CHAPTER 4

Indexing Procedures

The function of indexing in libraries and information retrieval systems is to indicate the whereabouts or absence of items relevant to a request. It is essentially a time-saving mechanism. Theoretically, we can always find the relevant items by an exhaustive search through the whole collection (assuming that we can recognize what is relevant when we see it). Since this is economically impossible, the size of the store to be examined is reduced by classification, using this term in its very broadest sense, i.e., as the recognition of useful similarities between documents and the establishment of useful document groups based on these similarities. So documents, or document surrogates, are assigned to a limited number of classes according to certain criteria, in particular, their subject content (although in machine indexing, utilizing complete text scanning, this 'limited number' can become very large - as large as the number of significant words used in the text). Search for relevant items is made via these classes (which are classes of documents); only those with a probability of containing relevant items are examined, and the rest (hopefully the vast majority) are ignored. Clearly, we need to know as much as possible of the nature of the classes to be recognised, and the degree to which they allow reliable predictions to be made as to the probability of relevant items being included in them.

Most library indexes, other than those to imaginative works (novels, music scores, etc.) are aimed ultimately at the retrieval of subject information. Even the great Author-Title catalogues, on which so much care has been lavished, serve for the most part the function of a diagnostic classification, i.e., an author's works are sought in the first place because they are about a certain subject and his name is a clue to locating it. The popularity of the author-title catalogue rests partly on its precision in retrieval. Classes determined by authorship or title are mutually exclusive; there is almost no overlapping, no ambiguity about them and requests can be met with 100% recall and precision. But they are useless if the author or title is not known and it is this situation with which IR is mainly concerned. So the classes investigated by this project are those designed for searching by subject prescription only.

There is one exception to this. Bibliographic coupling (including Citation indexing) establishes classes for much the same reason as author-title catalogues, as an oblique way of getting at subject content. Papers which have cited item x are assumed to have some connection with the subject of x. This particular device is dealt with separately in Chapter 7.

The terms which are used to express a request or a search prescription rarely coincide exactly with the terms used to describe a particular relevant document; this is likely to happen only at a relatively broad level, when a request may be answered by a treatise or monograph on the subject. For example, in the test Q.93 read 'What investigations have been made on the flow field about a body moving through a rarified, partially ionized gas in the presence of a magnetic field.' Two documents relevant to this question were

1296 'Waves through gases at pressures small compared with magnetic pressure'
1446 'Waves of a satellite traversing the atmosphere'

In both cases the match is very imperfect. It is made only by recognising that the

class prescribed in Q.93 can be adjusted in order to coincide at some points with the index descriptions of the documents. The class 'rarefied, partially ionized gas' must be seen to correspond or relate, after suitable manipulation, to 'gases at small pressure' in the one case and 'ionosphere' in the other. The class 'body' must be seen to relate at some point to the class 'satellite'.

So a subject index must provide facilities for adjusting and manipulating its classes; it must allow the index classes examined to be expanded or contracted, and in different directions, until a match with the search prescription is recognized. Index language devices are the agents of this manipulation. They are devices whereby class definitions may be adjusted to meet the requirements of different searches.

Index language devices

The index description of a document is a condensed (usually a highly condensed) statement of the document's subject content; it seeks to convey succinctly what the document is about. Its main, and sometimes only, constituent is the set of substantive terms (lexical elements) which act as clues to the subject of the document. These terms may be supplemented by some indication of the relations between them (syntactical elements), e.g., by the addition of roles, or facet indicators (explicit or implicit) or by such elementary syntactical devices as those of the Alphabetical subject catalogue. In a post coordinate index they are usually kept to a minimum, enough to remove serious ambiguity but no more.

It seems reasonable, then, to assume, as the simplest possible form of index description, a bare list of words, selected directly from the title and text of a document as being good clues to its content, and presented without any reference whatsoever to a control list for synonyms, related terms, etc.

The simplest way in which such a list of words could be used would be to regard each word as defining one of the classes to which the document belonged, without reference to the other words. Searches would then be made simply within these classes, separately. For example a document indexed as being about Wakes - Satellites - Traversing - Ionosphere would be seen simply as a member of four different classes (the class 'Documents dealing with Wakes', the class 'Documents dealing with Satellites', and so on). So a search on Satellite wakes would be made simply by examining all documents in the class Satellites, and all documents in the class Wakes. This is very similar, of course, to what a Permuted Title or KWIC index does. Recall performance figures for this crudest of all forms of index language were assessed in the first Cranfield project as 97%. Precision figures were not available but it is certain that they were very low. It is assumed that all the keywords constituting the question are examined. If a selection were made, recall would probably drop in so far as the exhaustivity of the searching would have dropped. The question of exhaustivity and specificity of searching and indexing is discussed later.

Now will be considered the ways in which, by the use of various devices, this simplest of all possible forms of indexing can be refined in order to increase its capabilities for meeting all the demands which search prescriptions may make on it. Such devices may be separated conveniently into two groups;
1. recall devices - those which, when applied to any existing class, increase the size of the class in terms of the documents responding to the definition; e.g., if the class Bakelite is expanded by hierarchical linkage to include all Phenolic resins, more documents are retrieved,

2. precision devices - those which, applied to any existing class, decrease the size; e.g., if we coordinate Bakelite with Extrusion and examine only the class defined by this simple relationship of intersection, we exclude those documents on Bakelite which do not refer specifically to this operation.
The operation of these devices on the sort of simple index description described above can be seen by a consideration of the relations such a description displays to the precise subject of the document concerned, and to the wider subject field of the information store from which we may wish to retrieve that document or something like it.

