GRAD

# Abstract Formulation of Data Processing Problems

*by* JOHN W. YOUNG, JR., and HENRY K. KENT

*Advanced Systems Research Section, Product Specifications Department, The National Cash Register Company*

THERE are three stages in the application of high speed digital computers to data processing problems:

1. Systems Analysis—the task of determining what is to be done
2. Programming—a statement of how it is to be done
3. Coding—a translation of this statement into machine language

A wide variety of automatic coding techniques have been developed. This paper presents a first step in the direction of automatic programming as well as a tool which should be useful in systems analysis.

Since we may be called upon to evaluate different computers or to find alternate ways of organizing current systems, it is necessary to have some means of precisely stating a data processing problem independently of mechanization.

The notation presented here provides such a precise and abstract way of specifying the informational and time characteristics of a data processing problem and should enable the analyst to organize the problem around any piece of hardware. This notation could be used in the input to a new type of automatic programming system in which a problem is stated not only independently of machine, but without specifying the structure of files or sequence of operations where such specification is not needed for logical correctness. Further, the application of a graphical version of the notational system will show the relationships among information in input and output so that redundancies can be eliminated and alternative ways of processing studied.

The context of our analysis is that the objectives of the data processing system have been stated in terms of the required outputs; these outputs are not considered as subject to revision. On the other hand, although the inputs may be organized in any desired fashion it appears necessary or at least convenient to state one of the possible input organizations from which any equivalent one can be derived. It should be noted that the input may supply any one of a number of equivalent pieces of information, e.g., either customer name to be copied directly onto an output or an identification number from which the name can be looked up.

## BASIC COMPONENTS OF DATA PROCESSING PROBLEMS

A data processing problem can be described in terms of four kinds of basic components:

1. Information sets
2. Documents
3. Relationships
4. Operational requirements

### INFORMATION SETS

An information set which we call $P_i$ is a list of all possible items belonging to the same class. From the items in these lists are drawn the data which will flow through the system. The members of an information set may be considered as all the possible entries which could be made in a specified blank on a document. Examples of information sets are customers' names and addresses, part numbers, unit prices, dates, invoice numbers, etc. Whether an information set or a collection of related information sets exists in the system as a file will depend on the ultimate mechanization of the problem. As far as an abstract statement of this phase of problem is concerned, it is necessary to state only the information sets and their relationships. Governed by these relationships the analyst is then free to choose whatever file structure will make the over-all system most efficient.

### DOCUMENTS

A document, which we call $D_j$, is a collection of related information items, $D_{jk}$ where $k$ is the number of the items on the $j$th document. Documents are either inputs or outputs; examples are shipping notice, invoice, daily sales report, etc.

### RELATIONSHIPS

The third element which must be described is the set of relationships among the information sets, the documents, and the items on the documents. A relationship between information sets specifies the correspondence between the items in one set and the items in some other set. For example, relationships exist between the corresponding items of the set of customers' names and customers' numbers, salesman's names and customers, day-month-year's and dates, etc.

The relationship which shows how an item on an output document is derived, or from where it is taken if it is simply copied from another document, is called the defining relationship. In addition, there are producing relationships which define under what stimuli the documents or certain items on the documents are produced,

Reprinted from the November–December 1958 issue of the Journal of Industrial Engineering, a publication of the American Institute of Industrial Engineers, Inc., 145 North High Street, Columbus 15, Ohio

and conditional relationships which specify under what conditions any of the foregoing apply.

### OPERATIONAL REQUIREMENTS

The final element in a data processing problem is the set of operational requirements of the system. These are the requirements which are not related to the logic of the problem, such as the volume of input documents per day, the number of copies of outputs, the size or color of the documents, the elapsed time between the receipt of an input and the production of an output, etc.

## ABSTRACT STATEMENT OF A PROBLEM

An abstract statement of a problem can be made by preparing two lists: one of the information sets and the other of the documents.

The list of information sets presents the name of each set along with certain data about the number of items in each set, the number of characters required for each item, and the qualitative and quantitative relationships among sets. The list of documents presents for each document the items on the document and the information set each item belongs to. It also lists the defining and producing relationships for each item on the documents as well as the operational requirements associated with the documents.

Thus the output documents are completely specified in terms of the transformations which are applied to the inputs. At the same time, the relationships among the information sets enable logical substitutions to be made in the input to achieve the same output.

In order to describe the notation and its properties, we will proceed by means of an illustrative example.

## VERBAL STATEMENT OF PROBLEM

Suppose that we have a manufacturer who maintains stock at several scattered warehouses. Sales are made and reported to the warehouses, which in turn make the shipments and at the same time send information to the central office for billing purposes. The central office sends invoices to the customers, receives payments, furnishes monthly statements, and prepares a daily cumulative sales report for management purposes. Each customer is served by a given salesman who in turn works out of a given warehouse.

There are five documents in this system, two inputs (shipping notice, customer payment) and three outputs (invoice, monthly statement, daily cumulative sales report).

Three hundred shipping notices are received daily and each has, on the average, five line items. An invoice is prepared for each shipping notice. After determining the unit price and extended price for each line item, the total price per shipment is calculated. The invoice must be sent within two days after receipt of the shipping notice.

About two hundred payments are received daily, each covering on the average one invoice. The customer includes the invoice number with the payment.

Monthly statements are sent out between the tenth and the fifteenth of the month to all customers with an open balance as of the tenth of the month. The invoices which were dated after the tenth of the previous month and which have not yet been paid are itemized on the statement. Old unpaid invoices are consolidated and shown only as an old balance.

