An Operating System of Building Information

A Case History of Applied Research, conducted jointly by teams at Washington University, St. Louis, the Universite de Montreal and the University of California at Los Angeles

Research Team: Professor Colin H. Davidson, Mr. Michel Jullien, Mr. John Roberts, Mr. Helmut Schulitz and Mr. Leonard Wert. Notes prepared by Professor Colin H. Davidson and Mr. John Roberts.

Abstract

The project is described as applied research; it has grown up, in response to the information system needs of "Industrialisation Forum" - a quarterly publication at Washington University and the Université de Montréal. Experience shows the advantages of co-ordinate indexing for information storage and retrieval, compared to subject classification; procedures are described and the importance of a controlled vocabulary common to the indexer and searcher is discussed. Generation of the Thesaurus of Common Noun Key Words needs careful study; new logical rules have been devised for establishing hierarchies of terms - these introduce the new relationship: Part Term/ Whole Term to the current Narrower Term/Broader Term. In addition, the concept of Proper Noun Keywords is introduced and their rôle in information storage and retrieval discussed.

Introduction: Reasons for the Project

Today, research workers and practitioners in almost every field are faced with the increasingly time-consuming problems of information retrieval. The building industry is no exception; indeed, the fragmented structure of the industry contributes to the problems of information retrieval and information flow. Members of the team are active in the field of applied building research and as a result of suggestions made at the 1968 ACSA (1) seminar, it was decided to confront this set of urgent requirements and respond to them. Our immediate response was to launch "Industrialisation Forum". (2) Since that time, our research work in information science, necessary for the production of "I.F.", has assumed an importance in its own right. However, our research findings can (and, through publication dead-lines, have to) be applied and tested in practice; risks of yielding to expediency, and taking short cuts, can be minimised by conscious planning. We feel that this relationship between our research and its application most useful.

In the early stages of planning "I.F.", two important decisions were made: firstly, that the publication would not only contain information, but should constitute, by its nature, an information system so that readers could retrieve the information as their need arose; secondly, that the publication should be aimed toward anyone interested in the current changes in building (which are commonly called "industrialization") and should not reflect the information habits of any single group. The result of these decisions (made in the early summer of 1969) are embodied in "I.F.", the first issue of which (October 1969) included a description of the system prepared by Leonard Wert. (3)

Since the publication of "I.F." vol. 1, no. 1, the project team has necessarily been applying its original decisions to subsequent editions of the periodical. However, our interest in, and emerging skills in, information have been leading into other task areas. These are described in greater detail later in these notes.

The Principles of Information Handling: Post-Co-ordination

Several of the members of the project team had suffered from the current building information classification systems - U.D.C., SfB and so on. All of these systems fail to operate as effective guides for information storage and retrieval for several reasons: firstly, no document treats one subject only; secondly, each reader approaches each document with a specific slant corresponding to his interests at a particular moment in time; thirdly, subjects and areas of interest change, going beyond the original scope of the classification system.

Consequently, it was decided to extend the method of post-co-ordination - developed in other fields - into our own subject area, which was beginning to be described as "building science and technology".

Post-coordination (which could also be called "post classification") allows decisions about the relevance of the information in any document to be determined by the person who retrieves the document and not by the person who puts it into storage, as is the case with subject classification. The routines employed are probably familiar to many people by now, and comprise the following steps (presented here in outline):-

A. Information Storage

1. A document (book, report, letter, etc.) is received into the library; it is given an accession number;

2. It is read, and the concepts it contains are described by the selection of up to ten to fifteen keywords chosen from the controlled vocabulary (see next section); these words are displayed somewhere on the document where (a) they serve a control function and (b) they constitute a list of the concepts in the document (quite often an abstract is prepared at this time, it will carry the same accession number and the same keywords);

3. Special cards are prepared, one for each new keyword, and the accession number of the document is entered on it (this may be written or may be punched out - as in the case of perforated cards of some sort). In the case of keywords that have been found in some document earlier, and for which keyword cards have already been made, the existing card is brought out, and the accession number of the new document is added to it.

4. The keyword cards are filed alphabetically, and the document is put away in a location which is uniquely described by the accession number given to it (the accession number is a "zip-code" for the "address" of the shelf or file where the document is located and nothing more).