An index description a, b, c, d, e, f (where each letter represents a substantive term or lexical element, e.g., Wing, Drag, Control) embodies two sets of relations: firstly, those internal to it, reflecting the local and temporary conditions peculiar to the subject of the document described; e.g., the fact that d is the product, whereas f is an agent of the process a which produces it, or, the fact that b qualifies a while c qualifies d, but that neither of these qualifiers is applicable to the object of the other; or, more subjectively, that a and b, rather than c d e or f, represent the dominant theme of the document. These are, broadly speaking, the interlocking relations between the substantive terms.

The second set of relations are those external to it, reflecting the more permanent pattern of relations in the wider field or subject area to which the document belongs: e.g., that a is a species of x, or that c is almost synonymous with q, or that a represents one participle of a term (e.g., Cooling) which may be usefully related to another participle (e.g. Cooled) in the subject concerned.

The two sets of relations can be utilized to add precision to the original description (using the first set) or to expand the description by reference to the wider relations (using the second set). In other words, they underlie the two groups of index language devices which will now be outlined briefly.

Devices which increase precision

(i) Coordination - i.e., the conjunction of two or more terms to produce a narrower class defined by the intersection; e.g. Shear and Flow to give Shear flow. This is the most important device in indexing. Whilst it is commonly associated with postcoordinate systems where it is implemented mainly if not entirely at the search stage, it is equally fundamental to precoordinate systems; but in these, only the products of selected coordinations are usually catered for conveniently.

(ii) Weighting - i.e., the assignment to a term of a figure representing the relative significance of that term in the total subject description of the document. So a term which represents the central theme of the document gets a high weighting and one which represents only a marginal element in the subject content of the document gets a low weighting. If now a question is also weighted, i.e., greater significance attaches to one or some of its terms than to others, then the search may be directed only to coordinations with that term or only to the same term when it has been given a similarly high value in indexing. In either case, the class of documents retrieved is made narrower.

(iii) Links - i.e., indicating a particular connection between two or more terms in a description where the lack of such an indication would produce ambiguity; e.g., if the same document deals with the hardness of copper and conductivity of titanium a link between Hardness and Copper on the one hand and between Conductivity and

Titanium on the other would make it clear which property referred to which substance. Clearly, a link must involve at least two terms.

(iv) Roles - i.e., indicating the role or function of a particular term in an indexing description. Sometimes a simple link is insufficient to remove ambiguity; e.g., Production of particle x by bombardment of particle y. To link, say, Production with x and Bombardment with y could still allow the description to be interpreted as Production by x, or Bombardment of y. The addition of a role-indicator makes the relationship more explicit, e.g. by labelling x as Product and y as Patient (and possibly a bombarding particle, z, as Agent).

Devices which increase recall

(v) Confounding synonyms - i.e., accepting items indexed by x when searching for y, and vice-versa, where x and y are regarded as synonymous.

(vi) Confounding word forms - i.e., accepting items indexed by different forms of the search terms, such as its singular and plural, participle and gerund; e.g., Injectant + Injected + Injection + Injectors. The most comprehensive operation of this device is where a stem or root is used to define the class and all words containing it are included in the class.

(vii) Hierarchical linkage - i.e., accepting items indexed by terms which are in some generic hierarchical relation to the search term. By this we mean terms which are either subordinate to, superordinate to, coordinate with, or collateral with the search term; e.g., the class Cooling might be extended hierarchically to include the subordinate term Sweat cooling, the superordinate term Heat transfer, the coordinate term Heating and the collateral term Radiation. This is a stricter interpretation of hierarchical linkage than is often used in the literature of IR, confining it to the relations between a thing and its kinds as distinct from numerous other relations such as those between a thing and its parts, its processes, its properties, the operations performed on it, and so on.

It is perhaps necessary to note that hierarchical linkage is essentially a recall device. This is not to say that it cannot be used in order to refine a question and give it more precision; e.g., an enquirer about to search the class Cooling might be led to realise (by the hierarchical display of related terms) that he really wanted Sweat cooling and would then narrow his search by confining it to this species of cooling. This is almost as though hierarchical linkage were acting as a precision device. But clearly, the question asked was wrongly put and the function of the hierarchy would have been to assist the question- programming. In testing devices, it must be assumed that the question has been accurately stated, otherwise a large and disrupting variable will enter the test. Therefore such adjustments as the one described cannot be considered, and hierarchical linkage must be treated as a recall device only.

(viii) Non-generic hierarchical linkage. The device of hierarchical linkage which has always featured prominently in subject indexing and is the central device in what is generally known as 'classification' is the result of selecting one particular relationship (the generic one, between a thing and its kinds) as the basis of various kinds of class definition. This raises the question: should we similarly regard the use of other particular relations (e.g., between a thing and its parts, a thing and its properties) as constituting separate indexing devices, each to be evaluated separately?

The importance of the genus/species, or thing/kind relation needs little explanation. Since Aristotle, it has been the central relation considered in traditional logic. In information retrieval its use as an indexing device ('hierarchical linkage') is that it provides an automatic relevance network based on the inclusion relation, e.g., if laminar boundary layer is a kind of boundary layer, then documents on the latter are likely to have some degree of relevance (possibly a high one) to questions on the former. It is also the basis of the notion of 'specificity', which is a fundamental parameter determining precision in index performance. It might be argued that the non-generic relations displayed in all library classifications also possess an 'inclusion' quality so far as documentary classes go; e.g., a precoordinate class with the index description Delta wing - Drag could be said to be 'included' in the class Delta wings just as much as the true species Sweptback delta wing is included. But the inclusion relation of the first example is weaker, in that it is a shared one; Drag on a delta wing is as much included in Wings-Drag as in Delta wings, whereas Sweptback delta wing is entirely included in the class Delta wing. That is to say, the non-generic relation is essentially one of conjunction ('coordination') rather than complete inclusion.

