1977

Data Base Concepts for System Analysis

Thomas I. M. Ho

Report Number:

77-219

DATA BASE CONCEPTS FOR SYSTEMS ANALYSIS

by

Thomas I. M. Ho
Purdue University
Computer Sciences Department
and
School of Management
West Lafayette, Indiana 47907

CSD-TR 219

January 1977

DATA BASE CONCEPTS FOR SYSTEMS ANALYSIS

Thomas I. M. Ho
Purdue University
Computer Sciences Department
and
School of Management
West Lafayette, Indiana 47907

Keywords: systems analysis, data base, requirements statement

Categories: 3.50, 4.33, 4.34

ABSTRACT

Systems analysis is the examination of a problem situation in order to define the requirements of a solution, often computerized, to that problem. The diversity of problems and the constraints of computing technology require that a problem be thoroughly analyzed in order to insure that the problem is clearly understood. Then, and only then, it can be determined how computing technology can be applied to solve the identified problem. A particular difficulty during systems analysis is determination of the contents of the data base and design of the logical structure of that data base.

Several recent developments offer new perspectives for systems analysis and data base design. These developments provide a conceptual framework for understanding information system and data base characteristics. This framework supports an improved methodology for systems analysis.

INTRODUCTION

The practice of information systems development has been conspicuously impacted by the recognition of data as a resource to be managed [7]. This recognition has been accompanied by the advent of data base management systems (DBMS), software systems for performing the effective management of data resources. DBMS technology has been extensively reported in the computing literature, but this coverage has been deficient in providing a framework for applying the technology to information system problems The literature contains numerous definitions of DBMS concepts and examples of familiar data structures that illustrate the concepts. However, the basic data definition facilities of DBMS do not provide a perspective for understanding the data requirements of decision-making and other activities that require data management support. The data definition facilities are only useful to those who already know their data requirements and who are ready to translate those requirements into a data definition schema that can be managed by DBMS. There is scant discussion of the role of the data base in an information system. The role of the data base is the primary motivation for the contents of the data base. Defining the contents of the data base is one of the major tasks in conducting a systems analysis in order to understand an information system and then to translate that understanding into a perspective that reveals how computing technology may be applied to implementing that information system.

Systems analysis involves understanding the two diverse aspects of a management information system [5]. The organization system characterizes organizational activity while the information system models a conceptual framework that enables translation of the organizational perspective into a computer solution.

A framework for the application of DBMS technology to management information systems includes:

1. Characterization of the role of data in the organization system
2. Characterization of the role of the data base in an information system
3. Translation of the organizational perspective into an information system representation.

THE ROLE OF DATA IN THE ORGANIZATION SYSTEM

The organization system is composed of functional subsystems that represent the various functions performed by the organization. In performing its functions, each subsystem seeks to fulfill certain objectives with respect to various persons, objects, and events. For example, the Order Entry Subsystem seeks to fulfill events called Orders from persons called Customers who desire objects called Products.

In pursuit of its objectives, each subsystem must perform certain actions and make certain decisions. For example, the Order Entry Subsystem must perform an order cost computation and make an order acceptance decision. For each action to be performed, the subsystem must contain a procedure that performs that action. For each decision to be made, the subsystem must contain a model that makes that decision. In order to perform the required action, a procedure requires data that describes the environment being affected by that action. For example, the order cost computation procedure requires data that describes the Order being processed and the Product being ordered. In order to make the required decision, a model requires data that describes the constraints and objectives of that decision. For example, the order acceptance decision model requires data that describes the Customer and the Order that he placed.

These data requirements motivate the need to know the values of relevant attributes of the various persons, objects, and events of interest to the procedures and models. For example, with respect to the event Order, the order cost computation procedure needs to know about the Order attribute Quantity and the order acceptance decision model needs to know about the Order attribute Cost. In addition, the data requirements motivate the need to know the instances of various associations among the various persons, objects, and events. For example, with respect to the event Order and the person Customer, the order acceptance decision model needs to know about the association between an Order and a Customer that indicates which Customer placed a particular Order.

The diversity of activities in the organization system imposes extraordinary demands on the data requirements. Diverse organizational activities often have common data requirements. These common data requirements are primarily motivated by the need to co-ordinate these diverse activities. For example, the need to co-ordinate the Inventory and Purchasing activities requires common knowledge of purchase requests. In order to insure that Purchasing procures appropriate quantities of materials so that Inventory suffers from neither surplus nor shortage, Inventory issues purchase requests to Purchasing. Furthermore, Inventory might determine the frequency and volume of purchase requests by virtue of Inventory's knowledge of Production's material requirements and of Purchasing's procurement lead times. Therefore, we see further instances of common data requirements among diverse organizational activities.

