the corresponding response is investigated.

The outer test can be used for control of inner activities of the system if there is some information concerning the system structure. A special kind of test is the response to exceptional inputs, outside the designated range. If such tests are not performed in advance, then they should be carried out during the experimental operation.

In the investigation of the system by outer test, not all the possible situations can be analysed; outer tests as the single methods of analysis are restricted to cases where the structure of the system is not accessible. The improvement of information systems requires, in most cases, a change in the system structure or a transformation in its elements. This approach is conditioned by inner tests.

The main *inner tests* are the tests of continuity and completeness. For instance in a system there is no relationship between representation of two elements, and this relationship is important, as indicated by inner tests. Then the structure of the system should be corrected. For testing the sequences of information flow in the system, flow charts are suitable and the tests by incidental matrices are applied using a computer.

In information systems that originated by the aggregation of activities previously performed separately and connected in the process of centralization due to computer data processing, inner *tests of compatibility* are important. The elements of the system have compatible relationships if output of one element is a suitable input into another element. Automation of the information process should respect compatibility not

only between elements, but also between subsystems so that a form of output should fit a form of input in further processing. If this condition is not fulfilled, e.g., in the application of the information service on magnetic tapes, a further element must be added to deal with this problem to secure compatibility.

Further tests that serve the correctness and effectiveness of flow of information in a system are called *communication, linkage and transformation tests.* According to the dynamic character of information systems the *loading* of individual elements should be tested. Data concerning the amount of flow, overloading or underloading, the frequency of requirements, their variability in time (continuous or impulse character) are collected and analysed.

The *matrix notation* of information systems structure helps to mechanize these tests. For instance consumptive elements can be picked out (i.e., the elements that receive more information than they transmit), elements that are sources of information, elements with a maximum variability of input and output, elements that are critical for the improvement of information systems, feedback in the system and mainly elements that produce a maximum delay of information.

Testing should start with simple and inexpensive tests and should proceed to more detailed and expensive tests. In processing a large quantity of data or in testing more systems with similar properties the methods of mathematical statistics and the theory of probability can be used.

Testing disturbances in the system is often called *diagnosis of the system.* In the design and building of a new system the knowledge of statistical characteristics helps to estimate the probability of disturbances and to enable proper measures to be taken to prevent the system from failing before it actually does.

The method of *tracing the signals* helps to find void places in information systems. Input standard values are used and the deviations from the standard values are sought throughout the system. For exceptional situations the method of *artificial overloading* is useful, which enables the discovery of the weakest points of information flow in the system.

11.5 THE LANGUAGE OF INFORMATION SYSTEMS

The *information language* is a tool for the description and transformation of information. In automated information systems it is used in the algorithmization of different phases of data processing.

In the description of information the process of abstraction is used mainly, i.e., information is transformed in the direction to greater generality.

Information storing is a component of information processing. In automated systems the form of storing must be suitable for both man and computer. The system should have a dynamic behaviour with a possible modification of information,

including its correction, supplementation, displacement and rearrangement. In information systems some part of the information should serve to retrieve information, which is the main aim of information systems. This part of information makes the greatest impact on the information language.

11.5.1 Characteristic Properties of Information Languages

For information retrieval the collection of documents needs to be in a certain order and there should be some means of searching, matching, and recognizing whether or not the document contains the desired information. To avoid reading the whole document file, a *catalogue* is used, i.e., a list of records with a description of the document content and their location in the file.

Each entry in this catalogue is called the *index record*, all index records form the *index file*. A special language is used in index records and index files. The requirements on this language in automated information systems are as follows:

– the language has to be understood by both man and computer; if it is designed for a broader community of water resources researchers, its syntax and semantics should be simple,

– the use of a computer requires more rigid and unambiguous rules as to format, syntax and vocabulary,

– the format of records should be suitable for their organization into files,

– this language should be apt for development and modification in application.

