In SYSTEMEERING 75, Festskrift till Børje Langefors. Eds. J. Bubenko and M. Lundeberg. Studentlitteratur. Lund 1975.

On the Design of Statistical Information Systems

SVEIN NORDBOTTEN

ABSTRACT

The aim of this note is to discuss some principles for design of statistical information systems. Section 1 introduces the concept of a national statistical information system. In section 2 the role of this system within the nation – considered as a cybernetic system – is discussed. The formalization of a design method is equivalent to the development of a designing system some components of which are discussed in section 3. These components are a subsystem for the specification of the class of alternative designs, the subsystem for selecting one design and a subsystem for compiling the information required by the designing system. In the last section conclusions are discussed.

1. STATISTICAL INFORMATION SYSTEMS

By statistics we mean information about the properties of groups in contrast to information about individual objects forming the groups. A statistical information system is an information system designed to produce statistics, and in this note we are considering nation-wide systems the aim of which is to produce statistics about the state and development of the nation for the use of government decision makers.

No government would be able to identify tasks, develop plans for alternative actions and select one plan for promotion without some statistical information about the nation, the ways it operates and reacts to different actions. There are also other important reasons for supporting a statistical information system for the nation, e.g. the needs of the business community for statistics, the educational needs and the general needs of the public to be kept informed about the activities of the government and their effects on the nation.

The traditional way of producing statistics was to consider parts of the nation as independent systems and to collect and compute statistics for each part independently. Few possibilities were left for future use of the collected data. The increasing needs for detailed statistics and statistical models as well as the introduction of the computer in the production of statistics promoted the idea of the integrated, statistical archive system about 1960 (4). The statistical

archive system, which in the terminology of the 1970's would be called a data base system, differed from the traditional system in several respect one of the most important of which was the addition of a data archive into which all collected data should be organized and stored, and from which data for statistical computations were retrieved when needed. The main concepts of the data archive organization were object identification, attribute code and value, and time specification (5).

The purpose of this note is to discuss the design of a statistical archive system as part of the system representing the nation. The discussion is limited to the infological aspects of the design, in the sense of Sundgren (6), i.e. the datalogical problems in connection with the representation of information sets and transformations are not within the scope of the present note.

2. THE NATION CONSIDERED AS A SYSTEM

We start our discussion by an investigation of the role a statistical information system plays in a nation considered as a cybernetic system (1). Figure 1 indicates the external properties of this system which is supposed to be observable at dicrete points of time indicated by the time index t. The design point of time is denoted by the index value t=0. At time t, the state of the nation is supposed to be described by variables represented as elements in the time-specified vector, r(t).

The system receives inputs from three sources considered as external to the system. Inputs from the first source are the goals for the nation which the system tries to realize and which are symbolized by elements of a goal vector, g. We will not engage ourselves in the explanation of how the goals are set and consider them therefore as exogenous. The second input source is the designer who has the power to introduce a new, or change an old, structure of certain operating parts of the system. The design specifications are represented by the elements of a design parameter vector, d, associated with the space D of all possible designs. One of the main purposes of these present lines is to consider the task of specifying the space D and select one specific design vector (2). The third source of input represents all the other external factors which affect the system, but which are not explained by the system. The effects of these factors are represented as elements in a vector e.

We will assume that the behaviour of the system is explained by the well-known state functions:

(2.1) $\quad r(t) = f(r(t-1), e(t-1); g, d)$

where f denotes a set of functional forms one for each element of r. The state of the nation at any point of time may thus be considered as a trans-

formation of its state and external inputs at the previous point of time given the goal and design parameters. This assumption seems to comprise current theories within social science and economics.

Let us take a look into the system. Figure 2 indicates a decomposition of the nation into four subsystems convenient for the present discussion. The material system, MS, comprises those objects which represent the aspects or properties of the nation reflected by the variables in the state vector, r. To the extent the properties associated with other subsystems also are relevant elements of r, e.g. their material resource utilization, these properties are supposed to be transmitted through the structure to MS and exposed in r. The effects of the external factors, represented by the vector e are supposed to act directly on MS and disturb its equilibrium.

The nation has a government which, according to the national goals, supplies MS with inputs represented by the vector x. The government has two main tasks: 1) to develop general directives for the treatment of objects in specified groups, and 2) to apply these directives in operative decisions concerning individual objects identified as belonging to the groups for which the directives are set. Examples of general directives are programmes, budgets, laws, etc., and the application of these directives means carrying out the programmes, disposing the resources and enforcing the laws. We will recognize a particular government subsystem for each of these two types of functions.