This characteristic of organizational activities is only one aspect of organizational activity that is impacted by data's role in the organization system. In addition to data's role in the co-ordination of organizational activities data is relevant to several other organizational concerns.

The multiple instances of any of the various persons, objects, or events of organizational interest are likely to possess different values for the same attribute. To insure correctness of the organization's knowledge of the relevant attribute values of a particular instance of a person, object, or event, the organization must be able to distinguish one instance of a person, object, or event from another. Furthermore, the existence of multiple instances motivates the need to categorize instances to distinguish some instances from other instances. For example, in a Personnel environment, there might be a need to distinguish salaried employees from non-salaried (hourly) employees. Finally, the existence of multiple instances motivates the need to find all instances that have some characteristic(s) in common. For example, in a Sales environment, there might be a need to find all the orders of a particular customer.

THE ROLE OF THE DATA BASE IN AN INFORMATION SYSTEM

The study of information systems is complicated by the dissimilarity of the organization system and the computer system. The organization system performs the activities that must be supported by the information system. The computer system performs computational and data management functions. There is no correspondence between the organization's activities and the computer's functions. The organization system contains the persons, objects, and events that are the subjects of organizational activities. The computer system contains the hardware and software facilities that perform computerized functions. There is no correspondence between the organization's subjects and the computer's facilities.

The gap between the organization and computer systems suggests the need for a conceptual bridge between these two systems. Such a bridge would guide and structure a systems analyst's activities while he formulates his approach to an organizational requirement. Such a bridge would not remove the necessity for the analyst to be familiar with the application domain with which he is dealing. Instead, the conceptual link identifies the concepts common to all applications of management information systems in order to supplement specific knowledge of organization and computer systems.

The conceptual link is an information system model [5] that provides a standard that enables organization system concepts to be expressed in a conceptual framework that is also compatible with computer system concepts. The information system model is itself a system composed of interacting subsystems:

1. Input subsystem
2. Output subsystem
3. Data base subsystem
4. Process subsystem.

The role of the information system model is illustrated in Figure 1.

The correspondence to the various subsystems of the computer system is clear and this is no surprise. Correspondence to the elements of the organization system can be established. The elements of the output subsystem correspond to the actions and decisions performed by each functional subsystem. The elements of the process subsystem correspond to the procedures and models used to perform each action or decision. The elements of the input subsystem correspond to the data received from the environment by the elements of the process subsystem to generate the elements of the output subsystem.

