THE FACT COMPILER: A SYSTEM FOR THE EXTRACTION, STORAGE, AND RETRIEVAL OF INFORMATION

Charles Kellogg
Intellectronics Laboratories, Ramo-Wooldridge
Canoga Park, California

Summary

The Fact Compiler is a system for the timely extraction of significant information from source data and for the storage of this information in an organized manner that permits rapid retrieval. In addition, the Fact Compiler can process or manipulate the stored data in a variety of ways, and it is adaptable for use with presend-day reporting techniques. The system is capable of orderly growth to meet the changing requirements of growing organizations.

Information is stored according to a logico-linguistic structure. This structure enables the system to: (1) directly interrogate personnel and request the reporting of specific information; (2) automatically present desired data at the appropriate time intervals; and (3) retrieve information according to subject, aspect, date, degree of specificity, and organizational unit.

The Fact Compiler

As society matures, production and use of information greatly increases. At the same time, individuals tend to narrow their fields of specialization in order to cope with this growth of information. Thus, the old adage about knowing more and more about less and less becomes more meaningful as time goes on. It becomes increasingly more difficult to assimilate and evaluate the mass of data contained in the paper work flooding many business, government, and university organizations. The techniques of automation have been applied in various ways to alleviate this problem. This paper will discuss an advanced system, the Fact Compiler, which uses new concepts to automate the storage and retrieval of information.

The Fact Compiler is a system for the timely extraction of significant information from source data and for the storage of this information in an organized manner that permits rapid retrieval. It has been designed to function as a centralized storehouse of important information. In addition to retrieving information, the Fact Compiler facilitates the processing or manipulation of the stored data in a variety of ways not readily possible with decentralized storage systems. The system is readily adaptable for use with present reporting techniques used in industry. At the same time, it can be used with fully automated input. The system is capable of orderly growth to meet the changing requirements of growing organizations. The Fact Compiler is an in-line processing system. That is, input data is immediately recorded in the memory and, therefore, is available for immediate use. The rapid response of the system permits the prompt discovery of trends in the stored data. It also encourages personnel to "browse" for information as an aid in finding all relevant data. The stored information, even if received from widely divergent sources, can be organized automatically and used in the generation of summary reports.

A key problem in designing the Fact Compiler was the development of a language capable of expressing the communication needs of personnel engaged in the input of information or in the formulation of a request for information. A large amount of linguistic and empirical study has resulted in a language structure that satisfactorily meets these objectives. Development of the language vocabulary required a precise formulation of the information needs of system users. These needs are expressed as a series of questions or interrogations which are indicative of the kind of information that should be stored within the system.

Interrogations are organized according to a logico-linguistic structure and are stored in the system memory. Input information consists of responses to the appropriate interrogations. The combination of an interrogation and its associated responses is called a factual statement. A fact is defined as an interrogation and one particular associated response.

General system features are shown in Figure 1. The system is composed of human and machine elements. In general, system personnel exercise judgment and perform decision-making functions, and the equipment elements perform the necessary routine processing. System personnel monitor and control operational functions, and aid in the input of data. The system uses a general-purpose digital computer and a high-capacity, rapid-access, magnetic memory. The greater portion of this memory is devoted to the file of facts. Other important files are the dictionary, or vocabulary of terms, and the interrogation vocabulary file. The input or extraction process is facilitated by the use of a communication console having an alphanumeric display tube and a special set of keys for data insertion. The output or retrieval process may use either the communication console or a high-speed printer.

Use of a Fact Compiler in a Business Firm

The concepts involved in the use of a Fact Compiler may be visualized in the application of the system to a hypothetical business firm. The organization structure of a typical business firm is illustrated in Figure 2. Some of the basic responsibilities of business management are the determination of the future course of the business by farsighted planning, the selection of qualified key personnel, development of a sound organization plan, and effective control through the delegation of responsibility. Management receives the data necessary for making decisions in these areas from a series of reports generated by line and staff departments. A common phenomena among

executive personnel is that of being "swamped" with reports. Development of procedures to filter information destined for management (i.e., selection of pertinent information and rejection of unnecessary data) is an important function usually performed by staff departments.

With a Fact Compiler, management and staff would, in effect, have an "extended memory" with which they could review particular aspects of the corporation's past history in various levels of detail. This type of review may permit better forecasts about future conditions such as costs, sale prices, raw material availability, and product demand. Also, the compiler could supply data indicative of the degree of control in effect within the firm. Correlations could be made between activities not previously compared and general economic conditions.