When automating information systems, the syntax and vocabulary of information languages should be investigated thoroughly. A vocabulary is a set of words that is used in a language. As the index language differs from the natural one, its vocabulary is restricted to words serving the description of documents in index records. The syntax is a set of rules for combining elements of vocabulary into language units, thus producing a meaning not expressible by elements of vocabulary.

The basic requirements on an index language are: expressiveness, unambiguity, compactness and low cost. *Expressiveness* measures the ability of the language to describe the subject. The Universal Decimal Classification is not very expressive, but it is not ambiguous; a natural language, on the contrary, is very expressive. *Unambiguity* requires a control of synonyms and homographs. The numeric code is most *compact*, but it does not meet the requirement of being easily understood by man.

The costs include the preparation of the index language, costs of indexing, training of indexers, maintenance and development of the language, costs of indexing errors, etc.

Examples of index languages are: hierarchical classification (e.g., Universal Decimal Classification), subject headings, and key words. The system of key words can

be fixed or free. The fixed system uses a *thesaurus*, i.e., a dictionary with interrelations in a fixed content vocabulary.

The syntax of index languages in information systems should have the properties of algorithmic languages for their representation on computers. The simplest syntactical rule is the *joining of terms*. The relationship of combined terms is described precisely. The added term narrows the region described by the first term. A higher type of syntax is, for example, filling out of items on a form.

Another example of syntax is the addition of descriptors or key words into terms to describe the content of a document in its proper context. This method helps in a system of key words with a fixed vocabulary to restrict ambiguity.

In the natural language retrieval is difficult. If for instance, the word "reservoir" is sought in the text, the whole text has to be read. In order to facilitate retrieval, different *systems of ordering* were developed. One such system is KWIC (key-word-in-context), where all significant permutations of the document title are alphabetically ordered.

11.5.2 Comparison of Information Languages

Information languages can be compared from different points of view, such as expressiveness, non-ambiguity, compactness, etc. The decisive aspect in the automation of information systems is their *compatibility* with computer algorithms.

A further requirement is their internationality, flexibility for the investigated subject (ability to express different branches of science) and universality (possibility of expressing different alternatives of information requirements).

No single language can fulfil all these requirements completely. Therefore different languages are constructed with properties suited for the main aim of the information system. Comparison of information languages is useful in the choice of the language for information system in WRS.

Information languages should cover the maximum quantity of subjects. The higher their *expressiveness*, the better their ability to meet these requirements. Classification according to expressiveness can be done in the following way:

1. hierarchical classification,
2. subject headings,
3. keywords with fixed thesaurus,
4. keywords with free thesaurus,
5. tagged descriptors,
6. analytic terms,
7. natural language.

Information systems with subject headings and key words with a fixed thesaurus are often used for those regions in which documents are expected to belong. The

flexibility of the thesaurus helps in filling the gap caused by the development of science.

The system of free key words can cover the new subject without difficulties. No synthetic language can provide an ideal degree of precision. Synthetic languages use combinations of descriptors to increase precision of subject description. The number of combinations that can be derived from the basic descriptors is enormous even with a fixed thesaurus. However, the coincidence of user's interpretation and indexer's idea is an issue often investigated.

Information languages are distinguished by their degree of *unambiguity*. By definition the hierarchical classification does not include any synonyms and has no ambiguity. In subject headings and key words with a fixed thesaurus, descriptors with very close meanings can occur. In key words with a free dictionary and further information languages, as discussed, some ambiguity may exist. The control of homograph is almost impossible in these information languages. Ranking information languages according to relative potential ambiguity is the same as ranking according to their expressiveness. Ranking by relative *compactness* is quite the opposite. A hierarchical classification requires fewer symbols, the natural language more of them. The *costs* of information languages, as measured by the selection and use of terms is difficult to determine. The whole process with its complexity should be taken into account, including time requirements of indexing, training of indexers, risk of indexing errors or misinterpretation. Some authors claim that minimum costs are required by information languages using key words and tagged descriptors when hierarchical classification is not necessary, syntax is not complicated and the automation of the information process can be achieved more easily.