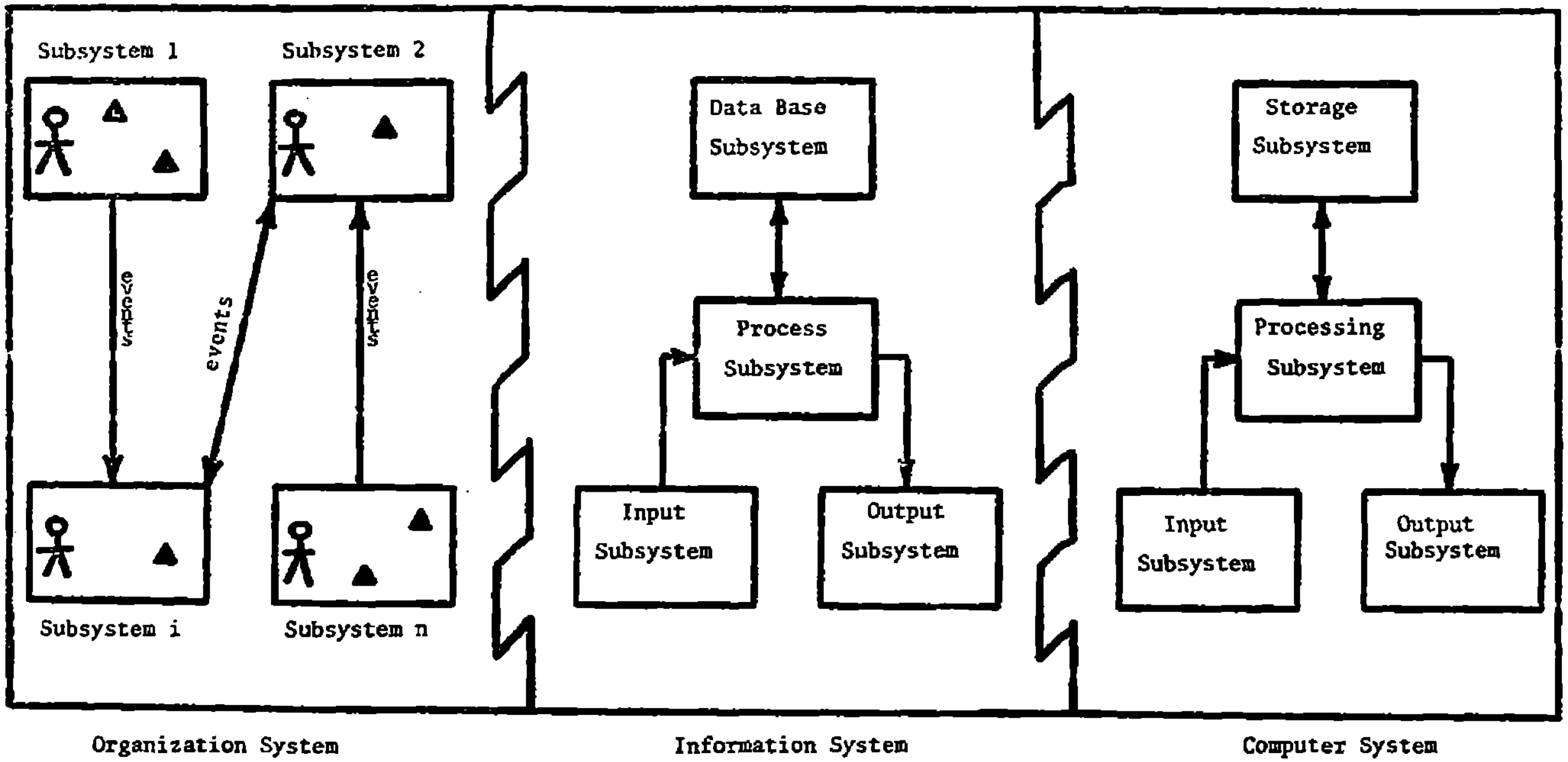

Figure 1. Management Information System

Data base subsystem

The data base subsystem serves as a decoupling mechanism between the input and output subsystems. The input subsystem gathers the data from the environment to be used to generate information to the environment through the output subsystem. However, the output subsystem does not necessarily generate information at the same time nor at the same rate as the input subsystem receives data. Therefore, the data base subsystem is an inventory of data resources. Furthermore, the output subsystem does not necessarily request information in a format that is identical with that of the data used to generate the desired information. Hence, the data base subsystem maintains a standard specification for data resources in order to decouple the incompatibilities between the input and output subsystems. The decoupling role of the data base subsystem in these respects motivates the residence of the data base subsystem in the storage subsystem of a computer system.

Elements of the process subsystem exist in a time dimension. For example, an organizational action is performed periodically by a procedure that corresponds to some process. Therefore, instances of that process occur periodically and so it is necessary to decouple an instance of a process in one time period from its instances in all other time periods, especially the next time period. In this respect, the data base subsystem serves as a decoupling mechanism between periodic instances of the process subsystem. An instance of the process subsystem in one time period receives data about some instance of a person, object, or event from the environment that is not necessarily received in any subsequent time period. Therefore, in order for an instance of the process subsystem in any subsequent time period to also use the same data, the data base subsystem serves as an inventory of data resources. In this way, relatively constant data that describes persons, objects, or events in the environment need not be received from the environment in every time period. For example, a sales information system maintains a data base representation of Customers and their common attributes, e.g. name and address. An instance of the process subsystem in one time period also computes data about some instance of a person, object, or event that must be used by another instance of the process subsystem in some subsequent, often the next, time period. Again, the data base subsystem serves as an inventory of data resources. In this way, cumulative data that describes an instance of a person, object, or event in some earlier time period need not be computed again by the process subsystem for use in the next time period. For example, a credit information system maintains a data base representation of Customers and their credit balances. In this way, the credit that has already been extended to a customer can be used as a factor in determining if additional credit should be subsequently extended.