Interrogations for a business Fact Compiler System are organized into interrogation lists representing major subject categories of interest. Typical interrogation list categories might be "factory," "regional sales district," "subsidiary," "budget," "taxes," "legal," "patent," and "corporate assets." Information to be stored in the memory for a particular factory, regional sales district, or department is placed in an individual record for that factory, district, or department. This record is called a Unit Record. Unit Records may store information derived with the aid of one or more interrogation lists. The use of interrogation lists for data extraction and Unit Records for storage and retrieval is fundamental to the operation of the Compiler System.

Fact Compiler Language-Restricted English

A solution to the communication problem between humans and the computer files was of critical importance in the development of the Fact Compiler. It was evident that neither natural English nor machine language could be used. A compromise had to be made. After much experimentation, a language was developed which has proved satisfactory for both personnel and the computer; the language is called Restricted English.

Restricted English consists of a vocabulary of specially selected English words that are familiar and meaningful to personnel as well as being directly related to the types of interrogations and information requests that the system is designed to handle. Words are carefully selected with respect to meaning, since synonyms are not permitted. Restricted English terms may be "simplex" (single words) or "complex" (a word grouping or phrase). Complex terms are formed to represent unique, commonly occurring concepts. Rules have been developed to aid in the translation process from natural English word groupings to Restricted English simplex or complex terms.

The language consists of five parts of speech. Each of these different term categories serves a unique function. The parts of speech are:

Substantive. A noun or noun complex; a name for a thing or object.

Descriptor. Describes or limits definitions of substantives. They may be ordinary adjectives or past participles with adjectival functions.

Activity Connector. Relates several substantives in order to describe an action. They are usually present active or present passive participles.

Relational Connector. Prepositions, used for defining relationships between two substantives.

Interrogative Operator. Interrogative adverbial phrases used to determine the magnitude, quality, or position of things or objects of interest.

Restricted English does not use conjunctions or pronouns. Typical examples of business-oriented terms are shown in Table I. Most interrogations may be formulated by using four or five terms; however, a few may require as many as seven or eight.

The total set of terms defined for usage comprises the Fact Compiler dictionary. It is estimated that a dictionary of a few thousand terms should be sufficient to handle most business storage and retrieval applications. Each term in the dictionary is coded and assigned a four-digit tag for internal computer use. The tag identifies the unique term and the part of speech of the term.

Interrogation File Organization

Just as the dictionary defines a vocabulary of terms for the computer, the interrogation file defines a vocabulary of interrogations. As mentioned earlier, interrogations are filed by major areas of interest into interrogation lists. Each interrogation list is further divided into a five-level generic classification structure.

Levels of Generality

1.	Interrogation List	subject categories
2.	Capital Topic	
3.	Major	interrogation categories
4.	Minor	
5.	Sub	
6.	Sub Sub	

By assigning codes to each level, a "generic address" is formed which uniquely represents any interrogation used in the system. This address serves a very important purpose in identifying the subject category and level of generality of an interrogation. It also permits insertion of new interrogations at the end of a level without requiring a revision of the entire address structure as would be the case if an absolute address structure were used.

Interrogation file structure is outlined in Figure 3. Each interrogation is represented by its generic address, a series of four-digit term codes denoting interrogation content, and a series of criteria. Criteria are defined at the same time as the interrogations and permit the computer to perform the following operations:

1. Present the interrogation automatically to an input analyst on the basis of a predetermined elapsed time since the interrogation was last answered.
2. Present current factual data to personnel at desired periodic time intervals.
3. Select next interrogation on basis of response to present one.
4. Determine if an answer is in the proper form; if so, perform any processing that may be desirable before answer storage.

Extraction of Information—Input

The computer may take an active or passive part in the input of new information. This depends on the situation and system requirements. With little or no historical data in its memory, the computer's role would be basically passive, waiting for personnel to request interrogation displays and input answers. However, once initial data is stored, the Compiler System can compare data "age" with interrogation criteria and begin actively asking for new input at appropriate periodic intervals. This last feature can be quite useful in assuring the reporting of important facts and decreasing the possibility of reporting redundant facts.