The operative government system, OGS, has two sets of input variables represented by the vectors r and z. The input vector r represents the potential information available to OGS about MS. How many and which of the variables of r are utilized, is determined partly by the ability of the subsystem to make individual observations and partly of the second input vector z which represents the directives given by the second government system. The output vector x is the result of the transformations performed by the subsystem on the two input vectors, and it represents the government's operative inputs to MS. The OGS comprises government activities such as educational activities, social assistance, hospital services, communication services, tax authorities as well as government information systems for administrative purposes such as population registers, information systems for taxation and social security information systems.

The aim of the second government system, the directive government system, DGS, is to produce general directives for actions on groups of objects in MS independent of the individual identities of the objects belonging to the groups. DGS has two kinds of input, the goal vector, g, and the available statistical information about the groups of objects in MS. Which of the available statistics are utilized by DGS, will depend on the state of the subsystem and the

goal parameter vector. The directives produced by DGS is represented by the already introduced vector z which is also an input vector to OGS.

The fourth subsystem considered is the statistical information system, SIS. The task of this subsystem is to make statistical information available for DGS. There are two sets of inputs to this subsystem. One is the state vector, r, of MS and the other is the design parameter vector, d. How much of the r vector is recorded will depend on the state of SIS and the vector d specifying the operational structure of SIS. The output of SIS is the different statistical products available to DGS, each of which is represented by an element of the output vector s. The SIS may of course also be considered as a third government system.

We can summarize properties and structures of all subsystems by the functions:

$$
\begin{aligned}
(2.2) \qquad r(t) &= f_1(r(t-1), x(t-1), e(t-1)) \\
x(t) &= f_2(x(t-1), r(t-1), z(t-1)) \\
z(t) &= f_3(z(t-1), s(t-1), g) \\
s(t) &= f_4(s(t-1), r(t-1), d)
\end{aligned}
$$

where f_1, f_2, f_3, and f_4 are sets of functional forms with one function for each element of the left-hand side vectors. The structure emphazises the assumed dynamic properties of the system, and that our design interest is concentrated to the statistical information system.

The behaviour of the national system for an interval from t=O to t=T, called the system trajectory, is denoted by the sequence vector, h=(r(O),. . . . . , r(T)), associated with the space H comprising all possible trajectories consistent with the system functions (2.2). Some trajectories are preferred compared with others. We assume the existence of some performance evaluation function for the system expressing how well the system satisfies its goals, i.e. we assume the existence of some function:

$$(2.3) \qquad w = w(h,g)$$

defined for all pairs of elements from the spaces H and G mapping the pairs to elements of an evaluation space W in such a way that one pair associated with a higher valued element of W is preferred to an other associated with a lower valued element.

Given a set of goals, the performance evaluation function measures the performance of the system expressed by its trajectory. One possible way of improving performance will be to redesign the SIS to produce more adequate information for DGS. The task of the designer of SIS will therefore be to select a design which will contribute to a future trajectory which combined with the goals is highly evaluated.

3. DESIGN OF STATISTICAL SYSTEMS

Because design of statistical information systems is a process to be repeated, we believe that there is a need of a formalized method for design of such a system. To formalize the principles for design may, however, be considered equivalent to the construction and inclusion of a designing system, DS, in our cybernetic system. Figure 3 indicates how DS can be coupled to the system we are considering in this note. The designing system will compile the information it requires from SIS and produce as its result a new SIS design represented by the design parameter vector d.

The designing system may, as figure 4 indicates, be considered as composed by four subsystems, one for compiling the necessary information from SIS, the second for specifying the space D of alternative feasible designs, the third for selecting one design from D, and the fourth for returning the selected design to SIS as a recommended design vector, d.

3.1 The framework of alternative designs

The framework of alternative designs is a model describing the different ways SIS may work. Figure 5 outlines the main activities and components in SIS, and represents the skeleton on which the alternative designs are specified.

SIS includes two main sets of information components, the archive of collected data and the archive of computed, available statistics. We can regard the collection of data from MS as an investment in the data archive. The data archive, similar to production capital, can be used without being consumed in the subsequent activity called computation of statistics. This computation activity can also be regarded as an investment activity in the second archive, the statistical archive, which contains statistical products like statistical tables, time series, demografic, social and economic models, etc. available to DGS (6).