B. Information Retrieval

1. A person requiring a document about some subject (and he may or may not know that such a document exists), describes the concepts that constitute his area of interest, and about which he seeks information, using keywords selected from the controlled vocabulary; we explain the significance of the keywords later on in these notes. Experience shows that the searcher may use from five to ten words;

2. He picks out the keyword cards corresponding to his selection. Any accession number common to all of the cards he has chosen will indicate the address of documents about his subject. Note that accession numbers common to most of the cards will indicate documents about a large part of his area of interest;

3. He goes to the shelves and retrieves the document(s).

In this sequence: (a) the person who enters the keywords does not classify the document, he merely describes it using the controlled vocabulary and this stage should be as objective as possible; (b) the person who retrieves the document "classifies" his interest by his choice of keywords from the controlled vocabulary. There is no a priori restriction on his interest; rather his unique interest is expressed in terms of combinations of keywords whose meanings are carefully controlled.

Search Patterns

Our experience through observation of other information systems under operating conditions e.g. KWIC, SfB, UDC, etc. has led us to believe that there are two distinctly different types of information search: firstly, there is the general search where the researcher is not sufficiently well acquainted with the search area he is defining to give any clues other than the general concepts of his requirements; he does not know if any information exists in the area he is attempting to describe nor is he aware of the existence of other researchers active in the area; at best he can describe his problem in terms of a set of Common Noun Keywords such as exist in published thesauri. Secondly, a search may be undertaken by the researcher who either has definite knowledge that a document exists or can suggest the names of organisations or people responsible for the document, its physical form or its approximate date of publication. For example, it is common to be asked a question of the form: "Do you have that paper published by the National Research Council on Modular Coordination?"

Proper Noun Keywords

Through our observation of information searches in practice, we realized that the "additional"clues provided by the researcher who was already familiar with his subject were extremely important. Since the aim is rapid and effective retrieval of information, these clues must be taken advantage of. Consequently our retrieval system makes considerable use of these clues, which we call "Proper Noun Keywords" (to distinguish them from the Common Noun Keywords described earlier in these notes).

Proper Noun Keywords can be coordinated in the same way as Common Noun Keywords, either together or in combination with Common Noun Keywords. Furthermore a PNK search can always be followed by a Common Noun Keyword search in answer to a subsequent (explicit or implicit) request "What else do you have in this area?"

One of the advantages of the use of Proper Noun Keywords is that they do not require defining and so the equivalence of words selected by librarian and researcher is automatic. Some cross-referencing may be necessary, however, where the order of words in a Proper Noun Keyword is not certain or where abbreviations are employed.

Having described the two different types of searches which were observed, it was realized that the first type (the general search) could be translated into the second type (the restricted search) with the help of the information system operator, for this person can fulfill the rôle of the experienced researcher and start the search by coordinating the Proper

Noun Keywords of sources familiar to him. In this way, the system operator acts in a manner which resembles the experienced librarian. It is important, however, that he should follow up his PNK search with a CNK search, in order to be sure that the information store has been thoroughly combed. The intervention of such a "specialist" may greatly increase the effectiveness of the system in the general search.

The use of PNK's, as opposed to CNK's, assumes particular importance in, for example, an office information system, where the names of clients, members of the staff, projects,etc.. can all become PNK's (though obviously some care is required in deciding which ones to use).

Controlled Vocabulary: Thesaurus of Common Noun Keywords

Two reasons are usually given to explain the need for controlled vocabularies: to restrict the number of terms allowed and to make clear the relationships between them. The degree of control required depends directly upon the information to be stored in the information system. For example, in the legal profession, where documents not only constitute the working material of the lawyer but also contain words and phrases the interpretation of which is the process of law, current research projects in legal information systems (4) have found it impossible to either restrict the number of keywords or to define relationships between them. In the building industry, this is not the case. Documents are only a medium for the transfer of information and it is possible to define terms ahead of time; indeed, this is desirable because practitioners in the building industry come from a wide range of backgrounds whereas those in the legal profession have all the same fundamental education.

The success of the storage and retrieval of information by this method depends very considerably on the equivalence between the choice of keywords by the librarian when describing each concept in any document and the choice of keywords by the searcher when describing each aspect of his subject. To ensure this equivalence in choice of words (or at least to minimize the risk of divergent choices) a common, or controlled vocabulary of Common Noun Keywords is indispensible.

A number of controlled vocabularies exist such as the E.J.C. thesaurus (5). Quite apart from the fact that these thesauri are not oriented towards the area of "building science and technology",they also present some serious weaknesses due to their having (i) too many terms and (ii) too loose a structure. In this situation, the risk that any one concept can be described by different terms (terms of differing generality or slightly different meaning) is very real.