The daily sales report is produced within two days after the close of business. It includes the gross sales for the date and a breakdown by salesman of daily sales and cumulative sales for the past month.

An abstract statement of this problem is presented in Table 1 (List of Information Sets) and Table 2 (List of Documents). A graphical representation of the information sets is presented in Figure 1 and a combined graphical representation of both the information sets and documents is presented in Figure 4.

TABLE 1

*Information Sets*

| | | $n$ | $L$ | Relationships |
|---|---|---|---|---|
| $P_1$ | Date | — | $6N$ | $P_1 = P_7 \times P_8 \times P_9$ |
| $P_2$ | Customer Identification No. | 2000 | $5N$ | $P_2 \simeq P_{10}$, $P_2 \sim P_4$ |
| $P_3$ | Ship to code | 9 | $1N$ | $P_{14} \simeq P_2 \times P_3$ |
| $P_4$ | Salesman No. | 50 | $2N$ | $P_2 \sim P_4 \sim P_{11}$ |
| $P_5$ | Model No. | 150 | $5A/N$ | $P_5 \simeq P_{12} \times P_{13}$ |
| $P_6$ | Quantity ordered | — | $2N$ | |
| $P_7$ | Day | 31 | $2N$ | $P_1 = P_7 \times P_8 \times P_9$ |
| $P_8$ | Month | 12 | $2A$ | $P_1 = P_7 \times P_8 \times P_9$ |
| $P_9$ | Year | 10 | $2N$ | $P_1 = P_7 \times P_8 \times P_9$ |
| $P_{10}$ | Customer $N/A$ | 2000 | $50A/N$ | $P_{10} \simeq P_2$ |
| $P_{11}$ | Warehouse name | 10 | $12A$ | $P_4 \sim P_{11} \sim P_{15}$ |
| $P_{12}$ | Part No. | 800 | $3A/N$ | $P_5 \simeq P_{12} \times P_{13}$ |
| $P_{13}$ | Color | 20 | $2A$ | $P_5 \simeq P_{12} \times P_{13}$ |
| $P_{14}$ | Ship to address | 6000 | $50A/N$ | $P_{14} \simeq P_2 \times P_3$ |
| $P_{15}$ | Pricing area | 8 | $1A$ | $P_5 \times P_{15} \sim P_{17}$ |
| $P_{16}$ | Invoice No. (Shipping Notice No.) | — | $5N$ | $P_{16} \simeq D_2$ |
| $P_{17}$ | Unit price | — | $5N$ | $P_5 \times P_{15} \sim P_{17}$ |
| $P_{18}$ | Salesman name | 50 | $15A$ | $P_{18} \simeq P_4$ |

$n$ = number of elements in set
$L$ = number of characters (numeric—$N$, alphabetic—$A$, or alphanumeric $A/N$) in each element.

## TABLE 2

*Document Descriptions*

### Shipping Notice—D1—Input

| Items | Verbal Description | Information Set | Defining Relationship |
|---|---|---|---|
| $D_{1-1}$ | Date | $P_1$ | |
| $D_{1-2}$ | Shipping Notice No. | $P_{16}$ | |
| $D_{1-3}$ | Customer Identification No. | $P_2$ | |
| $D_{1-4}$ | Ship to Code | $P_3$ | |
| $D_{1-5}$ | Salesman No. | $P_4$ | |
| $[D_{1-6}$ | Quantity of Order | $P_6$ | |
| $D_{1-7}]$ | Model No. | $P_5$ | |
| $D_{1-8}$ | Line Item | | $D_{1-6,7}$ |

Volume: $\overline{\sqsubset} D_1/P_{7E}=300$
$\overline{\sqsubset} D_{1-8}/D_1=5$

### Invoice—D2—Output

| Items | Verbal Description | Information Set | Defining Relationship |
|---|---|---|---|
| $D_{2-1}$ | Date | $P_1$ | $t_E(D_2)$ |
| $D_{2-2}$ | Invoice No. | $P_{16}$ | |
| $D_{2-3}$ | Customer Identification No. | $P_2$ | |
| $D_{2-4}$ | Customer Name and Address | $P_{10}$ | |
| $D_{2-5}$ | Ship to Address | $P_{14}$ | |
| $D_{2-6}$ | Warehouse shipped from | $P_{11}$ | |
| $[D_{2-7}$ | Quantity of order | $P_6$ | |
| $D_{2-8}$ | Model No. | $P_5$ | |
| $D_{2-9}$ | Unit Price | $P_{17}$ | $P_{17}(D_{2-8}\times P_{15})$ |
| $D_{2-10}]$ | Extended Price | | $D_{2-7}\cdot D_{2-9}$ |
| $D_{2-11}$ | Total Price | | $\Sigma\, D_{2-10}$ |
| $D_{2-12}$ | Line Item | | $D_{2-7 \text{ to } 10}$ |

Producing Relationship: $D_1 \rightarrow D_2$
Operational Requirement
Volume: $\overline{\sqsubset} D_2/P_{7E}=300$
Time: $t_E(D_2)-t_E(D_1)<2$ days

### Customer Payment—D3—Input

| Items | Verbal Description | Information Set | Defining Relationship |
|---|---|---|---|
| $D_{3-1}$ | Date | $P_1$ | |
| $[D_{3-2}$ | Invoice No. | $P_{16}$ | |
| $D_{3-3}]$ | Amount | | |
| $D_{3-4}$ | Line Item | | $D_{3-2,3}$ |