Upon receipt of a new document, an input analyst must determine the Unit Record to which it applies and then select the major subject category involved. This information is conveyed to the system via push-buttons, and the input analyst is presented with a set of interrogation subject categories on the display scope. A process roughly analogous to the game of twenty questions ensues. The analyst, now aware of the categories of interest, reads the document until he finds reportable information. Depressing a key on his console, he generates an answer to an interrogation in the selected category. On the basis of the analyst's response, more detailed interrogations are presented as long as affirmative answers are supplied; negative answers cause the generation of higher-level interrogations, different subject categories, or result in terminating the interrogation generation.

The extraction of significant information thus proceeds, with the analyst alternately scanning the document and then reporting data in as much detail as possible. This extraction procedure, of course, does not depend on the existence of a source document. Source data that is in any form recognizable by humans may be used. Personnel could directly report facts from memory if desired.

Fact Storage Organization

The choice of an information storage file structure depends on many considerations such as the type of memory device used, estimated size of file, retrieval time requirements, and knowledge of the frequency distribution of various types of retrieval requests presented to the file.

The Fact Compiler memory is capable of storing one million factual statements with an average of ten answers per statement, for a total of ten million separate fact items. With this type of memory, it is desirable to store facts so that one or several entries to the memory will select the information records necessary to satisfy typical requests.

If the frequency distribution of requests were known in detail and did not change with time, an optimum file structure could be developed. However, the distribution estimated before the system is in operation is usually only a rough estimate, and the distribution will change with time as interest in various aspects of the stored information changes.

If it is necessary for retrieval time to be minimized, factual statements may be redundantly stored under several storage schemes, or storage schemes may be revised by the computer as the distribution changes with time.

For most purposes, such a high service rate would not be necessary, and one of the two following storage plans would be satisfactory. One or the other plan would be chosen on the basis of the predominant types of requests.

Fact Storage by Unit Record. A fact consists of an interrogation and the answer to that interrogation, derived from a particular source document. It is also associated with a particular staff department, factory, sales district, or other Unit Record. A Unit Record storage plan is shown in Figure 4. In the Unit Record shown, all information extracted from factory XYZ source documents is stored. Major columns are the statement generic address column and the response columns. A response consists of the answer extracted from a particular document (represented by its document number) and the date (date on document). Rows indicate specific generic addresses of interrogations used to extract information. An intersection of row and column will provide an answer to a particular interrogation referenced to the associated document number and date. In actual storage, each row is a separate variable-length field, and responses are scanned for selection purposes. Responses are stored in inverse time sequence, so the latest answer is nearest its generic address.

Fact Storage by Substantive Record. The second storage plan is based on the part of speech which stands for the name of the thing or object of interest—the substantive. Storage of factual statements by the names of things referred to in the statements is a powerful method because these names or substantives play a dominant role in most retrieval requests.

The layout of a Substantive Record storage plan is shown in Figure 5. Each Substantive Record contains the generic addresses of all statements which contain the particular substantive. The addresses are arranged in sequential order as in the Unit Record storage plan. However, as each address refers to one or more Unit Records, code numbers occur as sub-entries under each address. These represent the Unit Records for which each generic address applies. Therefore, generic addresses act as main entries and have Unit Record code numbers as sub-entries. Each sub-entry has stored with it a set of responses in the same manner as the Unit Record.

2.3 Retrieval

The concepts and procedures discussed up to this point pertain to the extraction of data from source documents and the storage of this data in an organized manner in a digital file unit. The stored data consists of factual statements—each statement consisting of an interrogation and a variable number of answers to the interrogation. Each answer is referenced to the number of the document from which it was obtained and the publication date of the document. Each statement is generically related to other statements in an interrogation list and is filed in accordance with the Unit Record to which it relates.

The retrieval system is capable of selecting the set of statements, the set of answers to these statements and, if required, the applicable document numbers that satisfy the conditions in a request for information.

Retrieval requests may take many forms. If, for instance, a top executive were interested in receiving the most up-to-date information on production schedules or sales forecasts, a console could be provided for him with a push-button for each category of information. Whenever the executive pushed one of the buttons, a signal would be sent to the computer which would select the latest answers to the relevant statements and present this information on a display device in the executive's office. For highly standardized requests, this method would be simple and effective. However, if all requests were of this nature, the Fact Compiler System would probably not be economically justifiable, as this type of limited, highly predictable information retrieval could probably be performed by humans at less cost.

As mentioned previously, the Fact Compiler is able to produce routine reports and present timely information on a daily basis automatically. It is felt, however, that the most important advantages of the system are its ability to retrieve and relate items of information not ordinarily compared or mentioned in normal business reports, and its capability of reviewing the past history of various aspects of business information which can be stored in some detail is desired. The ability to "probe in depth" into historical business data would seem especially advantageous for staff department personnel engaged in analysis of data for presentation to top management.