On the first point, namely to avoid the generation of an excessive number of terms, present thesauri attempt to avoid synonyms through the USE instruction ("use for" (UF) being the reciprocal); thus a single word may serve for several where there is not a significant difference of meaning. We return to this point further in these notes.

On the second point, namely the loose structure, the usual thesauri have terms listed alphabetically, most of which are accompanied by their "broader terms (BT)", "narrower terms (NT)", and "related terms (RT)". For any term X, the narrower term describes "a type of X" (the broader term is the reciprocal). Any term in the same area (or areas) of interest is listed as an RT (with reciprocal entries); the RT form is used rather loosely and often includes near synonyms (the Case-Western Thesaurus of Educational Terms (6) tries to establish rules to cover the related term form).

Taking into account the peculiarities of the building industry (the fragmented nature of information flow) and having reviewed the work already accomplished in other fields, it was decided that the sort of thesaurus of Common Noun Keywords required was one with the following characteristics:-

(i) A minimum set of operational terms, the terms themselves to remain as closely as possible in user language; this can be achieved by a more frequent use of the USE/UF qualification.

(ii) An organised hierarchical structure to improve consistency in indexing, storage and retrieval of information.

In studying these aspects of establishing a controlled vocabulary, one further relationship (the whole term/part term, WT/PT relationship) was found to exist between terms; it is useful to make this clear in the thesaurus, to avoid (a)"straining" the Broader Term/Narrower Term relationship to include relationships which are not properly so described, or (b) "falling back" on the Related Term category, which can be better used. This WT/PT relationship has been introduced between words and a logical series of questions or rules to introduce words into the vocabulary has been established. These are described in the next section.

The purpose of the introduction of the WT/PT relationship is to enable rules for the selection of Common Noun Keywords (CNKs) to be set up more easily and in this sense there are two inter-related advantages. Firstly, an indexer should not be confronted with a large number of "narrower terms" to any one given keyword particularly if the relationships between these "narrower terms" and the main keyword are inconsistent as has been indicated. For example, in traditional thesauri, DOOR HANDLES and WOODEN DOORS are both "narrower terms" of DOORS, yet each bears a different relationship with DOORS. In more complex cases, an indexer may select a word which does not bear the same relationship to the "broader keyword" as the original document suggested and hence control is lost. Secondly, different indexing rules apply to WT/PT relationships than to BT/NT relationships and the application

of these rules minimises the frequency of "nil returns" or "noise". In all cases, we believe, keywords should be as specific as possible while adequately covering the concept. However, when the term chosen is part of a BT/NT relationship, an indexer should - after selecting the original term - progress upwards in the hierarchy; how far up will depend on (a) the word composition of the specific term chosen, and/or (b) whether other specific terms in the same area are chosen. For example, the relationship between FRAMES and CONCRETE FRAMES is BT/NT; if a document is indexed by the keyword CONCRETE FRAMES only, a searcher interested by frames would not retrieve the document unless the original document had also been key-worded with the Broader Term, FRAMES.

This situation does not exist with the WT/PT relationship. An article on DOOR KNOBS is not of interest to the researcher requiring information on DOORS. If a searcher is interested in other parts of doors, then he will select each of them independantly and organize his search accordingly.

To avoid the inclusion of an excessive number of keywords, it was decided to extend the USE/UF form to include words which may be different in meaning but which can be brought together for the purposes of information storage and retrieval. A specialist using this thesaurus could always over-ride the USE/UF form suggested for his own areas of specialisation if greater depth of concept definition is required (this can be called the construction of a "micro thesaurus"). This is done by taking some or all words out of the USE/UF category for a specific subject area, i.e. a set of keywords, and structuring them in the usual way (see next section) so that they bear a relationship with the word which previously was used in their stead. Such structuring can only be successful if the specialist is able to define the words accurately in his own terms. For example, while one recognizes that TOLERANCES and GAPS are different, for the purposes of the general purpose thesaurus, one can describe GAPS as USE:TOLERANCES because it is doubtful that any article including the concept GAPS would not also mention TOLERANCES nor that anyone searching for GAPS would mind being directed through the thesaurus to conduct a search through TOLERANCES. The specialist can, as we have pointed out, reverse this decision and maintain GAPS, and at the same time further introduce such narrower terms as MINIMUM GAPS. The aim is to minimize the risk of doubt when selecting keywords.

Procedures for Constructing a Common Noun Thesaurus

The subject area which was of concern to us, and therefore which was to be served by the thesaurus was defined as "building science and technology". In fact the area was defined by a set of general keywords, selected on the basis of usage and structured according to logical questions (see fig 1a & b). Having established these preliminary general keywords, it became possible to determine the extent of the work (an important question in terms of programming) and to decide upon a plan of action for completing the work in an organised and logical way such that the degree of completion would always be known.