Volume: $\overline{\sqsubset} D_3/D_{7E}=200$
$\overline{\sqsubset} D_{3-4}/D_3=1.5$

### Monthy Statement—D4—Output

| Items | Verbal Description | Information Set | Defining Relationship |
|---|---|---|---|
| $D_{4-1}$ | Customer Name and Address | $P_{10}$ | |
| $D_{4-2}$ | Date | $P_1$ | $10/P_{8E}(D_4)/P_{9E}(D_4)$ Statements are to be dated the 10th of the month. |
| $D_{4-3}$ | Customer's (old) balance | | $D_{4-8}(D_{4-1},\ P_8-1)-\Sigma\ \overline{C}_{4-1}D_{3-3}(D_{4-1})$ |
| $[D_{4-4}$ | Invoice No. | $P_{16}$ | $D_{2-2}$ |
| $D_{4-5}$ | Date of Invoice | $P_1$ | $D_{2-1}$ |
| $D_{4-6}]$ | Amount of Invoice | | $D_{2-11}$ |
| $D_{4-7}$ | Line Item | | $D_{4-4,5,6}$ |
| $D_{4-8}$ | New Balance | | $D_{4-3}+\Sigma\ D_{4-6}$ |

Producing Relationship: $P_2\times P_8 \rightarrow D_4 \mid D_{4-8}\neq 0$
(statements are produced each month for each customer with a non-zero balance)
$D_2 \rightarrow D_{4-7} \mid C_{4-1} \wedge C_{4-2}$
(an invoice is included in the statement if both condition $C_{4-1}$ and $C_{4-2}$ are true)

Special Conditions: $C_{4-1}$: $[P_8(D_2)=P_8(D_4)\wedge P_7(D_2)<10]\vee[P_8(D_4)-1=P_8(D_2)\wedge P_7(D_2)>10]$
(the invoice was dated after the 10th of the preceding month but before the 10th of this month.)

$C_{4-2}$: $\overline{\exists} D_3[D_{3-2}(D_{2-2})]$
(a payment has not been received for the invoice)

Operational Requirements:

Volume: $\overline{\sqsubset} D_4/P_{8E}=500$
(the average number of statements issued per month is 500)
$\overline{\sqsubset} D_{4-7}/D_4=4$
(the average number of invoices (line items) itemized per statement is 4)
Time: $10<P_{7E}(D_4)<15$
(statements are to be produced between the 10th and the 15th of the month)

*(Continued on next page)*

TABLE 2—*continued*

Daily Cumulative Sales Report—D5—Output

| Items | Verbal Descriptions | Information Set | Defining Relationship |
|---|---|---|---|
| $D_{5-1}$ | Date | $P_1$ | |
| [$D_{5-2}$ | Salesman No. | $P_4$ | |
| $D_{5-3}$ | Salesman Name | $P_{18}$ | |
| $D_{5-4}$ | Sales this date | | $\Sigma D_{2-11}(D_{5-1}, D_{5-2})$ |
| $D_{5-5}$] | Cumulative sales this month | | $D_{5-4}+C_{5-1}\cdot D_{5-5}(D_{5-1}-1)$ |
| $D_{5-6}$ | Line Item | | $D_{5-2,3,4,5}$ |
| $D_{5-7}$ | Total gross sales this date | | $\Sigma D_{5-4}$ |

Producing Relationships: $P_1 \rightarrow D_5$
$P_4 \rightarrow D_{5-6}$

Conditions: $C_{5-1}$: $P_7(D_{5-1}) \neq 1$

Operational Requirements:
Volume: $\sqsubset D_{5-6}/D_5 = 50$
Time: $t_E(D_5) - t_I(D_5) < 2$ days

TABLE 3

*List of Symbols*

| | |
|---|---|
| $P_i$ | A list of all possible information belonging to the same class |
| $p_i$ | A specific member of the class |
| $D_j$ | A document |
| $d_j$ | A specific document belonging to the $D_j$ class |
| $D_{jk}$ | A collection of entries on the document $D_j$ |
| [ ] | Line item |
| $\simeq$ | Isomorphic (one to one correspondence) |
| $\sim$ | Homomorphic (many to one correspondence) |
| $\times$ | Cartesian product, e.g., $P_j \times P_k$ means a pair of $p_i$ and $p_k$ |
| $\subset$ | Contained in |
| $\rightarrow$ | Produces |
| $E$ | Extrinsic time (real time) |
| $I$ | Intrinsic time, e.g., date written on document |
| ( ) | Function |
| $C_{m-n}$ | $n$th Condition relating to the $m$th Document |
| $\vert$ | If |
| $\overline{C}_{m-n}$ | Negation of $C_{m-n}$, i.e., $\overline{C}_{m-n}$ is true if and only $C_{m-n}$ is false |
| $\sqsubset$ | Number of |
| $\bar{\sqsubset}$ | Average number of |
| $\exists$ | There exists |
| $\bar{\exists}$ | There does not exist |

The following format is used for the information sets:

| $P_i$ | Name | Number of Items | Number of Characters | Relationships |
|---|---|---|---|---|
| | | | (Numeric, N; Alphabetic, A; Alphanumeric, $A/N$) | |
| $P_{10}$ | Customer name and address | 2000 | 50 $A/N$ | $P_{10} \simeq P_2$ |
| . | | . | . | . |
| . | | . | . | . |
| . | | . | . | . |
| $P_{18}$ | Salesman name | 50 | 15$A$ | $P_4 \simeq P_{18} \sim P_{11}$ |

In this problem, the customer's name is always linked with his billing address and $P_{10}$ is simply a name for the collection of the two thousand names and addresses in the system. The fifty alphanumeric characters required represent the maximum number for any one name and address. We could, if desired, also list the mean and standard deviation, the range, or any other statistic which is applicable. The other information sets are obtained by examining all of the data flowing through the system and simply listing the sets as they are found. In this problem, there are eighteen such information sets.