With respect to the organization system, the data base subsystem also functions as a decoupling mechanism. The various functional subsystems of an organization system are interacting subsystems that must communicate with one another to achieve the desired synergistic effect. Again, the data base subsystem serves as both an inventory and as a standard for the data resources that are generated by any functional subsystem and can be used by any other functional subsystem in pursuit of that subsystem's objectives. Similarly, the data base subsystem also decouples separate procedures and models within a single subsystem. However, it is the data base subsystem's role as a decoupling mechanism between functional subsystems that elevates it to its central role in an integrated information system.

The final role of the data base subsystem in an information system is motivated by the other roles of data in the organization system. This final role of the data base subsystem is that of a standard representation of the mechanisms necessary to fulfill these other roles of data in the organization system. First, the data requirements of the various procedures and models that perform organizational actions and decisions necessitate a standard representation of the entity mechanism in order to represent the values of relevant attributes of the various persons, objects, and events of interest to the procedures and models. Second, the existence of multiple instances of any of the various persons, objects, and events of organizational interest necessitates a standard representation of the identifier mechanism of these entities in order to distinguish one instance from another. Third, the need to categorize instances necessitates a standard representation of the subsetting mechanism that distinguishes some instances from other instances. Finally, the need to find all instances that have some characteristics(s) in common necessitates a standard representation of the relationship mechanism that links any instance to all other instance(s) that have a common property with respect to the first instance.

DATA BASE CONCEPTS FOR SYSTEMS ANALYSIS

The multiple roles of the data base subsystem require definition of key data base concepts that describe the components of the data base subsystem in the information system. These concepts are the shapes and forms from which the data base definition of the information system blueprint is drawn. This blueprint is the statement of information system requirements that must be satisfied by a computer solution that supports the organization system. The drawing of this blueprint is the task of systems analysis.

Entity

An entity type is a model of a person, object, or event of interest to the organization system. An entity occurrence is the representation of an instance of the person, object, or event represented by the corresponding entity type. Therefore, EMPLOYEE may be an entity type while JOHN DOE is an occurrence of EMPLOYEE. An entity type consists of attributes that describe the entity. An entity occurrence consists of facts that describe the instance being represented. Therefore, if EMPLOYEE consists of the attributes NAME and ADDRESS, JOHN DOE might consist of the facts NAME is JOHN DOE and ADDRESS is 123 MAIN STREET. One or more of the attributes must serve as an identifier whose value distinguishes one occurrence of an entity from another occurrence of the same entity. Hence, EMPLOYEE might have the identifier EMPLOYEE-NUMBER.

The scope of an entity is arbitrary. Part of one entity can be separately defined as another entity. For example, an object entity called PRODUCT can also be defined in terms of another object entity called SUBASSEMBLY. Conversely, a collection of entities can be separately defined as another entity. For example, the collection of object entities PART and PRODUCT can be defined instead as the single object entity MATERIAL. Hence, in any organization system, any number of entity types is possible. Some guidelines for the selection of attributes of an entity have been presented by Brown [3].

Relationship

An entity type may be associated with some other entity type, not necessarily different from the first, by a relationship type. For each occurrence of one entity type, a relationship type defined between that first entity type and some other entity type defines a set of occurrences of the second entity type that have a common property (implied by the relationship) with respect to the occurrence of the first entity type. For example, a relationship type defined between the entity types CUSTOMER and ORDER defines the set of ORDER occurrences that were placed by each CUSTOMER occurrence. An important property of a relationship is its connectivity. For each occurrence of one entity type, connectivity indicates the maximum size of the set of occurrences of the second entity type that have the common property with respect to the occurrence of the first entity type. For example, the relationship between CUSTOMER and ORDER has connectivity 1 to N (≥ 1) because each CUSTOMER may place more than one ORDER, but each ORDER is placed by only one CUSTOMER. In addition, a relationship may also have connectivity 1 to 1, e.g. between SHIPMENT and BACKORDER if a SHIPMENT fulfills only one BACKORDER and if each BACKORDER is completely fulfilled by a single SHIPMENT, or connectivity N(≥ 1) to N(≥ 1), e.g. between PRODUCT and SUBASSEMBLY if a PRODUCT may consist of more than one SUBASSEMBLY and if a SUBASSEMBLY may be used in more than one PRODUCT. A final characteristic of a relationship is its associated-data which is data that describes the relationship, but that does not describe either of the participating entity types in isolation from the other entity type. For example, the relationship between PRODUCT and SUBASSEMBLY has associated-data that specifies the quantity of the SUBASSEMBLY that is used in each unit of the PRODUCT. The significance of the associated-data is readily apparent when one realizes that the connectivity of the relationship makes it meaningless to specify the associated-data as an attribute of either entity.