Certainly, no other single relation rivals it in ubiquity or importance for retrieval, although collectively the other relations contribute significantly to class definition in indexing. In precoordinate classified indexes it is normal in question-programming to move from the terms of a category or facet, each one constituting a hierarchy, to the particular term which gives rise to the facet; e.g., to move from Conductivity (of titanium) in the Properties facet of the class Titanium, to Titanium. The object of systematically arranging classes in such indexes is to provide a permanent and constantly available mechanism for manipulating the classes in this way, and not only by movement within a strict genus/species hierarchy. Classes are expanded or contracted by moving also from one category to another. For example, in a question on compressor operation, the class examined may be expanded by moving from particular parts of the compressor (Blade, Shroud, etc.) to particular processes (Stage interaction, Stage stall, etc.) or to particular characteristics of these (Inlet blade angle, Stall limit line, etc.)

In fact, the term 'hierarchical' is frequently used as a synonym for the process of subordination which is the essence of a precoordinate system. In this view of hierarchical linkage a chain such as Delta wing - Sweptback - Transonic speed - Low angle of attack - Lift is regarded as reflecting a hierarchy every bit as much as a true genus/species chain like Metals-Titanium, or Aircraft-Heavier than air-Monoplane-with Delta wing. It is possible to go even further and refer to the precoordination of an Alphabetical subject catalogue (which, of course, subordinates some terms to others in its subject headings, e.g. Aircraft-Design) as a hierarchical system (Ref. 21).

Even if this last extreme view of the term 'hierarchy' is rejected, we are left with the fact that the identification of hierarchical linkage with the full range of relations displayed by a highly organized library classification system does not give us a basic device which can be measured in the same way as can coordination, weighting, etc., but is a varied mixture of relations capable of defining classes.

In traditional library classification it is well established that any relationship may be used to establish 'subclasses' of a given class. In the sense that library classification deals with classes of documents, and that documents on any aspect whatsoever of a subject x can be regarded as belonging to some subclass of 'Documents on x', then Extrusion of plastics is just as much a subclass of Plastics as

Bakelite. But recognition of this is of little use unless the further step is taken of organizing these subclasses according to their particular relations, in terms of subject content, to the class x., i.e., of recognizing that some are the properties of x, some are its parts, and so on.

In a modern faceted classification these relationships are systematically displayed. All terms standing in the same relation to the original class are marshalled together to constitute a category or facet of that class. So, whatever concept comprises the original class, all its Properties are assembled together, all its Operations, all the Agents of these Operations, and so on. So far as they lend themselves to the process, the members within each category are organized in a hierarchy and their relationship within the category is one of a thing and its kinds - kinds of properties of x, kinds of operations on x, kinds of agents of operations on x. But hierarchical linkage in the strict sense, that is generic hierarchical linkage, is established between the terms within a category, not between individual terms from different categories, or between the terms of a category and the original class.

However, it is undeniable that in indexing and searching, classes are manipulated (i.e. expanded and contracted) by going outside hierarchical linkage in the strict sense defined above. The two main paths pursued are those already indicated; firstly to move from a term in one category of a class to one in another category of the same class, e.g. Separation see also Boundary layer control (where Separation belongs to the Process facet of Boundary layer and Boundary layer control belongs to the Operations facet). Secondly, to move from a term in a category to the original class giving rise to the category, e.g. Blowing see also Boundary layer (where Blowing is an operation designed to accelerate the flow in the boundary layer, and belongs to the Operations facet of Boundary layer).

Both the above types of connection imply a more or less definite subject area in which the terms in question stand in some categorical or facet relation. Another type of connection is sometimes recognized, between terms which come from quite distinct areas and which are therefore not considered to have such a facet relation. The 'phase relations' of Ranganathan are one example, and some of the terms connected in syndesis constitute another. For example, a thesaurus might link Automatic control theory with Aerodynamic stability. Generally speaking, the view that this constitutes a quite distinct type of relation assumes a relatively arbitrary map of the field of knowledge in which subjects are assigned to one conventional class or another, such as those found in a general classification. It does not correspond to any fundamental relation between the terms which, if they have any connection at all, can be fitted into the framework of categorical or facet relations. For example, Control may be viewed as a term in an Operations facet of the subject Aerodynamic stability.

It is not feasible to consider syndesis (i.e., the adjustment of classes via a system of linking references) as a device in itself since it clearly uses a mixture of several quite different relations in indicating its further classes. Neither does it seem particularly profitable to make a rigid distinction between the two types of non-generic relations above since it is likely that they will overlap from time to time; e.g., Blowing may well occur in a general Control Operations facet rather than be subordinated solely to Boundary layer; in which case the linkage between Blowing and Boundary layer exemplifies the first type, not the second.

Also relevant to the question of whether these relations rank as discrete devices

is the fact that they are rarely, if ever, used alone; whilst all the devices given earlier may be thus used (at least in a postcoordinate system), the non-generic relations are invariably associated with hierarchical linkage, whether this is via a classified index or a syndetic network of connective references. In a thesaurus, for example, generic hierarchical relations and synonym relations are often indicated separately, but individual non-generic relations are never recognized separately.