## RELATIONSHIPS AMONG INFORMATION SETS

One relationship that frequently exists between sets of information items is a one-to-one correspondence. Whenever there is this one-to-one correspondence between the items in one set and the items in another set, we shall, borrowing a term from abstract algebra, call this relationship isomorphism. This relationship holds, for example, between $P_{10}$ (customers' names and addresses) and the information set that consists of customer identification numbers, $P_2$. That is, to each specific customer name and address corresponds exactly one customer identification number and vice versa. This relationship is read $P_{10}$ isomorphic to ($\simeq$) $P_2$.

If the correspondence between the items in the information sets is many-to-one rather than one-to-one, we shall call the relationship homomorphism. For example, there are many salesmen who work out of a given warehouse, so that we may say that $P_{18}$ (salesman name) is homomorphic to $P_{11}$ (warehouse name). Thus given the salesman, we can find the warehouse from which he works, but not vice versa. In other words, if $P_j$ is homomorphic to $P_k$ then given an element of $P_j$ we can find the corresponding element of $P_k$. However, unlike isomorphism, the relationship is not symmetric; given an element of $P_k$ there is a whole set (with one or more members) of elements of $P_j$ which correspond to it.

Since both isomorphism and homomorphism are transitive,[1] we can derive relationships which are not explicitly

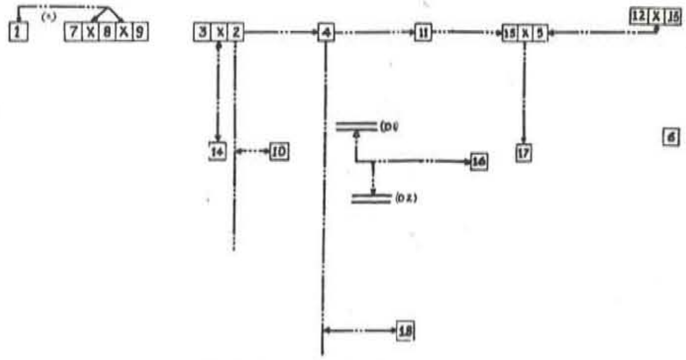


Fig. 1. Graphical Representation of Information Sets

[1] If $a$, $b$, and $c$ are three elements and $R$ is a relationship, then $R$ is said to be transitive if $aRb$ and $bRc$ together imply $aRc$.

given. For example, if $P_{10}$ (customer name and address) is homomorphic to $P_{18}$ (salesman name) which in turn is homomorphic to $P_{11}$ (warehouse name), we can then say $P_{10}$ is homomorphic to $P_{11}$.

Our graphical notation, shown in Figure 1, provides the simplest way to derive all of the relationships among the information sets.

The double-ended arrow indicates an isomorphism between sets and the single arrow, homomorphism. The relationships hold only in the direction of the arrows so that $P_{10}$ which is isomorphic to $P_2$ is also homomorphic to $P_4$, $P_{11}$ and $P_{15}$. For convenience in drawing, the arrows can be connected to a line extending from the squares.

An item of information is sometimes the result of combining several other items; e.g., the model number is determined by the combination of part number and color, or expressed symbolically $P_5 \simeq P_{12} \times P_{13}$; this is read $P_5$ isomorphic to $P_{12}$ cross $P_{13}$. Care must be taken to distinguish between this so-called cartesian or cross product, and arithmetic multiplication, which we always denote by $a \cdot b$ or just $ab$. The above says that there is a 1-to-1 relationship between the elements of $P_5$ and all pairs $(p_{12}, p_{13})$ where $p_{12}$ is a member of $P_{12}$ (i.e., a specific part number) and $p_{13}$ is a member of $P_{13}$ (i.e., a particular color).

Again the customer's actual shipping address ($P_{14}$) (store, warehouse, etc.) is determined by the combination of his identification number ($P_2$) and the ship-to code, i.e. $P_{14} \simeq P_2 \times P_3$.

In the case of date ($P_1$), we say that $P_1$ is equal to (rather than isomorphic to) $P_7 \times P_8 \times P_9$ (day, month and year) since an element of $P_1$ actually consists of the day, month and year taken from $P_7$, $P_8$ and $P_9$. Since certain triplets of day, month and year do not exist, it would be more correct to say that $P_1$ is contained in (symbolized by $\subset$ ) $P_7 \times P_8 \times P_9$. Operationally, however, this should be no problem since dates such as Feb. 31, 1958 should not enter the system.

$P_{16}$, the information set consisting of invoice numbers and shipping notice numbers (in this problem the same number is used for both) must be handled in a slightly different manner. Since these numbers are created at the same time as the documents they identify, they bear a definite relationship to these documents. In fact, there is a one-to-one correspondence between the numbers and the documents so that we may say that

$$P_{16} \simeq D_2$$
$$\text{and } P_{16} \simeq D_1$$

This implies a relationship between $P_{16}$ and any of the information sets to which the items on the document belong, viz:

$$P_{16} \rightharpoonup P_j$$

where $P_j$ is any information set to which an item on the document belongs. In other words, given an invoice number we can find the specific entries on the document.