The relationship concept is essential to the fulfillment of the decoupling role of the data base subsystem. The integration of the various functional subsystems of the organization system is promoted by relationships among the entities that represent the various persons, objects, and events of interest. For example, integration of the activities of the Order Entry, Inventory, and Shipping subsystems motivates the relationships indicated in Figure 2. A rectangle is used to represent an entity and a diamond is used to represent a relationship. The relationships indicate the creation of either SHIPMENT or BACKORDER occurrences to fulfill each ORDER occurrence. In addition, the relationship between BACKORDER and SHIPMENT indicates creation of a SHIPMENT occurrence when each BACKORDER occurrence is fulfilled. With respect to a customer's inquiry concerning any ORDER, the contribution to integration is apparent in the ability to respond with relevant information of either SHIPMENTs or BACKORDERs that fulfill that ORDER. Feedback is also apparent in the ability to inform a customer of the imminent receipt of his unfulfilled ORDER by virtue of the fulfillment of the responsible BACKORDER.

Effective organizational control is promoted by the relationship concept. Control is possible only if there exists a sensor mechanism to detect a system state that is at variance with some designated system standard. The sensor mechanism is enabled by the data base representation of a relationship that enables ready detection of the variance condition. As illustrated in Figure 2, a relationship type between BACKORDER and PRODUCT enables easy detection of the variance condition exhibited by excessive backorders for any particular product.

The relationship concept also restores the loss of structure that is apparent when the scope of an entity is narrowed. When part of one entity is separately defined as another entity, the original data structure can be preserved by defining a relationship between the two entities. As illustrated in Figure 2, when a PRODUCT entity is defined in terms of a SUBASSEMBLY entity, structure can be preserved by a relationship type that defines the set of SUBASSEMBLY occurrences that compose each PRODUCT occurrence.

Finally, the relationship concept promotes data non-redundancy and its recognized contribution to data consistency and storage savings. Figure 2 includes a relationship type between CUSTOMER and ORDER that avoids redundant representation of CUSTOMER data in multiple ORDER occurrences placed by the same CUSTOMER.