Within this framework, the procedures which have been developed for the generation of a controlled vocabulary can be described as follows:-

(i) A general keyword within the larger framework is selected e.g. ACOUSTICS, MODULAR COORDINATION, SYSTEM BUILDING etc.

(ii) Words are collected from earlier thesauri, glossaries, research reports, books, articles etc.

(iii) These words are divided into groups by, for example, their semantic nature. This is merely to reduce the random list of words into manageable sets.

(iv) Definitions are agreed (specialists are consulted where the research team does not have the appropriate knowledge); note that often, negotiation is necessary on some specific meaning which is declared through the annotations contained in a "Scope Note (SN)".

(v) Words which can adequately be covered by other terms through the USE/UF form are "eliminated" from the processing (note that these words are not "lost" since they will appear in the thesaurus not as keywords but accompanied by the appropriate annotation and its reciprocal.

(vi) Words are picked out in pairs (using "common sense" for guidance) and a logical sequence of questions is run through. This shows whether the terms under consideration have a BT/NT relationship or a WT/PT relationship, whether they are related terms (RT) within the same hierarchy or whether they are not related at all (see fig 1a & b).

(vii) A relationship chart is prepared showing the initial hierarchy (see fig 2a & b).

(viii) Groups of words bearing any of these relationships to each other are re-assembled by means of a computer programme (the purpose of which is simply to display all the words in the area having relationships between them all in clusters); a question is then asked of each of these sets of keywords as follows: "which single keyword describes the ideas contained within this set of keywords?" If the resulting keyword already exists within the hierarchy being compiled, then the hierarchy is as complete as the original set of words allowed. If on the other hand, some of the resulting keywords are not in this hierarchy, a search is made to see if they exist in other hierarchies. If this is so, then relationships between hierarchies are established. Note that if these key-

words do not exist in other hierarchies, it may be because the hierarchies have not been fully developed. Alternatively, the concepts generated by these keywords may fall outside the area of interest as defined.

(ix) If new keywords are produced as a result of re-assembling the original keywords, then steps (iv) - (viii) are repeated until such time as no new relationships appear. The process is thus iterative and the relationships become better defined with each cycle and more comprehensive with the increasing completion of the entire subject area.

(x) Individual cards are prepared for each keyword showing its immediate surroundings in the hierarchy and a revised hierarchical display is drawn (as completely as the advance of the work will allow see fig 3).

(xi) The alphabetical list of keywords is updated to include those just generated.

At the time of writing, the final format for the display of hierarchical relationships has not been designed.

Generation of Proper Noun Keywords

Words which describe other than the factual content of a piece of information may be described as Proper Noun Keywords. The following are categories of Proper Noun Keywords:-

(i) The names of authors responsible for the literature.

(ii) The names of organisations (e.g. universities, publishers, etc.) responsible for the publication of the information.

(iii) The date of publication.

(iv) A geographical reference to the information (usually by country).

(v) A set of Proper Noun Keywords describing the physical nature of the information as follows:

ABSTRACTS
ARTICLES
BIBLIOGRAPHIES
BOOKS
DOCUMENTS
FILMS(MOTION PICTURES)
ORGANISATIONS
PERIODICALS
PUBLICATIONS LISTS
SOUND RECORDINGS
STATISTICS

Whereas Common Noun Keywords have to be selected ahead of time by experts in the field and structured accordingly, Proper Noun Keywords are generated automatically avoiding the problems of selection, definition and structuring. Some cross-referencing is necessary e.g. when abbreviations are used, but cross references can be built into the Proper Noun Keywords either by cross reference cards or search instructions for computer operation. The use of PNKs has already been described, in the context of rapid searches by experts.

Proper Noun Keywords describing the physical nature of the information have already been generated for the abstracts published in IF and it should be noted from the list in (v) above, that it is sufficient to select only one Proper Noun Keyword to describe the physical nature of the information. Often this class of PNKs helps with the actual physical retrieval of the document, since quite probably they correspond to types of storage facilities; indeed it is possible to regard these PNKs as an adjunct to the accession number "addresses" of the documents.

Use of the Thesaurus: Relationship to Other Work

The immediate relevance of this work is to develop a workable Common Noun Thesaurus (i.e. controlled vocabulary of keywords) in our subject area. The hierarchy is not a classification of words (pre-classification anyhow is to be avoided at all costs); it allows the librarian and the searcher to recognise at a glance how general or how specific within and between hierarchies any word must be, and thus guide them in the choice of word that corresponds most exactly to the generality or specificity of the concept under consideration. At the present time, more comprehensive rules for indexing are being prepared.