We conclude that there is nothing in the practice of indexing to suggest that a separate evaluation of each non-generic relation is necessary, but the collective contribution to index performance of these relations compared with the contribution of generic hierarchical linkage is a matter of some interest, and it seems reasonable to group them together as a comparable device. It may be noted that generic hierarchical linkage is itself an aggregate of several particular relations, just as this group is. The problem is discussed further in the section on Concept hierarchies in Chapter 5.

(ix) Bibliographic coupling is a device for extending a class x (representing the subject of a particular document q) by accepting all, or some of the documents which have cited q: or, by accepting all documents in a particular universe which have a certain number of citations (6 or 7, say) in common with q.

(x) Associative indexing by machine ('clumps', etc.) The possibilities of automatic indexing now being explored by a number of investigators rest mainly on the assumption that classes useful for retrieval purposes can be established on the basis of the statistical characteristics of the index vocabulary (which may in fact approximate to the complete texts of the documents concerned). By using such features as the frequency of occurrence and co-occurrence of individual words and of particular word-clusters, their position in the text, their relative frequency compared with a standard word-frequency list in the subject area concerned, and so on, associations between terms are established which then form the basis of search programmes. The criteria defining the classes to be examined are thus quite different from any of those listed above and therefore the procedure constitutes an indexing device in its own right.

How far it might be feasible to distinguish particular procedures (e.g., the use of one statistical technique rather than another) is as yet uncertain. In particular, the purely statistical methods are in some cases replaced by methods using linguistic analysis, and insofar as these must overlap the 'semantic' devices already described (confounding of word forms, hierarchical linkage of various kinds) they may not merit the status of a discrete and unique index device.

(xi) "L'Unité" system described by te Nuyl (Ref. 22) is a somewhat exotic device whereby a reduced vocabulary is established in a quite mechanical way by lumping together all the terms in a given sequence of pages in the Concise Oxford Dictionary and treating their aggregate as a single class. As is the case with all drastically reduced vocabularies, it is argued that the theoretical absurdities which might arise (e.g., the appearance of documents on Acne in a search for Aconite, or on Conductivity in a search for Cones) do not arise in fact, since subsequent coordination eliminates them.

Any reduced vocabulary may be regarded as a recall device, in that it implies enlargement (by coalescence) of the classes which are formed initially by the individual index terms assigned to a document. Usually, reduced vocabularies are formed

by a mixture of the devices already described - by hierarchical linkage, by confounding word forms, etc. "L'Unité" uses none of these, although the last-mentioned will in fact normally be a prominent accompaniment of its class definition, since different word forms will usually appear in the same alphabetical cluster. So it must be admitted that it forms a discrete index device in its own right.

The above eleven devices may be compared with the list of techniques for controlling index languages given by B. C. Vickery in his book 'On retrieval system theory' (Ref. 9) (see Chapter 1). Of those listed above, three devices, namely Bibliographic coupling, Associative machine indexing, and L'Unité, were not mentioned by Vickery. The others include all the techniques given by Vickery, since a number of these were variants of the more broadly defined devices above.

We have tried to distinguish the basic device itself, as a method of class definition, from the different ways in which it might be implemented in different index languages. The latter may be regarded as different amalgams of the various devices, with further differences resulting from the various methods of file organization.

The different ways in which coordination is applied in precoordinate and postcoordinate systems, have already been mentioned. The fundamental difference is that in the former a limited number of coordinations are made and the resultant compound headings are then filed in linear order, and rules observed (as to citation order, etc.) to allow determination of the exact position of any particular combination. This difference has repercussions for the other devices. Even in a largely single-entry precoordinate system, for example, 'weighting' of an elementary kind is implicit in the restrictions placed on the number of entries which can be recognized. Links are fundamental to a precoordinate system by 'partitioning', e.g., separate entries would be made for Copper-Hardness and for Titanium-Conductivity.

Roles are indicated in a precoordinate system by citation order or by explicit syntactical devices; in a faceted index an index description such as

Wings - High aspect ratio - Drag - Low angle of attack

conveys by its citation order of Thing (Wing, etc.) - Property (Aspect ratio, etc.) - Process (Load, Drag, etc) - Condition (Aerodynamic parameters, Angle of attack, etc.) that High specifies Aspect ratio and Low specifies Angle of attack. In an Alphabetical Subject Catalogue, a similar function is served by such headings as

Children in art,
Pressure vessels - Heat transfer,
Acceleration - Psychological effects.

Synonyms are treated with varying strictness of interpretation in different systems. Often, it is a reflection of the degree of specificity sought, as when one system distinguishes Potential flow from Irrotational flow and another confounds them. The problems of synonymity occurring at the level of multiple term descriptions raises a particular problem for post coordinate systems, since it implies a degree of precoordination at some point in the system; e.g. a Ground effect machine is a synonym for Air cushion vehicle, although there is no synonymity between the individual constituent terms.

Confounding of word forms may be implemented at the indexing stage, via a controlled vocabulary (e.g., Conducting see Conduction), or by search rules (e.g., accept Conducting + Conduction + Conductor). In a precoordinate system, where the different forms are separated in different categories, confounding may be possible only at the search stage; e.g., by consulting the A/Z index of a classified index and observing the variant forms used.

Hierarchical linkage may be implemented in a post-coordinate system by generic posting, by coding based on a form of semantic factoring as in the WRU system, or by search strategies based on a classification schedule or thesaurus. In a pre-coordinate system it may be achieved to a large degree by the file arrangement, as in the systematic juxtaposition of related classes in a classified file; it may be achieved fragmentarily by inversion of subject heading in an Alphabetical subject catalogue (e.g., Drag, Base; Drag, Form; Drag, Induced; etc.) or by search strategies based on a syndetic network of see also references.