11.5.3 Thesaurus

A thesaurus is a dictionary containing words in relation to a precisely defined word content. As defined by UNESCO, it is a controlled dynamic dictionary of terms in semantic and generic relations that covers a certain area of knowledge. This dictionary contains an alphabetically-ordered system of descriptors and indexes of their relationships, a set of expressions with their relationships and rules for their use. The expressions in the thesaurus may be descriptors, uni-terms, key words or expressions from the natural language. It is based on some natural language (e.g., English). The following main types of relationships occur:

– relationships that handle synonyms that specify for each thesaurus entry one or more synonym categories or concept classes (instead of A use X),

– relationships that handle homographs by an additional descriptor; i.e., in the thesaurus several combinations are used (instead of A use AX, or AY or AZ),

– relationship of indication that specifies some different, possibly more accurate descriptor (if A compare X),

– hierarchical classification – if possible, each descriptor is followed by a broader and a narrower term.

A thesaurus is a grouping of words or of word stems into subject categories (concept classes). It controls the synonyms, gives rules for using standard descriptors (it contains terms whose use is prescribed for content analysis purposes) thus serving automatic retrieval systems.

11.6 MODEL OF AN INFORMATION SYSTEM

The objective of information systems is to predict and control its elements and relationships. It is based on the description and explanation of the reality modelled.

The organizational structure of information elements or the types of activities can be chosen as a basis for model classification. According to the organizational structure the following models are referred to:

– models of library activities, publishing activities, retrieval activities, data management, and catalogues and file organization.

The information process as a communication between the source of information, via information systems, to the user is used in classification by types of activities. The following models can then be distinguished:

– models of publishing primary information,
– models of documents acquisition,
– models of secondary information processing,
– models of information retrieval,
– model of selective dissemination of information,
– models of files organization,
– models of data base set up and management.

In the *model of library activities* there are areas where automation is not desirable (e.g., in the physical manipulation of books); this can be taken into account in world libraries, but not in information centres of water management.

The storage of documents by micro-storage techniques is suitable for automation of their handling and duplication. A highly automated system may be achieved by combining these systems with automatic communication systems and the remote transmission of documents with their print-outs by a user, on request.

Automation of further activities is included in the model of secondary information processing. The main problem – automatic classification and indexing – has not yet been solved. Its main assumption – the automatic coding of texts of documents for computers – is in the development stage. The translation of optical signals gained by reading texts, the recognition of individual symbols exists, and it is commercially available for texts on a particular form, but not for the reading of conventional books, journals and reports. This problem will be solved and a multi-purpose reading apparatus with the necessary storage for this information will be developed.

Progress in computer hardware should be accompanied by the necessary software with systems for the comprehension of texts by computer.

The systems of *automatic indexing* are often based on statistical analysis and the frequency of separate terms in text. Some words often occur in certain associations (e.g., in this chapter "information system"). Thus the associations of this term can be investigated as a whole and the frequency of this association can be determined. The terms or term associations are ordered in descending order of their frequencies and those with the highest frequency are used as descriptors in automatic indexing.

Some authors (e.g., Salton, 1968) criticize this method, as frequency can be a bad indicator of word significance. They claim that a word is significant if its frequency is higher than was expected *a priori*. This method requires knowledge of *a priori* probabilities of terms and term associations for each branch.

Other methods of retrieval of significant words and word associations in the text do exist. However, the problem of automatic indexation has not yet been solved. A promising way probably is in the systems approach with the requirement that the authors of primary information should create the basis for indexing. A summary of papers and books helps very much in indexing, some journals publish not only the primary information, but also the secondary information, including the key words. Then the task of transforming these indices into the form required by the information system and the computer is easier, but still has to be undertaken. These problems are also present in information systems for water management.

The systems approach shows that information systems influence the model of primary information that is separated from information systems, but forms its significant environment.