The thesaurus is needed above all for IF; it would be risky to continue to generate words in isolation. Also it is, in our opinion, a necessary (and unavoidable) first step in the setting up of any wider information services or data banks, which is something else that we are beginning to undertake (see next section of these notes).

It is essential that there should be as few thesauri as possible in use at any one time. Our group has been in contact with other people working on thesaurus generation on the North American Continent and elsewhere, though little work is being done on building industry thesauri. In fact it seems that although building information systems have been the subject of considerable theoretical study, few people have been fighting with the practicalities of operating an ongoing system for the building industry.

One major exception is the Thesaurus being developed for the Canadian Department of Industry, Trade and Commerce - a first edition of which is due for publication at the end of the summer of 1970. Having been in contact with the team working for the Canadian government, we believe and hope that there will be compatibility between our thesauri - at least as far as the allowable word lists are concerned.

Use of the Thesaurus: Information Systems

Our group is constantly aware of the problems of finding information - both in teaching and consultancy work - yet although a consi-

derable amount of research and development results are presently available, few people are in a position to collect them and compile them into a common post-coordination type system based on a commonly accepted thesaurus.

At the time of writing, our team is involved in three levels of initiative in the area of information systems. Firstly, the ground rules have been established for a formal information link between groups at Washington University, St. Louis, the Université de Montréal and the University of California at Los Angeles; this will involve sharing the work of information collection and the preparation of abstracts and exchanging abstracts with agreed keywords. Secondly, it is proposed to agree keywords with several institutions, particularly Laval Université, Québec who are setting up a housing information system (this will allow people to use a common routine for information searches at any of the universities in question). Thirdly, it is proposed to extend our work to include setting up and operating building science informations systems to government agencies through contracts.

As work on developing the thesaurus proceeds, many other advantages of the logical structuring of words are being discovered. Since the words correspond to concepts in the field of building science, the structuring of words implies a structuring of the concepts on the same logical basis. This would appear to offer all sorts of interesting indications in other areas, e.g. when setting up a curriculum or planning a research project or establishing checklists for various stages of design and planning.

Conclusions

These notes have described experiences in the practicalities of information handling - from the early design-publication stage with Industrialisation Forum to the current forays into building science information systems. The critical decision is to use post-coordination instead of any of the creaking classification systems with which the building industry is bedevilled. The critical research phase is then the development of rules for word generation and word selection.

The advantages of using a common thesaurus are many, though in certain circumstances it may be necessary for different users to develop certain areas to greater depth; even so, it is possible to agree "translation" rules to pass from one thesaurus to another, i.e. from one information stores.

Notes

(1) "The Impact of Industrialisation on the Design of the Built Environment", AIA/ACSA Teachers' Seminar 1968, Montréal.

(2) "Industrialisation Forum; Building: Systems Construction Analysis Research", published quarterly jointly at Washington University, St. Louis and the Université de Montréal.First issue October 1969.

(3) Wert, L.,"Information Retrieval and 'Industrialisation Forum'", "Industrialisation Forum" volume 1, number 1, pp. 11 - 17. (Includes a short bibliography).

(4) "DATUM", a system of information storage and retrieval developed by the School of Law, Université de Montréal.

(5) "Thesaurus of Engineering and Scientific Terms", Engineers' Joint Council, New York, 1967, 690 pp.

(6) Barhydt, Gordon C. and Charles T. Schmidt et al., "Information Retrieval Thesaurus of Educational Terms", The Press of Case Western Reserve University, Cleveland, 1968, 133 pp.