The activities indicated are time-consuming, and the time interval between the collection of data from MS and the time at which the products are utilized, may be substantial. One main task for the designing system will therefore be to allocate the activities at the time axis in such a way that the purpose of SIS is fulfilled as far as possible. When we in the following are discussing the different components and activities of SIS, it is important to remember that we are discussing information sets and information transformations, not their representations or associated algorithms.

We introduce the following notation:

$q_j^I(t)$: computation of the statistical archive component j in period t-1 to t,

$q_k^{II}(t)$: collection of the data archive component k in period t-1 to t,

$Q_j^I(t)$: statistical archive component j in the archive at time t,

$Q_k^{II}(t)$: data archive component k in the archive at time t.

for $j=1, \ldots\ldots\ldots, J$
$k=1, \ldots\ldots\ldots, K$
$t=0, \ldots\ldots\ldots, T$

The activity variables, $q_j^I(t)$ and $q_k^{II}(t)$, are all defined as binary variables the values 1 and 0 denoting present or absent activity, respectively. The component variables, $Q_j^I(t)$ and $Q_k^{II}(t)$, are defined similarly, the values 1 and 0 representing the presence or absence of the corresponding component at time t. These binary assumptions can be summed up by the conditions:

(3.1) $\left.\begin{matrix} q_j^I(t) \\ q_k^{II}(t) \\ Q_j^I(t) \\ Q_k^{II}(t) \end{matrix}\right\} = 1 \text{ or } 0$

The component variables and the corresponding activity variables are related since the former must be produced by the latter. We express the relations by:

(3.2) $$Q_j^I(t) = Q_j^I(t-1) + q_j^I(t)$$
$$Q_k^{II}(t) = Q_k^{II}(t-1) + q_k^{II}(t)$$

The relations imply that the components will be present from the end of the period during which they were produced and when they have once been produced, they never will vanish. For an information system as SIS, this may not be an unrealistic assumption. The relations (3.2) also imply that an activity can never be repeated and that:

$$Q_j^I(t) \geqq q_j^I(t)$$
$$Q_k^{II}(t) \geqq q_k^{II}(t)$$

Computing a statistical product may require the existence of several components in the data archive. We assume that the data components have to be present at the end of the period before the computing period. The requirements of the statistical computations to the data archive are represented by

a binary requirement matrix A the element a_{jk} of which represents the requirement of the computing activity j to archive component k. Requirement for the component is represented by the value 1, independence of component by value 0.

The requirement matrix A permits us now to formalize the conditions:

$$(3.3) \qquad a_{jk}(q_j^{I}(t) - Q_k^{II}(t-1)) \leqq 0$$

which say that no computation activity can be started if not all required information components are present.

The designer will frequently identify a number of different activities which result in substitutable information components. There is no need for producing several substitutable components, and we take care of this condition by means of the alternative matrices B^{I} and B^{II}. The element $b_{jm}^{I} = 1$ indicates that the computing activity j may be substituted by any other activity indicated by value 1 in column m of matrix B^{I}. Similarly, all the elements b_{kn}^{II} of column n in matrix B^{II} represented by value 1, indicate that the corresponding data collection activities can be interchanged.

The alternative matrices are used for formulation of the non-duplication conditions expressed by:

$$(3.4) \qquad \sum_t \sum_j b_{jm}^{I} \; q_j^{I}(t) \leqq 1$$

$$\sum_t \sum_k b_{kn}^{II} \; q_k^{II}(t) \leqq 1$$

for

$$m = 1, \ldots\ldots M \leqq J$$

$$n = 1, \ldots\ldots N \leqq K$$

We regard the variables $q_j^{I}(t)$ and $q_k^{II}(t)$ as the binary elements of the parameter vector, d, i.e. the design task is to decide which activities should be carried out and in which periods. The assumptions represented by the restrictions (3.1) – (3.4) combined with the initial information components $Q_j^{I}(0)$ and $Q_k^{II}(0)$ present at the design point of time delimit the class of feasible design parameter vectors. The activities decided to be carried out in the design are indicated by variables values 1 in the design parameter vector.