The following format is used for the documents:

| Name of Document | | Numb:r ($D_j$) | Input/Output |
|---|---|---|---|
| Items | Verbal Dzeciption | Info-mation S.t | Defining Relationship |
| . | . | . | .. |
| . | . | . | . |
| . | . | . | . |
| $D_{jk}$ | — | $P_i$ | — |

Producing Relationships:
Special Conditions:
Operational Requirements
Volume:
Time:

For each document we record its name, number and whether it is an input or output document. The items $D_{jk}$ are entered along with their verbal description, the information set to which they belong, and, for output documents, the defining relationship, which tells from where the item is taken or how it is derived. If no relationship is shown, the item is derived directly or indirectly from the producing document. Whatever producing relationships and special conditions apply are listed at the bottom together with the volume and time requirements.

One characteristic of a document is that it contains not only information that appears only once but also information, known as line items, which may be repeated any number of times. The line item for the shipping notice ($D_1$) is $D_{1\text{-}6}$ and $D_{1\text{-}7}$, quantity ordered and model number, and is distinguished by placing square brackets around it. A particular shipment may consist of 7 different models so that the line item would be repeated seven times, each containing a different quantity ordered and model number. In order to refer to a line item as a whole, without the necessity of repeating its elements, an item, $D_{1\text{-}8}$, is added and used whenever reference is made to an operation which pertains to each of the line items. For convenience, the items contained in a line item are listed in the defining relationship column. Thus $D_{1\text{-}8}$ is shown to consist of $D_{1\text{-}6}$ and $D_{1\text{-}7}$.

We distinguish between two kinds of time, extrinsic or $E$ time and intrinsic or $I$ time. Extrinsic time is the time at which an event occurs, e.g., the production of an output document, and intrinsic time is the time which is an information item on a document, e.g., the date written on the document.

Using this symbolism we can now express elapsed time relationships. For example, if the invoice must be produced within two days after the receipt of the shipping notice by the system, we would write:

$$t_E(D_2) - t_E(D_1) < 2 \text{ days.}$$

Absolute instead of relative time can also be expressed, e.g.

$$10 < P_7[t_E(D_4)] < 15$$

implying that the day ($P_7$) of production of $D_4$ must be between the 10th and 15th (of the month). Expressions of this kind will often be condensed from $P_7[t_E(D_4)]$ to

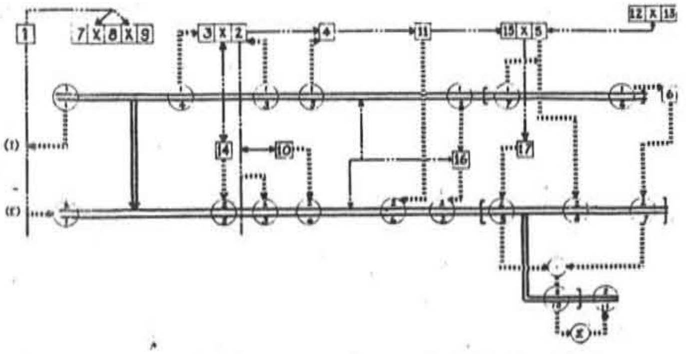

FIG. 2. Graphical Representation on $D_1$, Shipping Notice and $D_2$, Invoice

just $P_{7E}(D_4)$, since $P_7$ is an information set consisting of time units. We can apply arithmetic operations to dates as in the specification of $D_{5-5}$ in Table 2 where $D_{5-5}$ $[(D_{5-1}) - 1]$ means that we take $D_{5-1}$ (date), subtract 1 from it, and look up $D_{5-5}$ from the $D_5$ with this date.

The volume statements on the shipping notice state that the average number (indicated by the square $C$ with the bar over it) of shipping notices ($D_1$) per day ($P_{7E}$) is 300 and the average number of line items ($D_{1-8}$) per shipping notice is 5. The special symbols are used with an eye toward future automatic programming input.

A graphical representation of the shipping notice and the invoice together with the information sets is shown in Figure 2.

The items on a document are connected by a (generally horizontal) double line and the item numbers are entered in a circle. By convention the circles on input documents are connected to their corresponding information sets by a line with an arrow pointing toward the information set. The line items are enclosed by square brackets. The graphical representation makes it easy to pick up redundant information. For example, $D_{1-3}$ is a member of $P_2$ which in turn is homomorphic to $P_4 \cdot D_{1-5}$, which is a member of $P_4$, could therefore be derived from $D_{1-3}$ by a table look-up. In other words, it was not necessary to input the salesman's number as it could have been derived from a knowledge of the customer number. Note also that it was not necessary to know what meaning was given to $D_{1-3}$, $D_{1-5}$, $P_2$, and $P_4$ in order to determine this redundancy.

## PRODUCING RELATIONSHIPS

We have discussed certain relationships among the information sets; other kinds of relationship may exist among documents, and between documents and information sets. One of these relationships is that of "produces." In general, a document will be produced in one of two ways:

1. Periodically, once every day, week, etc.
2. Irregularly, once for each occurrence of an input document or of a condition dependent on input data.

For example, since an invoice ($D_2$) is created for every shipping notice ($D_1$) that is received, we say that

$$D_1 \rightarrow D_2 \text{ (read } D_1 \text{ produces } D_2\text{)}$$

The producing relationship is concerned with the existence rather than the content of the document involved. The daily sales report is produced each day, so that we say the date produces the report; however, this does not tell us what is on the sales report or from where its information is taken. In general, though, the items on an output document will be derived from the document which produced it, unless the defining relationship specifies some other way to derive these items.