Measuring the performance of index devices

In order to establish recall and precision performance figures for the different devices, both singly and in various combinations, it was first of all desirable that we established as far as possible figures for indexing in which none of the devices was operating. Then it would be possible to determine the impact on these figures of the introduction of each device in turn. This assumes, of course, a test collection and a set of questions to be put to it, where it is known just what documents are relevant to each question, as described in the previous chapter.

Performance figures for an 'unindexed' collection seemed to imply a situation in which the complete text of each item in the collection was searched for each question. This would have been too tedious an operation (although something like it, except that it was on a small scale, using computer facilities, has been described by Swanson (Ref.18)). The alternative which we decided to take, was to use, as the base situation, one in which the simplest known indexing device was used and to measure the impact on this of all the other devices. This simplest device was taken to be that of condensation of the full text into an index language consisting solely of the 'uniterms' thrown up by the title and text of the document itself, quite uncontrolled by any prior index language.

So the first step was to establish, by the indexing of the test documents, a crude, elemental index language from which all the other languages (each one characterized by the addition of a particular device or aggregate of devices) would be derivable. Before this could be done it was necessary to provide for the control of two major parameters in indexing, exhaustivity and specificity.

Exhaustivity and specificity

Exhaustivity in indexing refers to the degree to which one recognizes (i.e. includes in the index descriptions) the different concepts or notions dealt with in a document. Specificity refers to the generic level at which these concepts or notions are recognized. For example, suppose a report has as its main theme the subject 'Drag on swept wings at high subsonic speeds'. If one neglects, for the time being, the various subsidiary themes which are also dealt with, this report may be said to deal with three concepts - an aerodynamic characteristic, an aerodynamic structure and a flow condition. If these concepts were described in the above fashion in the index description, this latter would be exhaustive but not specific. If the description consisted only of Drag - High subsonic speeds it would be neither exhaustive nor specific; for whilst the terms retained are specific, the absence of any reference to Swept wings implies that the subject deals with aerodynamic structures in general (some structure is implicit, of course) and this is less than specific, since to be this a description must be exactly coextensive with the notion represented. There can be no reduction in exhaustivity which is not a reduction in specificity; but the reverse does not hold.

An exhaustive and specific description (either in indexing or in question formulation) is one which allows no further qualification or refinement - it is a completely precise statement of the class concerned (or classes, if the terms are considered individually). It seems clear that such a description should include not only the substantive terms or lexical elements (which are the essential and often sole constituents of most index descriptions, at least for post coordinate systems) but also the full range of interlocking relations, or syntactical elements, which convey the exact relations between these terms in that particular description. To take a rather far-fetched example, another report might refer to a high wing at subsonic speeds and unless, in the first example, High is interfixed or linked with Subsonic speed the two different subjects are not clearly distinguished. Unless we are to recognize these syntactic elements as a third parameter in the precise description of a document, they must be regarded as elements in exhaustivity and/or specificity. In the great majority of cases they do not refer to the generic level of substantive terms but reflect non-generic relations; they constitute 'relational' terms, analogous to the substantives. They are used as such in a few indexing systems and theoretically at least can themselves display varying generic levels; e.g., Influence could be replaced by the more specific Harmful Influence. It would seem, then, that these terms reflect both exhaustivity and specificity, but more often the former. Only when the relation is explicitly a generic one (as can be the case, for example, with Farradane's appurtenance operator, Ref. 23) can they be said to determine specificity.

Exhaustivity of indexing

Recall devices cannot be fully tested unless the indexing on which they are tested is exhaustive; otherwise, loss of recall at any point might be attributable to a lack of exhaustivity rather than to the device concerned. So maximum exhaustivity in indexing was attempted, at least as far as substantive terms (lexical elements) were concerned. At the same time, since it was clearly desirable to measure the effect of varying exhaustivity on different devices, it was necessary to note during the indexing which terms would in fact have been omitted if any level less than complete exhaustivity had been acceptable.

This problem was very conveniently solved by using the figures assigned to terms as a weighting device as indicators also of which particular terms would have been accepted at different levels of indexing exhaustivity. For the highest level of exhaustivity all terms would be acceptable, whatever their weight. For the lowest level of exhaustivity only those terms given the highest weight (i.e., those terms which would be regarded as essential even in relatively superficial indexing) would be acceptable.

In the result, on average, 31 terms were used per document, with 3 levels of weighting. (6, 8, 10). If only those terms weighted 8 or 10 were counted, it represented an exhaustivity level of 25 terms per document and if only those weighted 10 were counted it represented 13 terms per document.

Of course, 'complete exhaustivity' is a relative term here; strictly speaking only the use of the full text of the document including diagrams, tables and graphs, constitutes completely exhaustive indexing. But whilst the economics of mechanised aids in indexing may eventually make this feasible and its testing desirable, the degree of exhaustivity represented by 31 terms per document was thought to be a reasonable approximation to what would be regarded, for documents in this particular subject area, as extremely thorough indexing. The problem of syntactic elements ('relational terms') as an element in exhaustivity will be dealt with later.

Specificity of indexing

Precision devices cannot be fully tested unless the indexing language on which they are tested is of maximum specificity. For example, a question on Elliptical cylinders can only be matched specifically if, whenever that concept appears during indexing it is represented by the exact description Elliptical cylinders and not by a more general term such as Cylinders alone. If the indexing is not specific in the first place, there is nothing the searcher can do to improve precision by altering his search programme.