3.2 The design selection rule

The rule for selecting one design from the space D of feasible designs, should be constructed in such a way that it is consistent with the performance evaluation function (2.3). Given the values of the initial variables, we see that the trajectory, h, determined by the system (2.2), can be regarded as a function:

(3.5) $h = h\,(d, g, e(0), \ldots\ldots e(T\text{-}1))$

of the design and goal parameter vectors and the external factor inputs. The function form represents the system structure and subsystem functions.

To be able to specify a design selection rule, we need obviously knowledge about the goal vector, the effects of external factors and the form h in addition to the performance evaluation function.

Considering the goal vector as given and invariant in our context, two problems must be discussed. The first is due to the fact that the future effects of the external factors is unknown. Since we have no subsystem explaining these effects, we are also unable to make any definite prediction about their values. We may, however, assume that the effects can be considered to be realizations of a random process. To assume that the variables of the vector e are random variables, implies, however, that also h, through the function (3.5), and w, through (2.3), must be random variables.

We denote the probability function of the random variable w by:

(3.6) $p(W_i,D_j) = Pr(w = W_i \mid d = D_j)$

where $p(W_i, D_j)$ is the probability for the realization of a trajectory associated with an evaluation value W_i of space W if we select a design represented by the element D_j of the design vector space D.

To each design there may in other words, correspond several trajectories with different probabilities to be realized and different evaluation values. We define the expected evaluation value associated with a specified design as:

(3.7) $E\,w(D_j) = \Sigma_i W_i \cdot p(W_i, D_j)$

Remembering that the design vector consists of binary variables, the expected value must be a series of additive terms the first set of which contains single binary component and activity variables, the second set contains products of pairs of such variables, the third set contains products of triples of such variables, and so on. Considering only the first set of terms the function will be:

(3.8) $E\,w = \nu_0 + \Sigma_t\,(\Sigma_j \nu_{jt} \cdot Q_j'(t) - (\Sigma_j\, C'_{jt} \cdot q'_j(t) + \Sigma_k C''_{kt} \cdot q''_k)))$

where V_{it} may be interpreted as the expected positive contribution to the evaluation value from having the statistical product i available at time t. C'_{jt} and C''_{kt} may be interpreted as the expected costs associated with the activities $q'_j(t)$ and $q''_k(t)$, respectively, and which contribute negatively to the performance variable value because they are ressource consuming. The costs are assumed to be measured in units compatible to those expressing the performance value. We could also easily have introduced storage cost for the information components in (3.8).

Considering the elements of the evaluation matrices V, C' and C'' as expected values, imply that we must estimate the probabilities associated with the different values the corresponding random variables may take. The more knowledge we already have about the system (2.2) at the design point of time, the more precise estimates of the coefficients can be computed.

The next problem to be considered is the choice of selection strategy. Several strategies for forming a selection rule are available. Which to apply is a question to be answered outside the designing system considered in this note. One strategy may be to form a rule which says that we select that design vector which gives the highest expected evaluation variable value applying (3.8). In principle, we then have to compute the expected evaluation variable value for each design and recommend the one having the highest value.

3.3 Some other aspects of the design system

From the two previous sections we know that a designing system will need information for carrying out its task. The required information can be summed up as:

- A list of all potential activities of SIS.
- The requirement matrix A.
- The alternative matrices B' and B''.
- The initial information component vectors $Q'(0)$ and $Q''(0)$.
- The evaluation matrices, V, C' and C''.
- A starting stimulus.

We assume that the information must be retrieved form the existing SIS, and since the state of SIS is completely described by the information indicated by the initial information components, these are the source which is available. How to utilize the existing information in forming the above information required by the designing system is a question which has to be tackled and answered by estimation methods.

The information indicated by the last line of the above list has not so far been

mentioned. Since we have considered the designing system as a subsystem of the nation system, we will also be interested in how this system is activated. The designing system may for instance be started by a stimulus from SIS which might be considered as a call from SIS to DS for assistance. This implies that SIS contains a component which currently evaluate past trajectories and design parameter vectors by means of (2.3). When the result of the evaluation gives a value below a certain level, i.e. when the performance becomes too poor to be accepted, a call for assistance is issued to DS which retrieves the necessary and available information, forms the class of feasible designs, selects a design and recommends this design to SIS for implementation. Depending on the acceptance level and the effects of external factors, the design process may be periodical or continuous.