The model of documents acquisition is connected with the selection of documents and therefore with a model of secondary information processing and a model of information retrieval. In these models there is a centre of automation of information systems for the *selective dissemination of information*, retrospective retrieval, thesaurus operations, automatic classification, data base manipulation, etc. The computer with its peripheries is the main tool of automation (see section 11.7).

In the model of library activities in the conventional form, the advertisement activities and preparation and publication of bibliographic quotations can be automated; full automation is possible in activities using modern materials (microfilm, microfiche, etc.). This process is dynamic with step by step mechanization of the system components and their interconnection in the system.

If secondary information is stored on magnetic tape, then the handling of catalogues, indices volumes, glossaries, directories, etc., is possible using contemporary computer hardware.

The development of computers with big centralized systems and also of decentralized *minicomputers* adds further possibilities to information systems. The minicomputer does not require air-conditioning and uses conventional cassettes, floppy

discs, etc., and can be more easily adapted to the user's needs and to the needs of a library. Therefore, besides the centralized information service for water management, decentralized information centres and libraries with close contact with users are progressive forms of information systems.

A further condition of the success of information systems is provided by special purpose equipment such as *copying equipment*, graphic storage devices, printers, etc. The ideal is copying equipment with remote control operated by the user or a group of users, including equipment designed to transform information from one medium such as magnetic tape to another, e.g., microfilm with various kinds of expansion and reduction ratios. However, using the conventional copying equipment the time lag from information requirement to information retrieval can be reduced substantially by the automation of the most critical activities using modern microstorage, copying and conversion devices.

The *model of publishing activities* concerns both primary and secondary information. The automation of primary information concerns printing and editing; creative activity cannot be automated, and experiments with synthetic intelligence are in their initial stages.

Automated information systems help to reduce undesired investigation of the same problem by several researchers and to co-ordinate the methods and forms of publishing. They also help to shorten the time gap between the manuscript and the published book, or journal, by automation of the printing process. If the basis for secondary information (abstracts, key words, etc.) is included in the primary information, then the process of their publication can be fully automated. This procedure can be realized by contemporary computer hardware and equipment and it does exist in some branches to a limited extent.

Models of analytical and research activities cannot be fully automated for all the diverse activities involved. Processing of data on computers is an integral part of these activities; for example: *relevance* judgements, precision determination, questionnaire evaluation, computation of economic effectiveness of systems of selective dissemination of information with various indexing systems and the determination of their advantages for selective dissemination and retrospective retrieval, computation of costs per user, per one profile, per one descriptor, per one relevant document, evaluation of various policies in retrieval based on magnetic tape services, tests of data base organization, estimation of operation of on-line and off-line systems, etc.

Information systems are an important part of *cybernetic control systems* and management information systems. The modelling technique is a basic method of cybernetics. In the cybernetic concept the main stress is given to feedback in models and between models, application of the theory of algorithms for models, especially in quantitative expression and coding for computers, the application of certain parts of automata theory (e.g., finite automata), heuristic programming and approaches

to synthetic intelligence, the theory of games and its application to information systems, etc. In the future development of information systems other methods of cybernetics can be applied, e.g., investigation of self organizing systems, systems with automatic pattern recognition using optical or magnetic reading techniques. Self-organization and self-control can be the progressive means of automatic software generation. The process of learning common to living organism's can be utilized for cybernetic automata in information systems, e.g., in automatic indexing, translation from one language to another, etc.

The application of cybernetics (with the exception of systems approach, the use of computers, and modelling) has been used more intensively in control and regulation of technical systems and its use in information systems in water management is in the developmental stage.

11.7 COMPUTER AS THE MAIN INSTRUMENT OF INFORMATION SYSTEMS AUTOMATION

The development of computers made their application to information systems possible. In the automation of complex systems the computer is the main, but not the only tool of automation. Computerized information systems are used for *selective dissemination of information* and information retrieval when the systems are based on magnetic tape services, for the automation of set-up of lists, catalogues, bibliographical references, etc. In these activities problems can occur more in capacities for assembling and coding of input data and necessary software than in methodological issues and computer hardware.