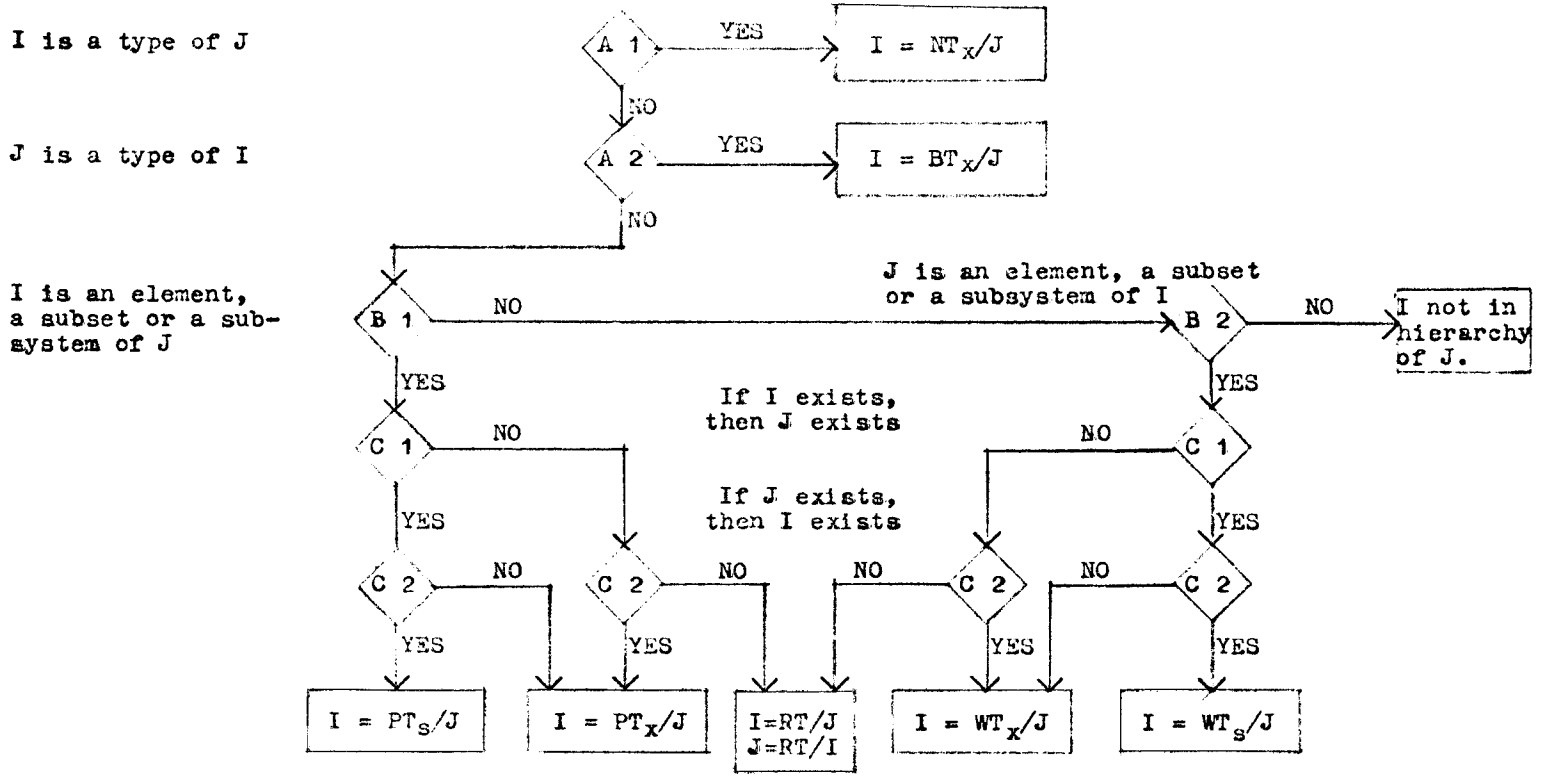

Fig.1a. Question-sequence used to determine what is the relationship between two terms ("I" & "J").

Notation: NT = narrow term, BT = Broad term, PT = part term, WT = whole term, RT = related term.

suffix 's' denotes specific PT or WT (next level in hierarchy), 'x' denotes general NT, BT, PT or WT at an unknown level above or below the term (to be determined by other word pairing).

Fig. 1b Examples of the Question-Sequence.

Example 1.

I = "Modular Dimensions"
J = "Tolerances"
A_1: Are "Modular Dimensions" types of "Tolerances"? No
A_2: Are "Tolerances" types of "Modular Dimensions"? No

B_1: Are "Modular Dimensions" elements, subsets or subsystems of "Tolerances"? No
B_2: Are "Tolerances" elements, subsets or subsystems of "Modular Dimensions"? Yes

C_1: If "Modular Dimensions" exist, do "Tolerances" exist? Yes
C_2: If "Tolerances" exist, do "Modular Dimensions" exist? No

Therefore: "Modular Dimensions" is a whole term (WT_x) of "Tolerances"

Example 2.