4. CONCLUDING REMARKS

The purpose of this note was to investigate some problems in connection with the design of nation-wide statistical information systems. An ultimate aim for work on design problems is to obtain general principles formalized as a system in order to leave some of the future work to less trained people or to automata.

The basis for our work on a designing system must be the knowledge of the role to be played by the information system in its environment, and one important task for a nation-wide statistical information system will be to supply the government system with the information needed to form general decisions for regulating the activities of the objects forming the nation.

Two major components of a designing system will be the component forming the class of feasible designs, and the component selecting one of these for recommended implementation. The designing system may be considered as a system component which is periodically active for improving the performance of the total system.

The points of view exposed in this note may probably be valid for the task of designing information systems in general. They may also be interpreted in the terms of the information system theory of Langefors (3). The role played by the information system in its environment and the necessity for analyzing this role has been emphasized by Langefors, who also argues for the need of a formal theory for information system design. There is a obvious resemblance between forming the class of feasible designs and selection of one design to be recommended as outlined in this note, and the two tasks described as information analysis and system computation by Langefors. Our requirement matrix is very similar to the incidence matrix and may easily be transformed to a precedence matrix. Langefors considers the task of selecting a system design under the heading system design computation. Based on a class of feasible de-

signs derivable from a given incidence matrix and specified hardware restrictions, Langefors discusses the problem of finding that design of process groupings and file consolidations which minimizes costs represented by the data transport. In our presentation, the alternative process groupings and file consolidations are represented by alternative activity matrices while the hardware restrictions and cost minimizing are represented by the evaluation matrices and the selection rule strategy, respectively.

REFERENCES

1. ASHBY, W.R., *An Introduction to Cybernetics,* University Paperbacks, London 1971.
2. CHURCHMAN, C.W., *The Design of Inquiring Systems: Basic Concepts of Systems and Organization,* Basic Books Inc., N.Y. 1971.
3. LANGEFORS, B., *Theoretical Analysis of Information Systems,* Studentlitteratur & Auerbach, Lund 1973.
4. NORDBOTTEN, S., *Elektronmaskinene og statistikkens framtidige utforming,* Nordisk Statistisk Skriftserie, No. 7, Helsingfors 1961.
5. NORDBOTTEN, S., *Purposes, Problems and Ideas Related to Statistical File Systems,* Proceedings from the 36 session of the International Statistical Institute, Sydney. Reprinted as Artikler No. 40, Statistisk Sentralbyrå, Oslo 1971.
6. SUNDGREN, B., *An Infological Approach to Data Bases,* Urval No. 7, Statistiska Centralbyrån, Stockholm 1973.