The function of computers in systems of selective dissemination of information and information retrieval is the basic function of modern information systems.

11.8 INPUT OF INFORMATION SYSTEMS IN WATER MANAGEMENT

International co-operation in scientific, technical and economic information, together with the co-ordination of their processing, is the main prerequisite of effective information systems. A basic condition of an effective information system is its data base. The automation of information systems in water management requires the set-up of a data base, its maintenance and replacement. This system is to be co-ordinated with information systems of other branches and with international information systems.

For WRS primary, as well as secondary information is important, especially hydrological data and data concerning the basin where the WRS is located. The data base of such an information system is often developed together with some big

project and water plan. An important requirement in this activity is the interface between their information systems and the compatibility of data structure.

Data structure elements are called *bits* and are often organized into bytes (e.g., IBM) or words (e.g., ICL) that can be used as representation of *characters*. A *field* is built up from characters, and a *record* from fields. The field is a set of characters with a defined meaning, a record is a set of fields; the records are descriptive of individuals, the file is a class of individuals. Grouping of records in a file is done according to some organisational structure.

The representation of a character by bits depends on the type of computer. Therefore, the basic level for the information system is the character. In data structure on magnetic tapes the item of structure is one block, on a disc the item is a cylinder. This is a physical structure. It is advantageous if this physical structure can be used as a subclass of the file structure of information systems.

The arrangement of fields in relations in a record, the arrangement of records in a file, files in set of files, etc., is called *data structure*. Data structure can be defined in different ways. The requirements of data structure are similar to the requirements of the index language, as they both serve the same purpose – adequate information storage and retrieval.

The expressiveness of the record structure is the ability to describe information using this structure. If the structure is too simple, it hinders expression of complex information. If some information is sought for in a record, the field of descriptors in this record has to be found. If some difficulties occur in this procedure, the process of retrieval is delayed and if some ambiguity is present, the results may be incorrect. Therefore positional ambiguity must be excluded as far as possible.

Records with high compactness use a relatively small number of characters for coding information. Requirements for storage are then smaller, which is economically better. However, a high level of compactness makes the retrieval procedure more difficult.

Forms of structure: invariant structure, bit coding, (yes-no questions), fixed and repeated fields, tagged fields, adressed fields, and the structure of the natural language.

Invariant structure does not provide any information, and for a given file, it is used for the identification of this file from the rest.

Bit coding is used in systems of fixed fields for identification, to ascertain whether or not a particular index is present. In *fixed fields structure* the position and length of each field is given. However, the content of a field can change. If the length of the content is highly variable, this structure does not make full use of the storage in computers. If the form of the fields and their sequences are given, but their numbers change, this structure is called a *structure of repeated fields*. The *structure of addressed fields* (that are more economical as their length is variable) is organized into two groups of fields – descriptive fields and fields of adresses that delineate the limits of descriptive fields, or their length and the type of their content (heading of fields).

In the application of these structures combinations are preferred which make full use of their advantages. These advantages are derived from their properties such as expresiveness, compactness, etc. For example, the expressiveness of these structures in ascending order is as follows: invariant structure, bit coding, structure of fixed fields, structure of repeated fields, structure of adressed fields and structure of the natural language. In this order the ambiguity, both positional and semantic, grows. Compactness is highest in the structure of fixed and repeated fields, other structures are less compact.

Processing of a file is, in fact, its transformation and involves a search for the parts to be processed. *Searching* is the process of isolating one particular area of the file. The result of this search may be a statement as to whether a given information area is present or not. The processing of files depends on their structure, which is related to their physical storage in the memories of computers.

According to how accessible stored information is, the memories of computers may be classified into *equal*, *direct* (random), and *sequential access memories*. The inner memory of a computer has equal access as every addressable cell of the computer has the same access time. Magnetic discs (and drums) have direct (random) access, other storage devices such as magnetic tape, paper tape, punched cards, etc., have sequential memory access.