At the level of substantives, or lexical elements, specificity was fairly easy to achieve, since, by adhering closely to the language of the document and indexing exhaustively, it was reasonably certain that the specific subject of a theme or concept would be brought out. Even if the author used a more general term in the title or summary, as was often the case, the specific term would nearly always appear somewhere in the text. For example, the title might refer to a 'Laminar boundary layer', the summary to an 'Incompressible boundary layer' and the body of the text to a 'Steady, laminar, incompressible boundary layer'; the indexing would give Steady, laminar, incompressible boundary layer.

Effect on indexing procedures of methods of measurement

Having provided for the control of the major parameters of exhaustivity and specificity, the problem arose of how the different devices might be added, one by one, to the basic natural index language, so as to allow for their measurement.

Several possible methods of proceeding now presented themselves:

(1) To make one index completely devoid of any devices, and concurrently, to make a number of separate indexes, each one embodying this first index modified by a single device, e.g. one index in which the varying word forms of a term were confounded, another in which hierarchical linkages were established, etc.

(2) To make a device-less index and measure the impact of devices entirely by variations in search programming; e.g., the result of confounding synonyms could be measured simply by programming a search for 'Disturbance' as 'Disturbance + Perturbation' whereby the expansion of a class is achieved simply by making a sum of the constituent parts of the expanded class. Similarly, measurement of the effect of confounding word forms could be effected by programmes such as 'Injecting + Injection + Injector ...'

In comparison with (1), this method, obviously, would be much less laborious clerically, even in the case of hierarchical linkage. It might be thought that this could be best measured by constructing a classified index: but the latter is an amalgam of several devices and the measurement of strict hierarchical linkage in isolation is measurable quite effectively by such programmes as: Wave + [(Wave x (N + Standing + Blast + Shock)] to expand an initial class like 'N Wave' to the generic containing class 'Wave'.

Of course, such search programmes required the compilation of code dictionaries - of synonyms, of word-forms, of hierarchies (i.e. of classification schedules). But the indexing itself, so far as these recall devices were concerned, could be done without any regard to these devices whatsoever, since the relations concerned did

B 1590	AUTHOR STONE, A. TITLE Effect of stage characteristics and matching on axial flow compressor performance	Indexer JT.	Date 17-6 63
Base Document A137	REFERENCE Trans. American Soc. of Mechanical Engineers 80, 1958, p1273		

THEMES (partitioning)		CONCEPTS (Interfixing)		CONCEPTS (Interfixing)		TERMS & WEIGHTS		TERMS & WEIGHTS	
A	cd effect of a use of ef	a	Stage characteristics	t	Range of operations	Stage	10	Blade	5
B	——— b ———	b	Stage matching	u	Stage flow coefficient	Characteristic	10	Range	8
C	cdh	c	Axial flow compressor	v	Mass flow	Matching	10	Operations	8
D	cdi effect of j with k	d	Stage performance	w	Choking flow coefficient	Axial	9	Mass	8
E	cdl effect of g	e	Test data	x	Surge line	Flow	10	Choking	8
F	cmp effect of n	f	Analysis	y	Change in slope	Compressor	10	Line	6
G	——— vo	g	Mach number	z	Knee double valued performance curve	Performance	9	Slope	6
H	cmdq	h	Velocity distribution	aa	Unstalling hysterisis	Test	9	Knee	6
I	cmdr	i	Temperature co-efficient	bb	Inlet guide vane stagger	Data	9	Double	6
J	cmdt effect of sg	j	Flow co-efficient	cc	Uprating stage one	Analysis	9	Curve	6
K	cmv effect of g v	k	Constant flow angle	dd	Uprating stage two	Mach	7	Unstalling	6
L		l	Cascade losses	ee	Blade stagger	Velocity	7	Hysterisis	6
M	cbv effect of w	m	Idealised compressor	ff	Stage loading	Distribution	7	Inlet	6
N	cbx effect of o	n	Total pressure ratio	gg	Annular area	Temperature	6	Guide	6
O	cbxy effect of o	o	Percentage of design speed			Coefficient	6	Vane	6
P	cbz effect of aa	p	Performance			Constant	6	Stagger	6
Q	cv effect of bb	q	Stalling point			Angle	6	Uprating	6
R	——— cc	r	Compressor surges			Cascade	6	One	6
S	——— dd	s	Pitch line blade speed			Loss	6	Loading	6
T	cx effect of ee					Idealized	8	Annulus	6
U	——— ff					Total	5	Area	6
V	——— gg					Ratio	5	Number	7
W	——— vq					Percentage	8	Line	6
						Design	8	Valved	6
						Speed	8	Two	6
						Stalling	8	Pressure	7
						Point	8		
						Surge	8		
						Pitch	6		

FIGURE 4.1 INDEXING SHEET FOR DOCUMENT 1590

not depend on the context of the individual report being indexed but on the context of the index language as a whole; in other words, the relations were paradigmatic rather than syntagmatic. It may be noted that all the devices measurable by Method (2) are recall devices - i.e., devices for elaborating or expanding the classes given in the index description of a document.

(3) To incorporate particular devices in the original indexing but in such a way as to make them detachable when required; e.g., to attach weights to terms which could be counted when figures for weighted indexing were required but ignored for unweighted indexing. This method, also much less laborious than Method (1), would be particularly appropriate for those precision devices which depended on the context of the individual report being indexed, and which could not therefore be measured by Method (2).