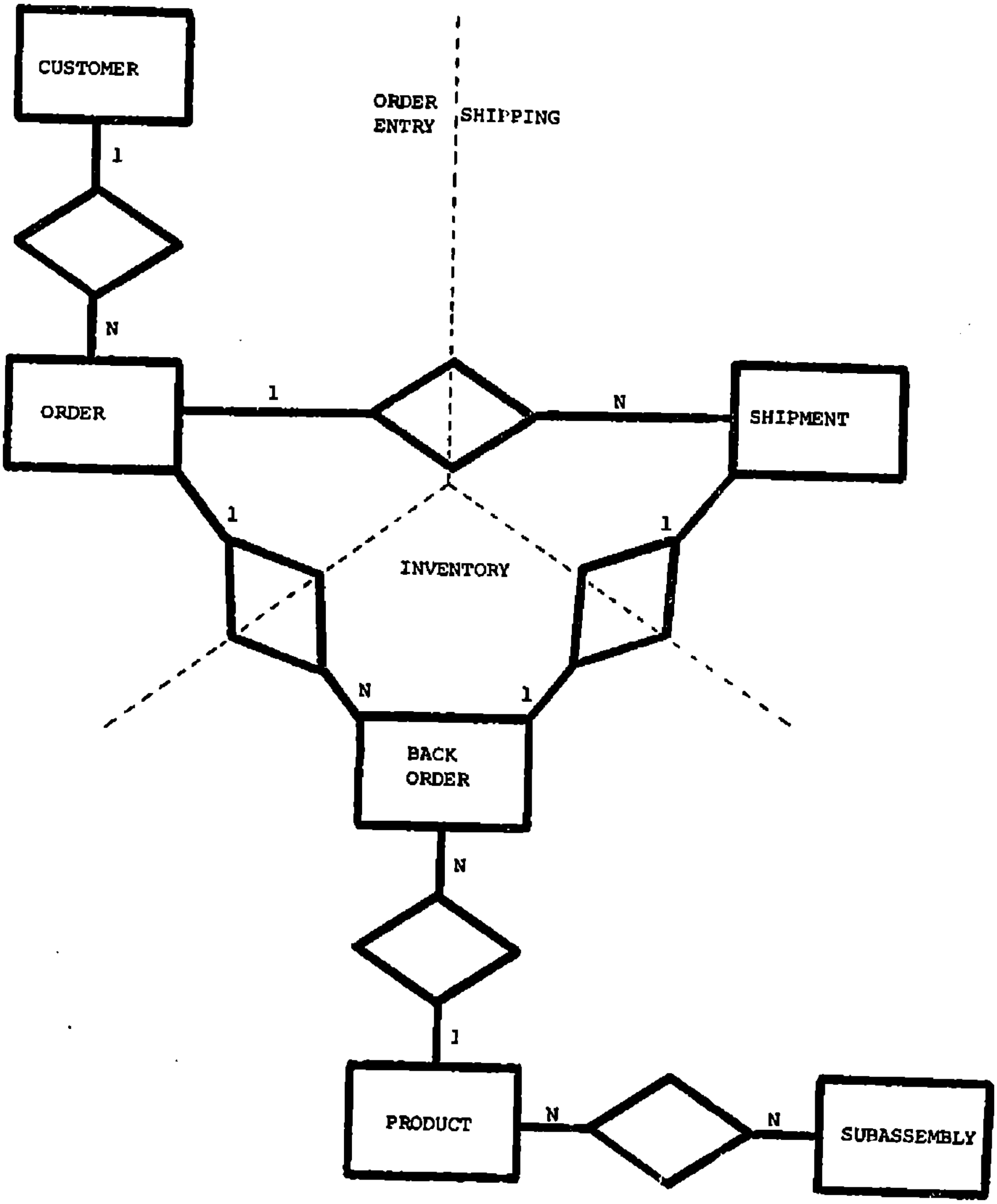
CUSTOMER
1
N
ORDER
ENTRY
SHIPPING
ORDER
1
N
SHIPMENT
1
INVENTORY
1
N
1
BACK
ORDER
N
1
PRODUCT
N
N
SUBASSEMBLY

FIGURE 2.

Subsetting criterion

The set of occurrences of an entity type may be partitioned into subsets according to some subsetting criterion. The characterization of subsets essentially distinguishes the entity occurrences in any particular subset from the entity occurrences in all other subsets. Similar in concept to a relationship, a subsetting criterion seems especially appropriate whenever a relationship with an entity representing some abstract idea is indicated. Instead of defining a relationship between some entity and an abstract entity, the abstract idea is designated as the subsetting criterion of the set of occurrences of the first entity. For example, designating DUE-DATE as the subsetting criterion of the set of BILL entity occurrences is equivalent to defining a relationship DUE between the entity DATE and the entity BILL. The relevance of a subsetting criterion to the control function is evident in the ability to recognize overdue bills. An alternate representation applicable to the control function might be the designation of DELINQUENCY-STATUS as the subsetting criterion for the set of occurrences of the BILL entity.

Timing

The role of the data base subsystem as a decoupling mechanism between periodic instances of the process subsystem motivates definition of the timing characteristics of each entity and relationship. The timing characteristics include frequency of data base update and lifetime of data base residence.

Frequency of data base update specifies the periodicity of the process that maintains the data base representation of each entity or relationship. In this context, time also serves as an identifier of data that distinguishes occurrences of one time period from those of another time period. For example, in a credit information system that processes customer charges and payments daily, the CUSTOMER entity that includes credit data has a daily frequency of data base update.

Lifetime of data base residence specifies the length of time that each occurrence of an entity or relationship is maintained in the data base. This characteristic is particularly important for calculating the volume of an event entity since the entity's lifetime and the event's rate of occurrence determine the maximum number of entity occurrences to be maintained in the data base. For example, under the assumption that the ORDER entity is maintained in the data base until the ORDER is paid, e.g. maximum of ninety days, and that a maximum of 500 orders are received each day, the maximum number of ORDER occurrences equals 500 ORDER occurrences/day X 90 days = 45000 ORDER occurrences.

Volume

Volume definition specifies the cardinality or maximum number of occurrences of each entity or relationship. Other than the cardinality of event entities previously described, the cardinality of an entity is determined through examination of the corresponding object or person in the organization system. For example, an organization that consists of a maximum of 1500 employees implies that the EMPLOYEE entity has cardinality of 1500.