To provide this kind of capability, communication with the Fact Compiler must take place in a language powerful enough to express the many different and varied requests and yet simple enough so that humans can readily express what it is that they want from the file. An important part of the retrieval language is the vocabulary of terms used in defining the interrogations themselves. Terms are chosen from this vocabulary and are combined to express the factual information desired. A thesaurus will facilitate the correct choice of terms.

Restriction of the request to Interrogation List or Unit Record categories is made by including the names of the lists or records of interest. Periods of time for answer selection may be incorporated by using date operators such as: before-date, after-date, during-month (or year). Grouping within a request will be provided by the use of the logical operators AND-OR and appropriate use of punctuation. Certain arithmetical operations are allowable for selection and processing of answers. Examples are: Answers $\geq$ or $\leq$ a constant (for selection); sum, average (for processing a set of related answers). A retrieval request may be visualized as an algorithm which is capable of causing the Fact Compiler to select and present desired information to the requestor. The Fact Compiler retrieval language of English terms and operators permits appropriate connection of terms, restriction of request scope, and definition of processing operations. The combination of appropriate terms and operators will allow the ready generation of algorithms for the retrieval of information. The retrieval process is outlined in Figure 6.

As an example of a retrieval request, suppose it is necessary to have production and sales information on all dishwashers produced after June 1950. It is desired to relate this information to changes in the Gross National Product, Index of Industrial Production, and dishwasher development cost. The request may be formulated as follows: (Dishwasher, model number of, actual-forecast sales of monthly, actual-forecast production of monthly, development cost of; Gross National Product; Index of Industrial Production) after June 1950.

1. Parenthesis indicate all answers are to meet the date operator requirement. The date operator is "after June 1950."
2. Commas separate related terms.
3. Dashes separate modifiers.
4. Semicolons separate unrelated terms.

The decoding section of the retrieval program would break down this request, via the operators used and a knowledge of the allowable syntax patterns in statements, into the following phrases which may be partial or complete statements in the file.

1. Model number of dishwasher.
2. Actual monthly sales of dishwasher.
3. Forecast monthly sales of dishwasher.
4. Actual monthly production of dishwasher.
5. Forecast monthly production of dishwasher.
6. Development cost of dishwasher.
7. Gross National Product.
8. Index of Industrial Production.

The model number and production data may be stored in a Factory Unit Record. Development cost may be located in a Research and Development Unit Record. Sales information may be contained in several Regional Sales District Unit Records. Items 7 and 8 are reportable under a General Economic Parameter Unit Record.

In a file organized by substantives, all information is found under the three Substantive Records "Dishwasher," "Gross National Product," and "Industrial Production." However, in order to find the information, the generic addresses of the pertinent statements must be known. These generic addresses are found through the use of a fact

index. This index contains entries for all terms in the vocabulary and, for each entry, lists the generic addresses of each interrogation which contains that particular term.

Consider phrases 1 through 6 above. Each phrase consists of three separate terms. All addresses of facts that pertain to one of the phrases may be obtained by matching the addresses for each of the three entries and selecting those that are common to all three. When the appropriate sets of addresses are selected, the three Substantive Records may be directly addressed, the generic addresses located, and the date restrictions applied to date-answer combinations.

It may happen that retrieval requests that would be unanswerable may be formulated. This would most likely happen when too specific a request is made, so that the combination of terms used does not occur in any statement in the file. A request of this type would immediately be detected since conjunction of terms in the fact index would result in a zero set of generic addresses. This fact would be immediately presented to the requestor so he could modify his request.

Conversely, if a request were quite general, a tremendous amount of information would be presented. Before printing selected information, the computer would indicate the approximate volume of information that has been selected.

Monitor and Control

Figure 1 shows feedback to Monitor and Control from the input or extracting process, and the output or retrieval process. This feedback is an important part of the self-correcting procedure necessary to keep system performance at a high level of efficiency.

Feedback from the input process occurs when an input analyst finds that he cannot enter what he considers significant data into the machine memory because there are no appropriate interrogations in the interrogation vocabulary.

Feedback from the output process occurs when system users are not satisfied with the information contained in system-generated reports or answers to retrieval requests.