The most advantageous is the inner memory that influences the structure of records less. Its capacity is, however, limited and the security of information is low (e.g., it may be destroyed by interruption). Sequential organisation, which is necessary on magnetic tape, is useful in sequential processing only. Duplicating or triplicating of files, which is relatively cheap on magnetic tapes, increases the security of stored information. Magnetic discs provide a lower access time, its security, however, is more complicated. The storage medium is influenced by the frequency of file processing.

The file *structure* is an important factor in the retrieval of information. The simplest organization of records in a file is the *sequential structure* where the records follow each other logically and physically. The addition of records inside the file is, however, more difficult and is not effective.

In further structures the logical and physical structures are different. In a *random structure* the physical structure is random and the ordering principle is given by addresses. In a *chain structure* these addresses are contained in records. A *branching structure* is an extension of the chaining concept. In a *list structure* (or directory structure) addresses are physically separated in a directory. The list structure is the most advantageous for retrieval, where direct access storage devices are necessary.

11.9 EFFECTIVENESS CRITERIA FOR THE PERFORMANCE OF INFORMATION SYSTEMS

Testing is the best way to evaluate the effectiveness of information systems. Its main components should be: the determination of test criteria, design of the test, performance of the test program, results analysis, and design of systems innovation or amendments based on this analysis.

The criteria for the evaluation of systems operations should be contained in answer to the following questions: Are the users satisfied with the system? Does the contemporary system of indexing encompass the main branch problems? Does the delay between publishing the primary document and getting the information to the user influence the performace of the system and diminish its value? Could the user be informed sooner about this document by another means? Are there any substantial differences in indexing methods between the files? Are the descriptors specific enough? Is the applied method of relevance evaluation appropriate? Are the terms of the thesaurus adequate? What are the requirements of users concerning the indices of precision and recall? Are there policies for their increase and can these policies enter the system? How are the outputs of information systems and other services utilized by users? What methods are used by users to increase the indices of precision and recall? Can these methods be automated and used in the information system? What are the most effective methods of interaction between users and the information system? Does the necessity for this interaction influence the system? Do input data of the system (mainly magnetic tape services) include errors, if so what kind of errors? Can the system handle these errors? Are the programs in information systems software flexible enough for possible changes?

The user is interested in *high relevance* and a *short time lag.* The basic problem of the selective dissemination of information and retrospective retrieval is the high degree of relevance. Relevance is often evaluated by *precision and recall indices.* There is an approximately indirect relationship between them. The precision index (PI) is the proportion of relevant documents to retrieved documents, the recall index (RI) is the proportion of relevant documents retrieved by the system to all relevant documents in the file. (For instance 20 documents in the file are relevant. If 10 documents are retrieved and 8 of them are actually relevant, the precision index $PI = 8/10 = 0.8$ and recall index $RI = 8/20 = 0.4$.) If no section is performed and all the documents are retrieved, the recall index will equal one and the precision index will be almost zero. If the retrieval is made by strict requirements a few documents will be retrieved and possibly all the documents will be relevant ($PI = 1.0$), but a relatively high number of relevant documents will not be present in the retrieved part, and the recall index will be low.

In the design of tests for systems evaluation more users should be involved as their evaluation of relevance can differ from that applied in systems software. Some

aspects of systems operation policy may be geared to satisfy the users' requirements. The system operation policy is, however, based on the requirements of an average user, i.e., it tries to satisfy the majority of users, but not all of them. It would not be efficient to complicate the system for the specific needs of one, or several users.

Further aspects of systems operation policy are revealed by comparison of *operational effectiveness* and *overall economic efficiency*. Some means serve to increase both forms of efficiency, e.g., automation of information systems. However, in some situations the effort to increase both forms leads to competitive actions and some optimization is necessary to determine the optimal degree of satisfaction of these aspects. This phenomenon can be observed in system hardware elements and in software too, e.g., in indexing methods, information languages, retrieval policy, processing of outputs, methods of their delivery to users, etc.