The cardinality of a relationship is calculated from the relationship's connectivity and the cardinality of the participating entity types. For a relationship between ENT1 and ENT2 with connectivity n^1 TO n^2, the cardinality of the relationship equals

$$\min\{n^1 \times c^2, n^2 \times c^1\}$$

where c^1 is the cardinality of ENT1 and c^2 is the cardinality of ENT2. For example, the relationship between PRODUCT and SUBASSEMBLY with connectivity 2 TO 5 has cardinality of

$$\min\{2 \times 1000, 5 \times 300\} = \min\{2000, 1500\} = 1500$$

where PRODUCT has cardinality 300 and SUBASSEMBLY has cardinality 1000.

Summary

The data base concepts defined herein form a conceptual framework for the determination of data requirements during systems analysis. The objective of the systems analysis is to develop a perception of an organization system that enables statement of the information system requirements. The diversity of the organization system precludes a definitive technique for determining data requirements. However understanding the role of data in the organization system and the role of the data base subsystem in the information system provides a useful perspective for determining data requirements. Such a perspective recognizes the organization situations that inspire application of data base concepts to characterize the data requirements of an information system that supports the designated situation.

SYSTEMS ANALYSIS TECHNIQUES FOR DATA BASE CONCEPTS

The application of these data base concepts is supported by a variety of systems analysis techniques that enables the complete and consistent statement of data requirements.

Entity-relationship diagram

A useful perspective on data requirements is afforded by an entity-relationship diagram that supports the entity-relationship model of data proposed by Chen [4] and the ANSI/X3/SPARC Study Group on DBMS [1]. An entity-relationship diagram consists of rectangles that represent entities and of diamonds that represent relationships. Each diamond also consists of arrows that point to the entities that participate in the relationship represented by that diamond. An entity-relationship diagram is illustrated in Figure 2.

Requirements Statement Language

A Requirements Statement Language (RSL) is a high-level language for describing information system requirements that are determined during systems analysis and fulfilled during systems design. An RSL is not a programming language since an RSL statement expresses what requirements must be fulfilled rather than how those requirements are implemented in a hardware and software solution.

The most advanced RSL is the Problem Statement Language (PSL) developed by the Information Systems Design and Optimization System (ISDOS) Project [8]. PSL facilities [9] for statement of data requirements conform to the entity-relationship model of data [4]. The use of PSL for describing data requirements is described by Kahn [6]. The translation of a PSL statement of data requirements into a DBMS Data Definition Language schema is described by Blosser [2].

CONCLUSION

Organization system and information system perspectives on data requirements support the data base definition activity of systems analysis. In particular, systems analysis techniques to support the data base definition activity are available. Finally, these techniques can be employed to design the DBMS solution to the data requirements specified during systems analysis.

REFERENCES

1. ANSI/X3/SPARC Study Group on DBMS. Interim Report. Doc. No. 7514TS01, CBEMA, Washington, DC (February 1975), II-14 - II-15, VIII-17.

2. Blosser, P. A. An automatic system for application software generation and portability. Ph.D. dissertation, Purdue University (May 1976), 67-87.

3. Brown, A. P. G. Modelling a real world system and designing a schema to represent it. Data Base Description, Douque, B. C. M. and Nijssen, G. M., (eds.), North-Holland Publishing, Amsterdam, The Netherlands (1975), 339-347.

4. Chen, P. P. S. The entity-relationship model: Toward a unified view of data. Trans. Database Systems 1, 1 (March 1976), 9-36.

5. Ho, T. I. M. Systems analysis perspectives. Proc. 14th Annual Conference on Computer Personnel Research, ACM-SIGCPR, New York (July 1976), 12-19.

6. Kahn, B. K. A method for describing information required by the database design process. Proc. 1976 ACM-SIGMOD Intl. Conf on Management of Data, ACM-SIGMOD, New York (June 1976), 53-64.

7. Martin, J. Principles of Data-base Management. Prentice-Hall, Englewood Cliffs, NJ (1976), 3-4.

8. Teichroew, D. and Sayani, H. Automation of system building. Datamation 17, 16 (August 15, 1971), 26-30.

9. Teichroew, D. et. al. PSL user's manual. University of Michigan, Department of Industrial and Operations Engineering, ISDOS Working Paper No. 98 (March 1975), 2.3-1 - 2.3-9.