In either event, operator personnel must decide whether system capabilities should be changed to permit extraction, storage, and retrieval of new data. In making such decisions, they must be guided by firm management policy on the use of the system.

Systems capability may be modified by making changes in the interrogation vocabulary or criteria, the Restricted English vocabulary, or the Retrieval Operator vocabulary. Such changes are easily made at a special display console operated only by monitor personnel.

Additional functions of the group are to translate natural language requests into Restricted English, and to control and operate the system in accordance with the wishes of management.

Conclusion

The Fact Compiler System represents one solution to one category of information storage and retrieval problems. The predominant philosophy involved in its development has been that of an appropriate division of labor between humans and machines. Humans are most efficient in matters of judgment, machines in routine and precise processing. A machine will always need human-generated criteria in some form for determining what kind of information is significant or important since an absolute measure of the importance of information does not exist.

The value of a Fact Compiler System will be proportional to the amount of judgment and effort exerted in producing and maintaining a set of interrogations that adequately express the real requirements of system users. Producing such a set of interrogations is not an easy task but it is presently being done and is greatly facilitated by the use of a series of rules for translation from ordinary English to Restricted English Interrogation Lists.

A clearly defined unambiguous vocabulary in a specialized field of interest permits direct storage of factual information within a computer memory file. This file may be searched in several "dimensions" of index space. They are:

1. time,
2. subject category,
3. organizational unit (Unit Record),
4. aspect (substantives and their modifiers), and
5. degree of specificity (generic level).

In addition, the Fact File may be used as a powerful index to source documents. Requests may be formulated in a vocabulary which is easy to learn and to use. Generally-oriented or "browsing" requests may result in the discovery of relations between data never before apparent because of departmental report boundaries or because of the long time periods involved.

In many instances, management should be able to retrieve facts from the file quicker than they can find a report in a desk or file cabinet. Immediate access to this "extended memory" could help management in speeding up the decision-making process.

The Fact Compiler System equipment and computer program is capable of handling any interrogation vocabulary and set of facts that are generated in an appropriate form. For example, use by a large university might consist of allocating system time of several weeks each to medical, business, and engineering schools. Thus, the medical school might generate a Fact File of information about the diagnosis of various diseases, the business school might use the Fact Compiler to compare data from major industries, and the engineering school might use the system to store and retrieve data on use and development of digital computers. Each school could read in its own vocabulary and data, and then extract, store, and retrieve for the duration of its allocated time. At the end of the six-week cycle, the process could be repeated.

The organization of information within the Fact Compiler will permit the development of statistical procedures for automatic correlation of data within the file. This will be one of the objectives of future

system studies. It is expected that these studies will result in a capability for automatically finding cause and effect relationships between various types of information in the file.

Acknowledgement

The author wishes to acknowledge the valuable suggestions of S. Rothman, J. Dealy, and R. Penne of Ramo-Wooldridge, and the important work of Planning Research Corporation in the area of language studies.