It was finally decided that the indexing proper should be done on the basis of Method (3); that is to say, the indexing would be basically postcoordinate, and take into account only the precision devices of weighting, links and roles (whilst observing a high degree of exhaustivity and specificity). Method (2) was to be used in the measurement of the recall devices of Synonyms, Word-forms and Hierarchical linkage (generic and non-generic). Associative indexing could not be measured within the conditions above, but it was hoped that the indexing would be sufficiently exhaustive to allow some tests of associative techniques to be made by other investigators. Te Nuyl's device was also ignored at this point, since it was clear that our indexing language could always be translated into dictionary-based clusters when necessary for measurement by Method (2). Bibliographical coupling, since its classes are not defined by subject descriptions, required quite separate measurement, and is discussed in Chapter 7.

The major precision device of coordination is, in a postcoordinate system, purely a search device and its measurement does not fit exactly into either Method (2) or (3). It is perhaps necessary to mention at this point that post coordination in itself does not constitute an indexing device. It is essentially a method of recording subject descriptions in a physical form which allows equally free access to whatever combinations of terms are requested. A precoordinate system, on the other hand, allows direct access only to certain selected combinations of terms. Other combinations constitute distributed relatives and access to them is to this extent made less convenient (although it is by no means forbidden, or made impossible, as is sometimes suggested). But the class defined by coordinating two or more terms is exactly the same, whether the operation is performed at the indexing stage (precoordination) or at the search stage (post coordination). The relative convenience with which access to such a class is gained was not something with which this investigation was concerned.

The form in which the indexing was recorded is best shown by Fig. 4.1 which shows the index sheet for document 1590. Author and title details were printed on the sheet. The indexer then analysed the document in four stages: firstly, 'concepts' were distinguished as a first-stage interfixing device ('link'); these are not easily defined (and this must be recognized as a theoretical weakness) but their practical function was reasonably clear. This was to remove the first level of vagueness and ambiguity inherent in words taken singly, by not accepting adjectival forms alone but only in conjunction with the terms they qualified. So terms which in isolation are weak and virtually useless as retrieval handles were given the necessary context; such terms as High, Number, Coefficient, Main, Trailing, Angle, Aspect which in practice do not form classes for which requests are made, appeared in conjunction

with other terms, to produce meaningful class terms - e.g., High subsonic speeds, Mach number, Pitching moment coefficient, Main wing, Trailing edge, Low angle of attack, High aspect ratio. Not only 'weak' terms were combined, however, for the interfixing function of concepts often produces phrases whose constituent terms are quite potent, index-wise even in isolation. For example, Cruciform wing, Circular body, Tail fins, Span loading theory, Force divergence Mach number, Wing vortex field, Rectangular wing surface, Wind tunnel wall. Such combinations make it clear, for example, that in the one document Surface relates to Rectangular wing, not Cruciform wing; that Mach number relates to Force divergence rather than to some other phenomenon; that Low relates to Angle of attack and High to Aspect ratio (and not vice-versa). Or, as in Document 1590 that Distribution relates to Velocity and Ratio relates to Total Pressure, and not vice-versa.

Secondly, the concepts were now grouped into a second-stage link device in order to display the distinct 'themes' into which the document could be partitioned. 'Partitioning' of a document is a well-established procedure in traditional precoordinate indexing and is often referred to as analytical cataloguing. Owing to the exigencies of space in the precoordinate index, such analysis is usually confined to items in which the constituent chapters, sections, etc. can stand alone; examples are symposia of various kinds, festschriften, and collections of plays. In such cases, 'standing alone' could be interpreted almost literally in the sense that each theme is dealt with in a distinct, self-contained physical section of the document. In such circumstances, partitioning could allow greater recall (resulting from greater exhaustivity of indexing) with almost no loss of precision. This was rarely the case in the aerodynamics reports constituting the test collection. In these, a particular concept might run as a thread throughout the document, appearing at different times in different contexts. So themes were not necessarily, or usually, mutually exclusive. This diffusion of various concepts throughout a document seems to be an important cause of many problems in retrieval and particularly that of the inverse relation between recall and precision; for a document whose index description contains all or most of the terms of the question prescription may yet feature those terms in an unacceptable pattern.

The example in Fig. 4.1 (Doc. 1590) demonstrates the salient features of the two stages of linking embodied in the indexing. To economise in the writing down of themes, the concepts were labelled with lower case letters and only these appeared in the themes. Where the relationship between concepts appeared to be potentially ambiguous, it was indicated in an elementary fashion by verbal quasi-role devices such as 'effect of', 'by means of' or 'use of'. So the first theme of document 1590 is to be read as Axial flow compressor - Stage performance - Effect of Stage characteristics - Use of Test Data - Analysis.

Normal practice was to give as the first theme (or first few themes where necessary), the general subject of the document considered as an integrated whole. Subsequent themes would then bring out the particular subjects which made up the whole. In document 1590, for example, themes A and B jointly represent a formal statement of the title, with the addition of Test Data and Analysis. It also demonstrates a fairly common situation whereby the title provides a reliable and succinct statement of the document's general theme.