11.10 SOFTWARE OF INFORMATION SYSTEMS

The software of information systems should be computer-independent. This requirement is, however, seldom fulfilled. Therefore, the software of information systems will differ for individual computers. This problem can be solved by the unification of hardware with centralization of information systems. Beside these tendencies to centralization, the development of information systems with *minicomputers* leads to decentralized systems for special purposes. The software for these systems is *often supplied by the producer, and it uses an interactive mode.*

Two basic types of information systems can be distinguished – integrated (centralized) systems and distributed (decentralized) systems. If the same data base is used by several subsystems, the integrated information system is often preferred. The key component is the common data base. The basic characteristic are fast response to queries via remote terminals, on line mass storage, continuous up-dating of files, centralized data processing. An integrated information system reduces redundancy, secures more protection of data, allows more than one user to retrieve documents, increases the overall effectiveness due to provision of more timely, relevant and accurate information, reacts better to long-term planning. The integrated system has some disadvantages – special personnel requirements (system analysts), co-operation of subsystems, lower responsiveness to user's needs, the breakdown of the information system may have catastrophic results, high development costs, difficult modifications.

The distributed information system, on the other hand, does not use the method of the universal system (that leads to low efficiency in software overheads and administration), but a modular system. An alternative to the common data base in a distributed system is an aggregation of information systems arranged in such a manner that a set of subsystems is formed and tied together by means of a communication interface. To develop the data base is difficult but not so expensive,

and this data base is more suitable to the needs of users. The cost effectiveness with minicomputers is often higher in distributed systems, the overall costs are lower, and the information system can be built in units that are able to function separately.

The system can easily be modified to meet users' requirements. Recovery and control are easily handled, simple software is necessary, the errors in input data are processed in their place of origin, the breakdown of the subsystem can be overcome, new subsystem can be added.

On the other hand, more co-operation is necessary, some redundancy occurs, more communication channels are necessary, the retrieval of data from several subsystems is more complicated, users can gain access easily on one subsystem only, etc.

The effectiveness of computer-based information systems can be measured by the cost-benefit ratio as in other branches of technology. In information systems not only the direct costs and benefits, but also the indirect costs should be considered. The direct benefits involve time reduction in tasks, the direct costs involve costs of computer purchase or rental, costs of computer operation (payroll of employees, costs of supplies, maintenance, power, insurance, etc.).

In information systems many benefits are of intangible nature that can hardly be evaluated in monetary units and stem from better information acquaintance and knowledge of users. It is also important to liberate people from routine work. Indirect costs also include costs of information system reorganisation, personnel training, and transition costs.

The increase of the effectiveness of computer use in information systems is concentrated on its hardware, software, systems approach to co-ordination of software, input data, data base, use of output and actions that are related to this output. The systems approach increases effectiveness by the unification of input and output, thus creating conditions for unique interface.

The software of information systems should be set up on the basis of the project of the information system, containing a preliminary or feasibility study, analysis of the existing system (or simulated system), and the design of the system, system innovation or system amendments. The feasibility study has to state the conditions for further information processing, form a team for project elaboration and find the budgetary constraints. The problems of analysis have already been discussed. The design of the system should be elaborated in alternatives with economic analysis to an extent to satisfy approval by decision-makers.

The computer is the main component of information systems. Therefore the alternatives differ according to which choice is made. For the designated type of computer and the necessary software the design should specify the main activities in the system, the possibility of their integration, a rough flowchart of data processing by computer, define the files and their organisation, estimate the cost and time of their processing and transformation, design the content and form of output, identify the individual runs of data processing, estimate the number of operations in the sub-

systems of information systems, the periodicity of these runs and operations, list the standards that should be respected, prepare specification for programmers, define methods for data conversion and preparation, etc.

Some of the methods of information system analysis quoted are used not only in system design, but also in the evaluation of its operation. An information system in operation is not a static object. The input data change in content, quantity and form. After exceeding certain limits the combined effect of these changes requires a modification of the system. Further, the requirements of users change with the development of science, technology and their experience with the system. The fast development of computer hardware and software influences information systems by providing new possibilities.