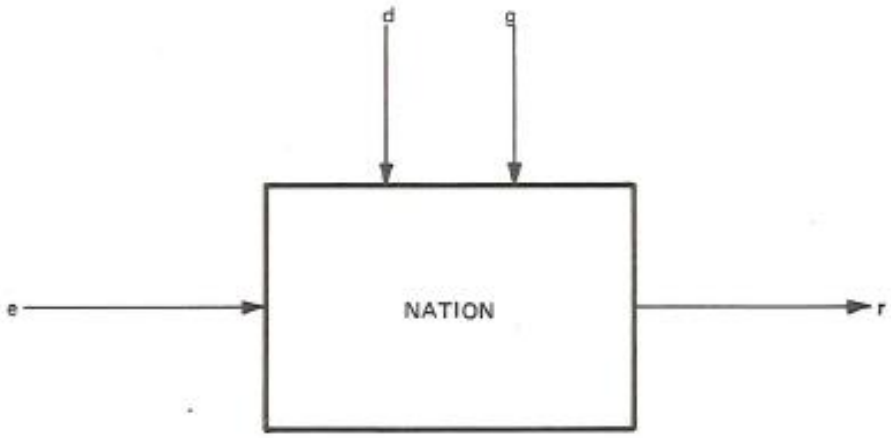

Figure 1: The nation as a system

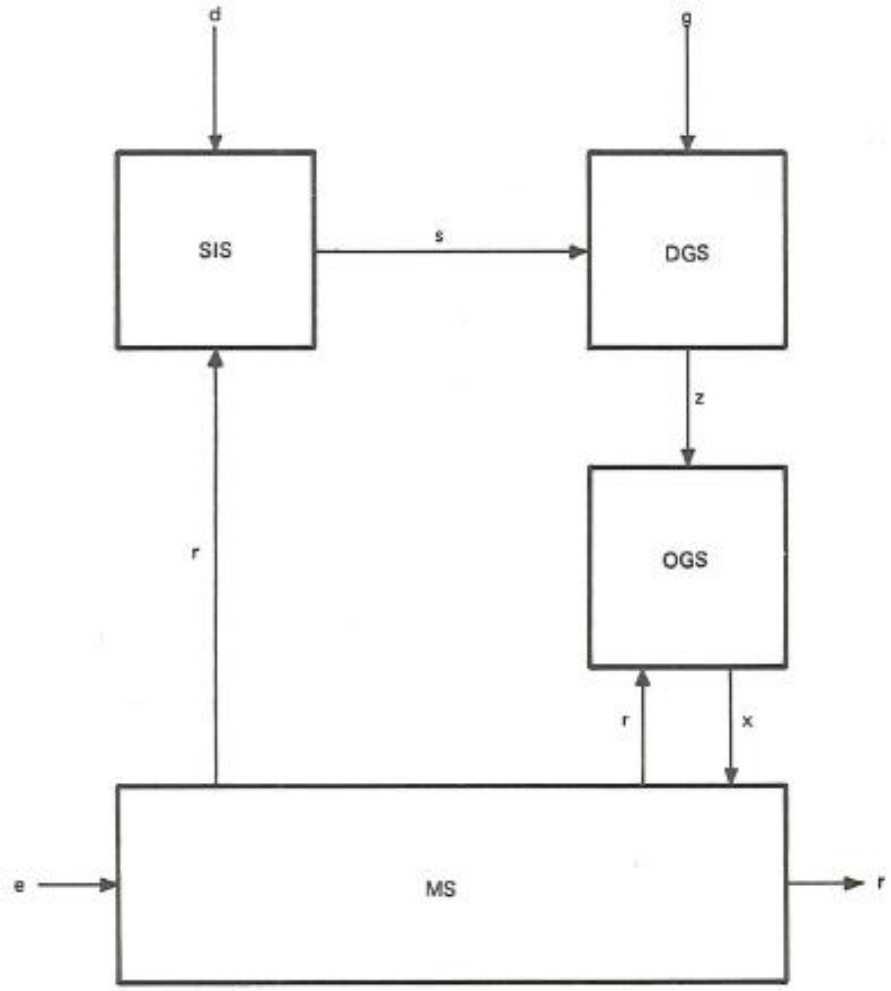

Figure 2: Subsystem of the nation

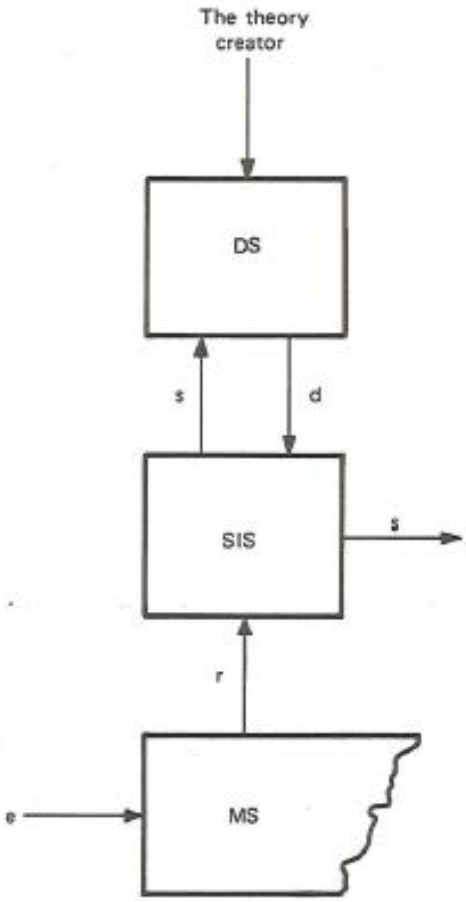

Figure 3: Designing system included as a part of the object system.