I = "Tolerances"
J = "Manufacturing Tolerances"
A_1: Are "Tolerances" types of "Manufacturing Tolerances"? No
A_2: Are "Manufacturing Tolerances" types of "Tolerances"? Yes

Therefore: "Tolerances" is a Broader Term (BT) of "Manufacturing Tolerances"

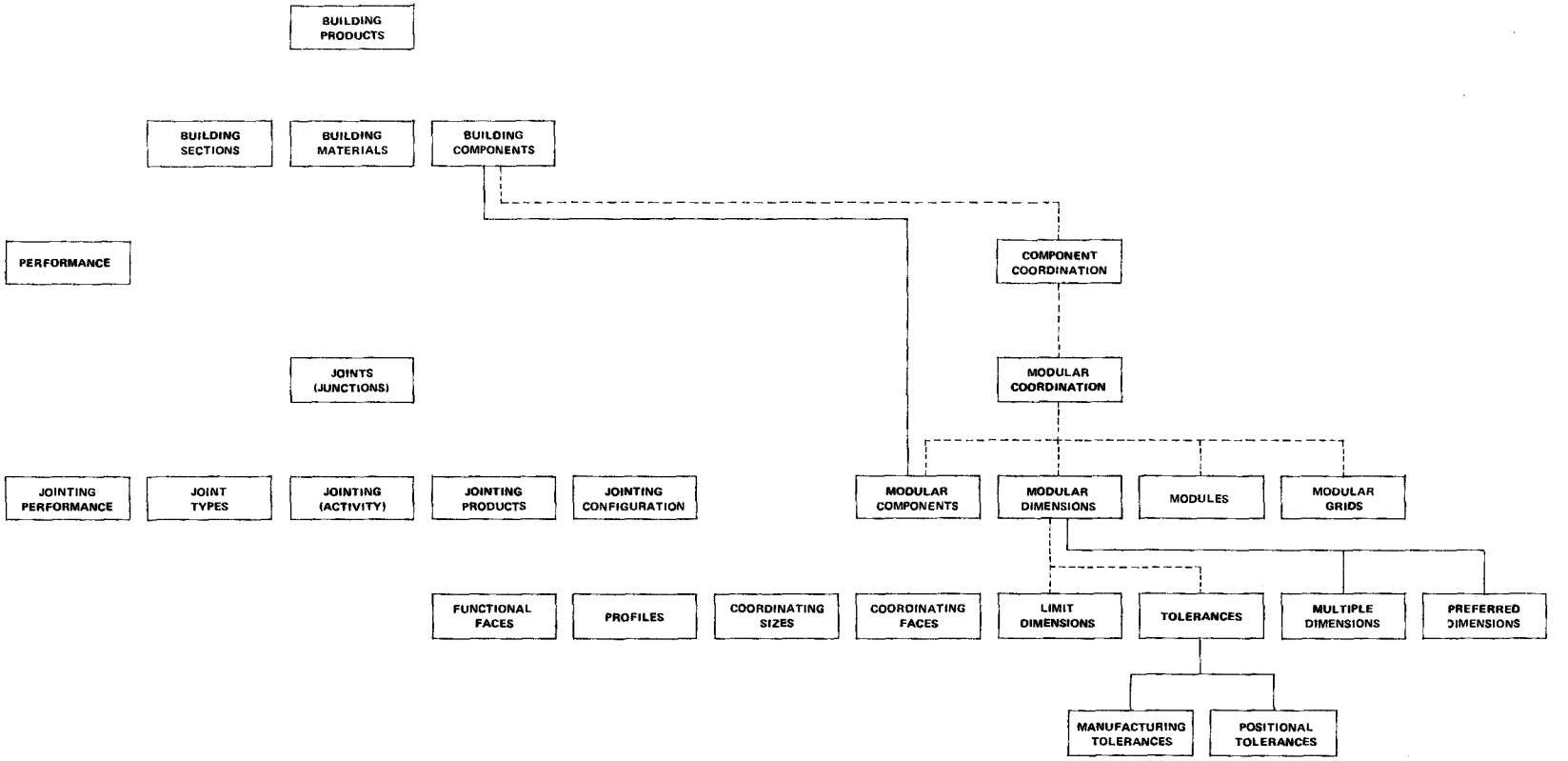

Fig. 2b. Part of a hierarchical display of key-words (in-house working format).