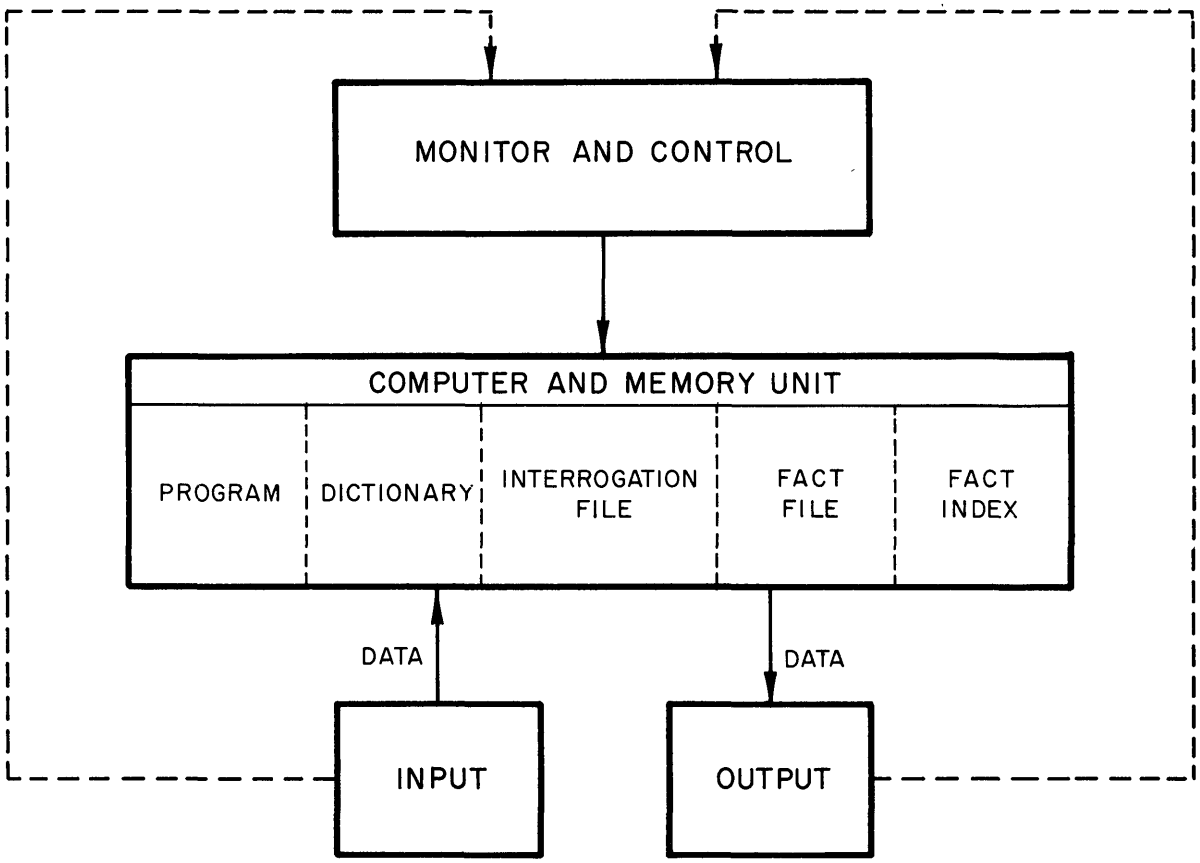

Figure 1 The Fact Compiler System

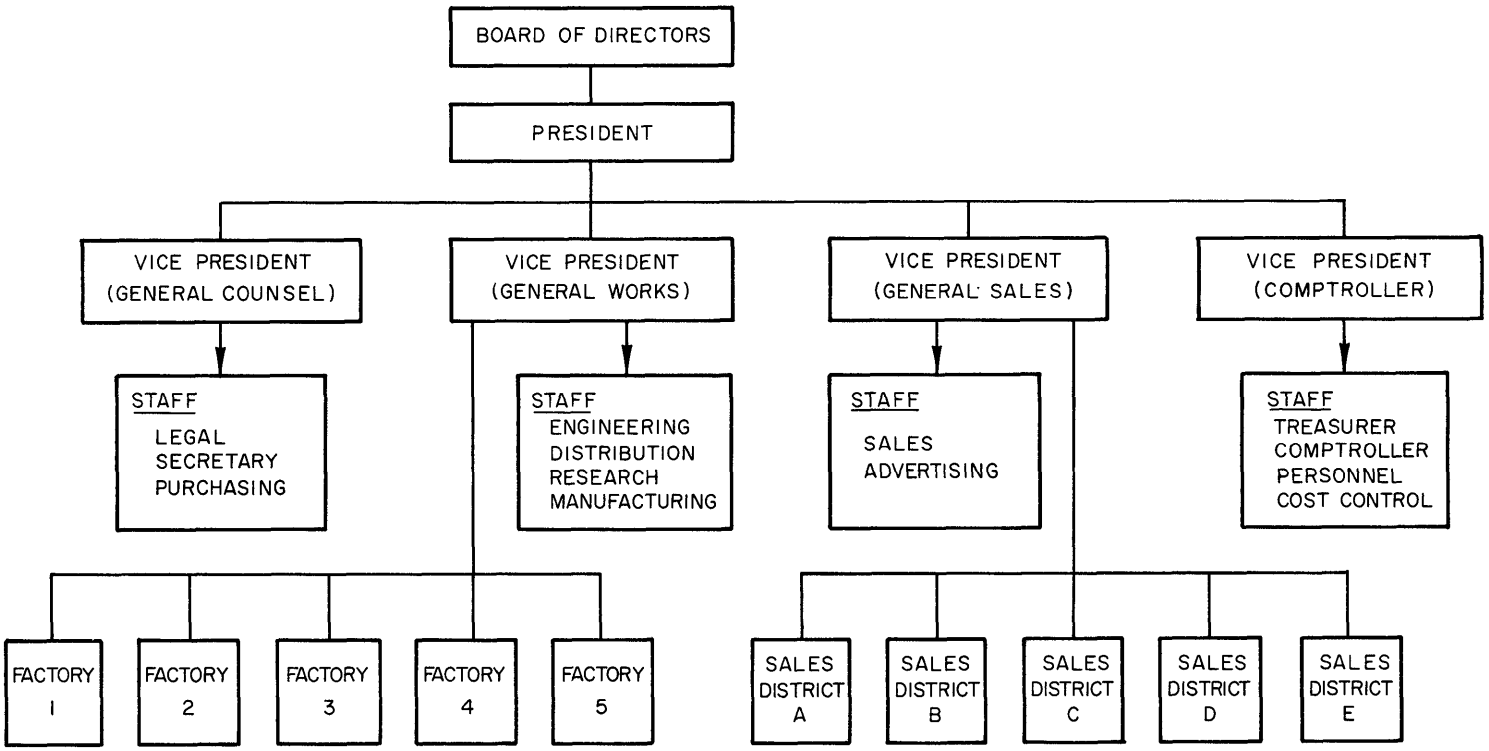

Figure 2 Business Organization Chart

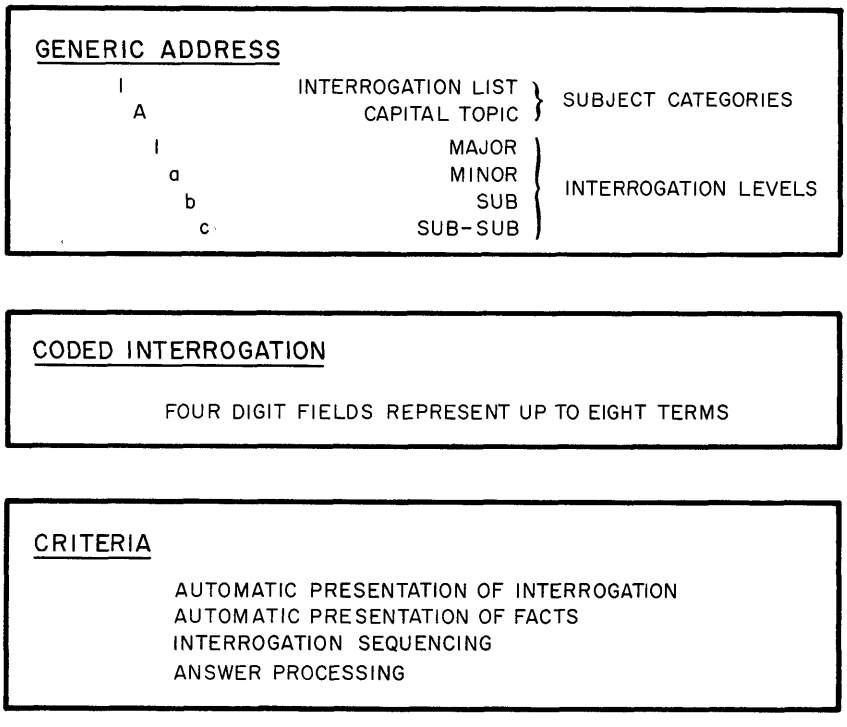

Figure 3 Interrogation File Organization

TITLE: FACTORY XYZ						
GENERIC ADDRESS	RESPONSE N			RESPONSE N-I		
	ANSWER	DOCUMENT NUMBER	DATE	ANSWER	DOCUMENT NUMBER	DATE
1A1aa						
1A1aaa						
1A1aab						
1A1aac						
1A1ab						
1A1aba						
1A1abb						

Figure 4 Unit Record Storage Plan

TITLE: OPERATING BUDGET							
GENERIC ADDRESS	UNIT RECORD CODE	RESPONSE N			RESPONSE N-I		
		ANSWER	DOCUMENT NUMBER	DATE	ANSWER	DOCUMENT NUMBER	DATE
1A1aa							
	03						
	05						
	09						
1A1aaa							
	03						
	05						
	09						
	10						
1A1aab							
	03						