## DEFINING RELATIONSHIPS

The defining relationship tells us the content of the output items. The exact specification of how this content is to be generated is left to the subsequent mechanization of the problem. Where no defining relationship is given, it is understood that the items are to be copied or derived from the input document that produces the output. These items are copied directly from the input when both input and output items belong to the same information set. They are derived indirectly from the input when the input and output items belong to different but related information sets. For example, the customer name and address ($D_{2-4}$) on the invoice ($D_2$) belongs to $P_{10}$ (the information set consisting of customer names and addresses). No defining relationship is given opposite $D_{2-4}$ in Table 2; thus for the derivation of $D_{2-4}$ we must look to the input document ($D_1$) which produces $D_2$. However, since no item in $D_1$ is a member of $P_{10}$ (that is, the customer name and address does not appear on the shipping notice), we must find some item on $D_1$ whose information set is related to $P_{10}$. Referring to Figure 1, we note that $P_2$ is isomorphic to $P_{10}$ and that $D_{1-3}$ is a member of $P_2$. Thus we know that $D_{2-4}$ can be derived from $D_{1-3}$ via the relationship between $P_{10}$ and $P_2$. In other words, given the customer identification number on an invoice we can find the customer name and address for the corresponding invoice. Note that nothing need be said as to how this is done. It may be desirable to maintain a file within the system and find the customer's names by a table look-up or this may be done manually