———————— NT/BT relationship

---------- PT/WT relationship

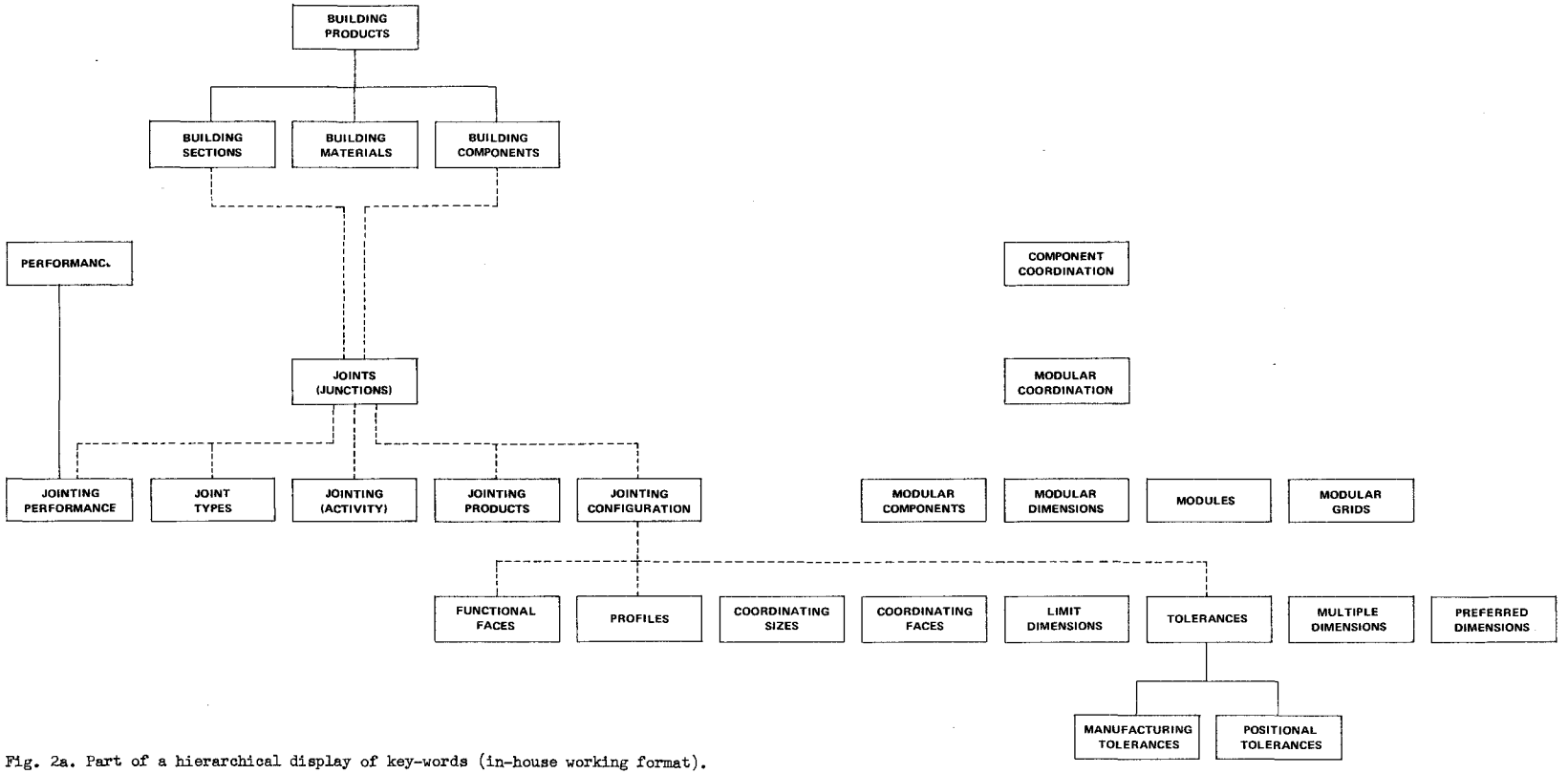

Fig. 2a. Part of a hierarchical display of key-words (in-house working format).

——————— NT/BT relationship

- - - - - - - - - - - - - PT/WT relationship

JT 2	JOINTS(JUNCTIONS)
UF	CONNECTIONS FASTENINGS LINKS COUPLINGS BINDINGS
BT	
WT	BUILDING COMPONENTS BUILDING SECTIONS
NT	
PT	JOINTING PRODUCTS JOINT PERFORMANCE JOINT CONFIGURATION JOINTING(ACTIVITY) JOINT TYPES
RT	

MOD 2	MODULAR COMPONENTS
UF	MODULES IF COMPONENTS MODULAR ELEMENTS MODULAR UNITS
BT	COMPONENTS
WT	MODULAR COORDINATION
NT	
PT	COORDINATING FACES FUNCTIONAL FACES PROFILES COORDINATING SIZES
RT	MODULES MODULAR GRIDS MODULAR DIMENSIONS

Fig. 3. Examples of Thesaurus Entries.