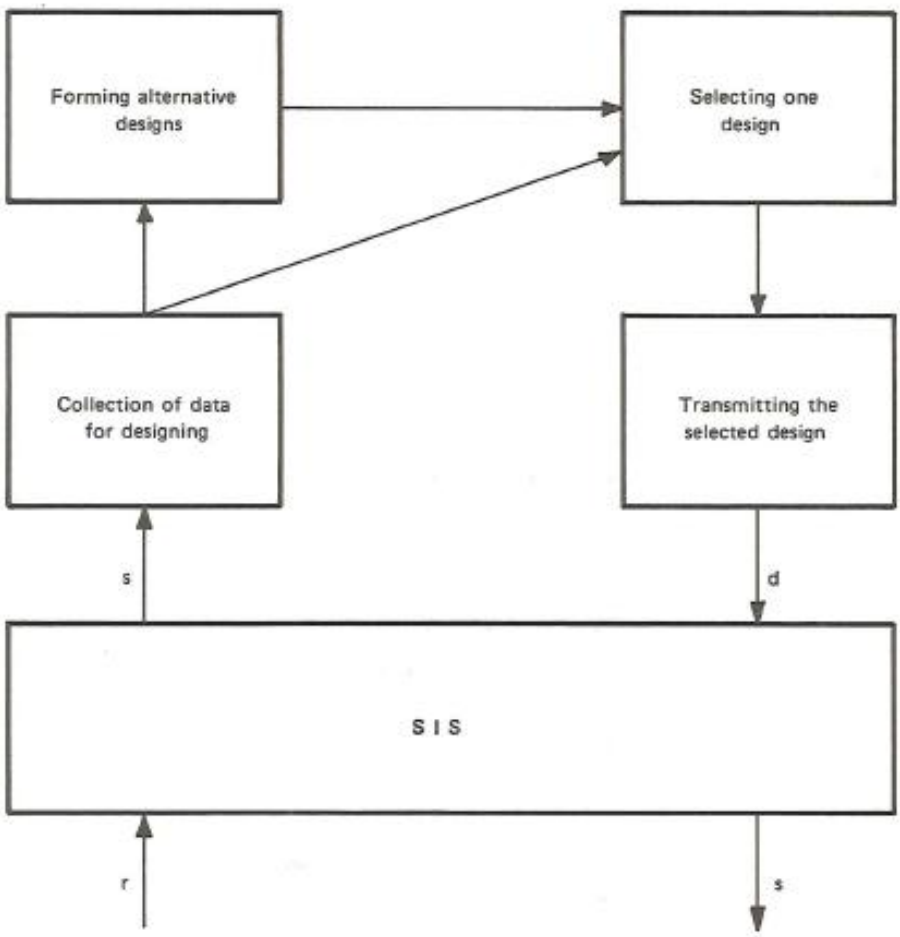

Figure 4: The four subsystems of DS.

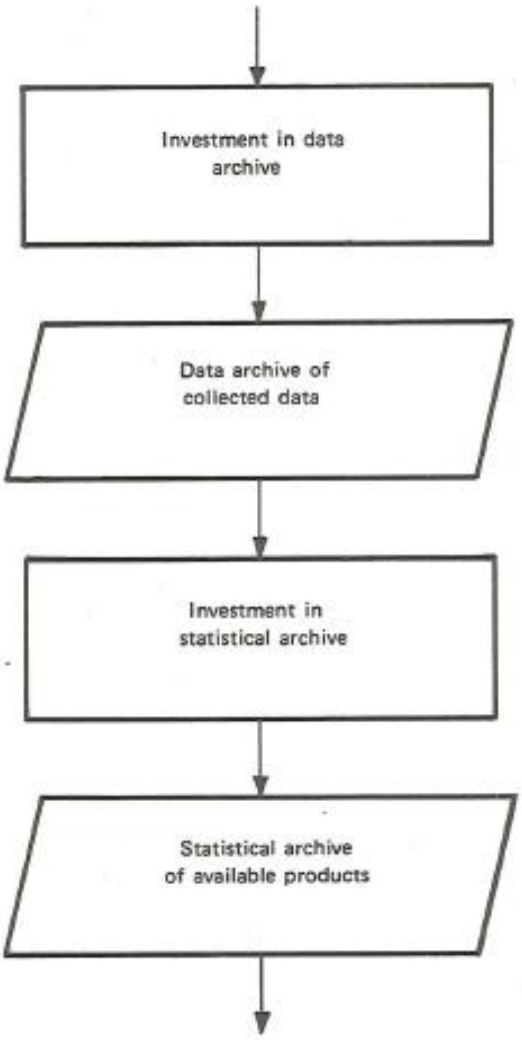

Figure 5: Main activities and products in SIS.