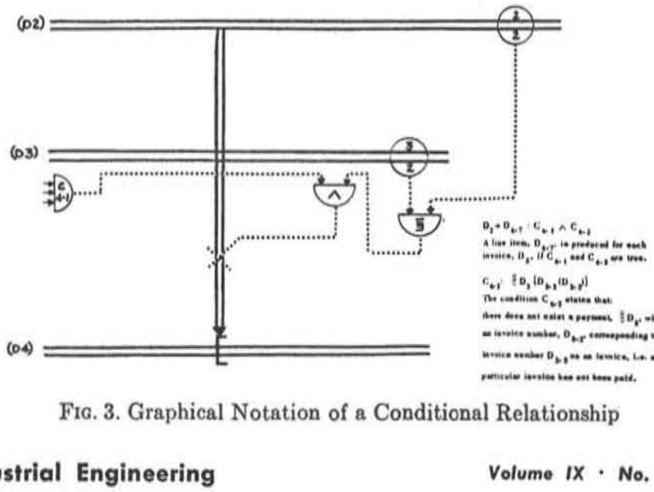


FIG. 3. Graphical Notation of a Conditional Relationship

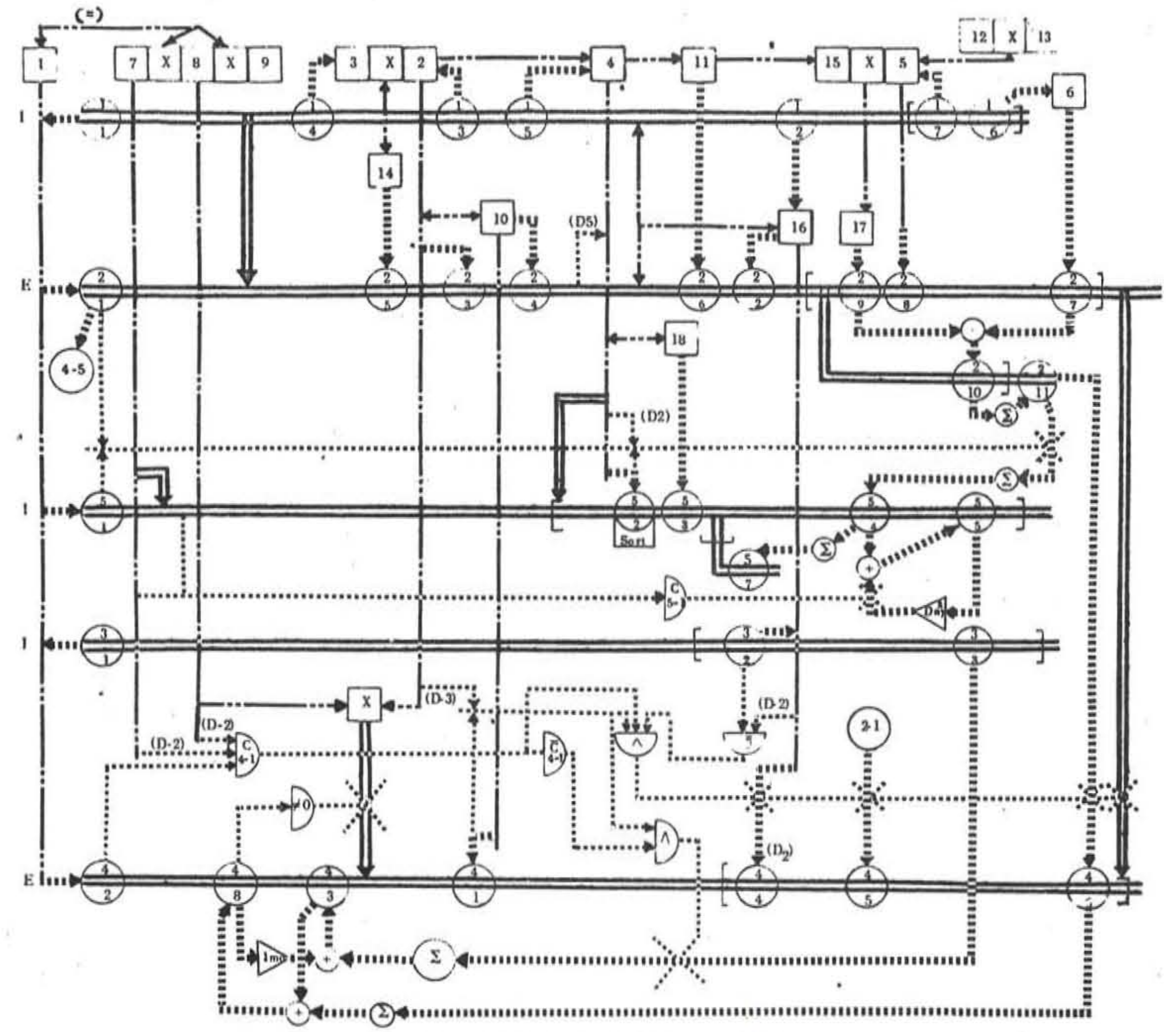

FIG. 4. Complete Graphical Representation

outside of the system. We have merely stated the requirement that it be done.

One way of symbolizing that an item in one information set is to be derived from an item in another set is by means of the function notation of mathematics. For example, the unit price depends upon both the model number and the pricing area in which the item is sold so that $D_{2\text{-}9}$, the unit price, (which belongs to $P_{17}$) is specified in Table 2 as $P_{17}(D_{2\text{-}8} \times P_{15})$. This is read as $P_{17}$ of $D_{2\text{-}8}$ cross $P_{15}$, and means that the system is to look at the combination of a model number ($D_{2\text{-}8}$) and a particular pricing area (belonging to $P_{15}$) and put the corresponding member of $P_{17}$ (unit price) in the space designated as $D_{2\text{-}9}$ on the invoice.

Note that the pricing area ($P_{15}$) may be derived in a number of ways, e.g., from a member of $P_{11}$ (via $P_{11} \frown P_{15}$), from a member of $P_4$ (via $P_4 \frown P_{15}$), or from a member of $P_2$ (via $P_2 \sim P_{15}$). That is, the pricing area can be found from either the warehouse name, the customer identification number, or the salesman number. Again, the particular choice of deriving the pricing area is left to the subsequent mechanization of the problem.

The defining relationship may be a simple mathematical relationship between various items, e.g., the extended price ($D_{2\text{-}10}$) is found by multiplying the quantity ordered ($D_{2\text{-}7}$) by the unit price ($D_{2\text{-}9}$). Thus, $D_{2\text{-}10}$ is defined as $D_{2\text{-}7} \cdot D_{2\text{-}9}$ (read $D_{2\text{-}7}$ times $D_{2\text{-}9}$). In the graphical notation, Figure 2, this is indicated by drawing information flow lines from $D_{2\text{-}7}$ and $D_{2\text{-}9}$ to a small circle which contains the arithmetic operation to be performed. The output of this operation is then shown going to $D_{2\text{-}10}$. The fact that $D_1$ produces $D_2$ is shown by the vertical double line with the arrow pointing from $D_1$ to $D_2$.

The defining relationships may become very complex. They may contain conditional statements or may involve arithmetic operations on time units.

### CONDITIONS

Decision making is another important part of any data processing problem; here we have available the help of symbolic logic (1). For example, the condition $C_{4\text{-}1}$, the first condition on the fourth document, is of the form ($p$ and $q$) or ($r$ and $s$) where $p$, $q$, $r$, and $s$ are equalities or inequalities on certain dates; $\wedge$ is used for "and," and "V" for "or." This condition states that the invoice was dated after the 10th of the preceding month but before the 10th of this month. $C_{4-1}$ is written as:

$$[P_8(D_2) = P_8(D_4) \wedge P_7(D_4) < 10] \vee [P_8(D_4) - 1 = P_8(D_2) \wedge P_7(D_2) > 10]$$

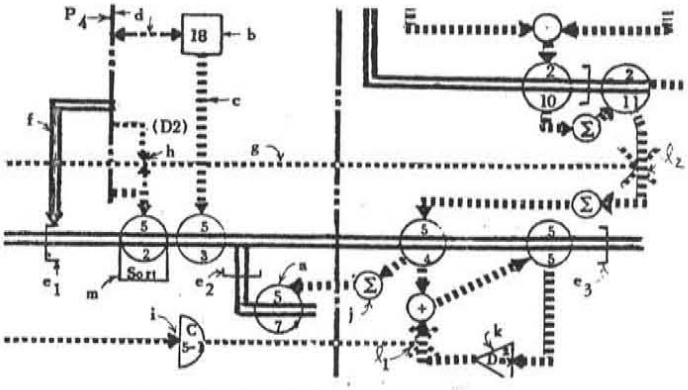


FIG. 5. Explanation of Graphical Notation

EXPLANATION OF GRAPHICAL NOTATION (FIGURE 5)

a. This circle represents an information item on a document, viz., $D_{5-7}$, with the document number above the line, and the item number within the document, below. The heavy double line (generally horizontal and possibly with branches) connects all the items on one document.

b. Each information set, $P_i$, is shown as a square, in this case $P_{18}$.

c. An item on a document which is an element of some $P_i$ is connected to it; the arrow runs out of the document for an input, into it for an output (i.e., one may think of the flow of information to and from the system as following the arrows).

d. Isomorphism (or homomorphism) between information sets is shown by a double (or single) headed arrow. In this case we are saying that $P_{18} \simeq P_4$, with the square standing for $P_4$ having been extended downward by the broken vertical line.