Figure 5 Substantive Record Storage Plan

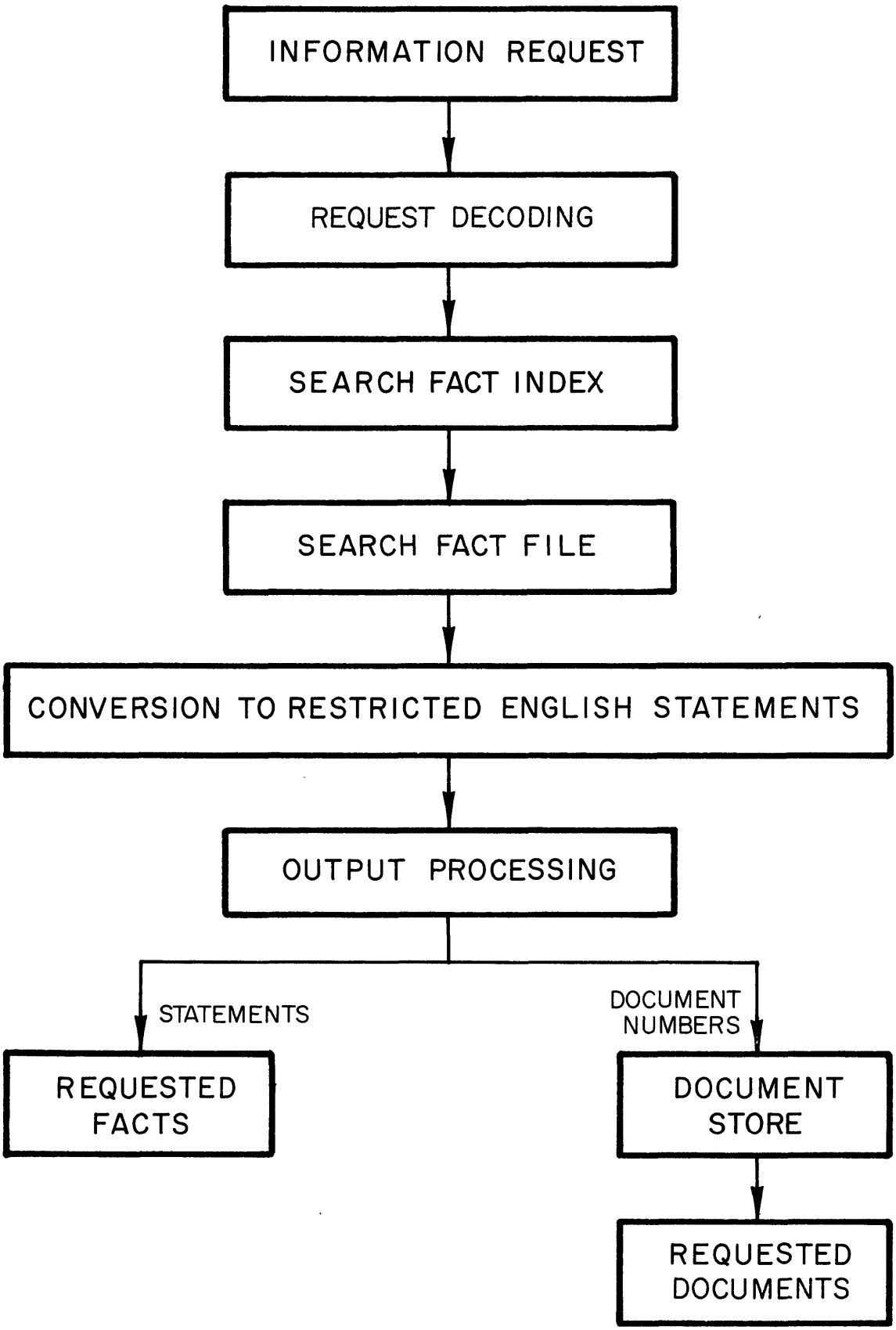

Figure 6 Retrieval

SUBSTANTIVE	DESCRIPTOR	ACTIVITY CONNECTOR	RELATIONAL CONNECTOR	INTERROGATIVE OPERATOR
GROSS NATIONAL PRODUCT	NEW	MOVING	IN	FORECAST VARIANCE OF
BUDGET	NAME	ARRIVING	AT	INVENTORY OF
CORPORATION	DECREASED	UNDERGOING	FROM	AMOUNT OF
KEY PERSONNEL	ESTIMATED	INCLUDING	FOR	PERCENTAGE COMPLETION OF
FOREIGN OPERATION	HEAT-TREATED	BEING MODIFIED	BY	NUMBER OF
SALARY	STATISTICALLY-CONTROLLED	BEING CONSIDERED	BEFORE	DOLLAR VOLUME OF

Table 1 Typical Restricted English Terms