e. The small square brackets enclose those items making up a line item of a document, in this case $D_{5-2,\,3,\,4\text{ and }5}$. Note that $D_{5-1}$ is not a line item and is excluded by the bracket $e_2$.

f. The producing relationship is shown by a double heavy line with an arrow pointing to the document or line item produced, e.g., $P_4$ produces a line item on $D_5$. The line runs from the extension (the vertical broken line) of the square box representing $P_4$ to the bracket representing a line item on $D_5$.

g. The dotted line is used to connect various elements of a condition.

h. The condition at $h$ is that $D_{5-2}$ is equal to $P_4$ $(D_2)$, represented by condition lines from $D_{5-2}$ and $P_4$ with arrows to the condition line $g$. That $D_2$ is involved is shown by writing in $(D_2)$.

i. This is the standard symbol for a gate and states the condition involved, viz., $C_{5-1}$. The input to the gate (the flat surface) is connected to the item or items involved in the condition. The output goes to the operation affected by the condition.

j. The summation sign indicates that $D_{5-7}$ is equal to the sum of $D_{5-4}$.

k. The triangle represents a delay equal to the time indicated within it. Thus yesterday's $D_{5-5}$ is added to $D_{5-4}$ to produce today's $D_{5-5}$.

$l_1$ & $l_2$. The dotted circle through an information transfer line ($l_1$) or a sum ($l_2$) (or any other operation or production) indicates that this transfer or operation is performed if the condition is true.

m. The sort notation implies that the document or line item must be sorted by the item indicated.

A bar over a condition indicates the negation of that condition, i.e., $\overline{C}_{m-n}$ is true (or false) if $C_{m-n}$ is false (or true). The statement in a defining relationship of a calculation involving a condition is often simplified by interpreting $C_{m-n}$ as a number which is 1 if the condition is true and 0 if it is false. For example, one entry on the monthly statement $(D_4)$ is the customer's old balance $(D_{4-3})$. This is defined as last month's new balance for this customer minus payments against invoices dated prior to the period covered by this statement. $D_{4-3}$ is then defined as:

$$D_{4-8}\ (D_{4-1}, P_8 - 1) - \Sigma\, \overline{C}_{4-1}\, D_{3-3}\ \ (D_{4-1})$$

Thus the system is to find the new balance, $D_{4-8}$, on last month's, $P_8 - 1$, statement for this customer, $D_{4-1}$, and subtract from it the sum of this customer's payments, $D_{3-3}$ $(D_{4-1})$ which do not meet the condition that the payments were made during the time period covered by this statement, $\overline{C}_{4-1}$, where $C_{4-1}$ is the condition which specified the period.

The second condition, $C_{4-2}$, on the monthly statement states that a payment has not been received for the invoice; here "∃" is used for "there exists" and "∄" for "there does not exist" so that $C_{4-2}, \nexists D_3 [D_{3-2} (D_{2-2})]$, implies that the system started with $D_{2-2}$, and invoice number, and found that there did not exist in the system a customer payment, $D_3$, with the same invoice number $(D_{3-2})$.

Conditional statements may be involved in the producing relationship. In describing the conditions to be met before some action is taken, we use a vertical line to symbolize "if," e.g., in saying

$$D_2 \to D_{4-7} \mid C_{4-1} \wedge C_{4-2}$$

or an invoice $(D_2)$ produces a line item on the statement $(D_{4-7})$, if $C_{4-1}$ and $C_{4-2}$ are true.

Figure 3 illustrates this producing relationship. $D_2$ is shown producing the line item $D_{4-7}$ by having the producing arrow run to the square bracket indicating the line item. The condition on the producing relationship is indicated by the $X$ on the producing line leading to a gate. The input to this gate are the conditions $C_{4-1}$ and $C_{4-2}$. The condition $C_{4-2}$ is diagrammed completely showing that the input to the gate containing the "there does not exist" symbol is $D_{3-2}$ and $D_{2-2}$.

If the condition is simple it may follow the vertical line itself as in

$$P_{10} \times P_8 \to D_4 \mid D_{4-8} \neq 0$$

or a statement, $D_4$, is produced each month $(P_8)$ for every customer $(P_{10})$ who has an open balance $(D_{4-8} \neq 0)$.

## CONCLUSIONS

We have shown how a simplified problem can be stated in pseudomathematical terms. Such a tool has many advantages for a systems analyst. In reviewing an existing problem it provides the ability to insure that all of the input entries are used to produce outputs. By presenting all of the alternative methods of deriving information, it facilitates making decisions on the best organization of the inputs. It also helps in determining the cost of producing outputs, one criterion for their inclusion. For programming purposes, it provides an unambiguous statement of the problem to the programmer. The maximum number of files, record lengths, file densities, volumes, amount and type of computation required, etc., can be easily determined. The graphical presentation, which can be modified to suit the needs of the user (e.g., by including descriptive labels), should be helpful in determining the best organization of files and subroutines and in providing a check on redundant and superfluous information.

At first glance it may seem that we have made the problem statement more, rather than less, complicated. However, the statement of a problem in these terms clearly cannot be more complex than the problem itself. The cost of current data processing no longer permits incomplete or partial solutions, and ultimately, we would like the computer to solve the problem of its own best use. The first step toward this goal must be a statement of the problem which the computer can understand.

**REFERENCE:**

(1) Ackerman, S. A., "Symbolic Logic," *Journal of Industrial Engineering*, Vol. 8, No. 5, 1957, pp. 293-299.