The third step in indexing was to give weights to the concepts (and subsequently to the terms) - i.e., to allocate to each concept a value indicative of its relative importance in the document. Such a value can be regarded as an assessment of the probability that, should the concept concerned happen to be the subject of a question,

the document indexed would be of some relevance to it. It might be argued that all indexing which involves a selection of terms to describe a document (and that is all indexing - even the use of full text by a computer, with deletion of articles, prepositions, etc.) implies weighting in that the use of a subject term indicates a reasonable probability that the document is of some relevance to questions on that subject whereas the rejection of a term indicates that the probability is low or non-existent. Weighted indexing extends the range of values from two (worth using, not worth using) to whatever number of different weights are recognized. For example, if an index description contains the terms a b c d e, and weighting is assigned to each term in the scale

3. Most important terms
2. Less important terms
1. Least important terms

If a, b and c are now each weighted 3, while d and e are each weighted 1, the implication is that the probability of the document being relevant to a question a b c is that much greater than the probability of its being relevant to a question on d e.

It has already been noted that the weights given to terms could be used as the basis for measuring exhaustivity i.e., when a figure for a high level of exhaustive indexing was required, weights could be ignored and all terms regarded; when a figure for less exhaustive indexing was required, those terms which had low weights could be ignored and treated as though they were not indexed. It should be noted, however, that the use of weighting as a measure of exhaustivity is purely an evaluation technique and plays no part in normal indexing, for in the latter, weighting only comes in as a device when it is applied also to the question. Then a question term with a given weight will accept as a match only those index terms which have the same (or a higher) weight. This procedure inevitably alters the boundaries of the classes defined in searching and proves weighting to be an independent index language device and not just a reflection of exhaustivity.

The rejection of an index term as irrelevant is now performed at the search stage, whereas exhaustivity of indexing is decided, of course, at the indexing stage. For example, suppose a question containing terms a b c d e, and a relevant document which has been indexed with weights as a^3 b^3 d^1 e^2 g^2 etc. If we were simply measuring the effect of exhaustivity, we would say there was a match of four terms (a b d e) when indexing was fully exhaustive (all weights accepted). If terms with the lowest weight (1) are now ignored, the match is reduced to three terms (a, b, e); if only the highest weight of terms is accepted (i.e. the lowest level of exhaustive indexing), then the match is reduced to two terms (a b). Here we have been using weighting purely as a measure of exhaustivity of indexing, but to consider its effect as a precision device, assume that each term in the question is weighted, and that the search specification is now a^3 b^3 c^2 d^1 e^3. The term c is rejected because it does not appear in the indexing; e is rejected because the search requirement is for a term with a minimum weighting of 3, whereas in the document e has been indexed with a weight of 2. This now means the class accepted has altered to a b d, whereas with variations of exhaustivity it was respectively a b d e, a b e or a b.

As to the problem of how to weight, two general approaches seemed possible and both were made. Firstly, for three hundred documents, weights were assigned on a quasi-statistical basis, dependent on the sections of a document in which a term appeared. If a term appeared in the title, the summary or abstract, the introduction, the body of the text, the conclusion, and the list of cited references, it received a

weight of ten. If it failed to appear in any one of these positions its weight was reduced by one point for each failure to appear. Since the indexing was strictly according to the natural language of the document this meant that all terms received some weight insofar as no terms were used which did not appear somewhere in the document.

Secondly, for the rest of the collection, weights were assigned subjectively. In most of the literature on weights, notably the pioneering articles by Maron and others (Ref. 24), weights were assigned subjectively to individual terms. This proved unsatisfactory; e.g. in a document in which 'Low aspect ratio wings' constitutes a central theme, can the single terms Low or Aspect or Ratio possibly be regarded as crucial in themselves? The significance of a term in a document is very often lost if the term is robbed of its context. So weights were assigned to concepts and the individual terms within a concept received the weight given to that concept.

A range of six different weights was again adopted. This time it attempted to combine a measure of the importance of the concept in relation to the total message of the document (which is another way of referring to the document's probable relevance to a question entailing the concept) with an assessment of its significance in retrieval terms, i.e., its potency as a retrieval handle in the particular collection indexed. The subject significance was measured by reference to a trio of values: these assume that a document can normally be regarded as consisting of its integrated subject as a whole (its main theme), together with one or more subsidiary themes, which may vary in importance from quite major component themes to quite minor, marginal themes. This assumption is a simple extension of the analysis into themes already described as 'partitioning'. According to the status of the theme in this scheme it received weights between 9/10, 7/8 or 5/6:

Weights

9/10 For concepts in the main general theme of the document
7/8 For concepts in a major subsidiary theme
5/6 For concepts in a minor subsidiary theme.

When assigning the weights to the individual terms, the higher weight of the pair assigned to the concept concerned was used if the term was considered to be a very potent one; potency here was regarded as a mixture of word frequency in the total collection (indicating roughly the generality or otherwise of the request in the context of the particular collection), and 'concreteness', whether the term was likely to be requested as the focal point of a subject or whether it was too vague to be the object of a direct and separate request. For example, in Document 1590 (Figure 4.1) the concepts of Themes A and B were each allocated the top weight; the individual terms which were considered potent (e.g. Stage, Matching, Compressor, etc.) received a weight of 10; Flow (on the grounds that it was a very common term) Test, Data, Analysis (on the grounds of vagueness) received the lower top weight of 9.

One small point to be noticed is that an individual term was always given the weighting of the more heavily weighted concept if it appeared in more than one concept. In the example, concept m is Idealised compressor, and the concept is given a weight of 8. However the term Compressor has previously occurred in concept c, Axial flow compressor, for which it received a weight of 10, and this it retains.

The fourth step in the indexing was to write out the terms individually and attach the weights to them as explained above.

The fifth step was to assign roles to the terms. At this stage a temporary failure of plan has to be recorded. If roles were to be assigned, it was obviously desirable to assign them at the time of indexing. Preliminary analyses were made of a number of complete indexing descriptions in order to establish the major variations of role occurring amongst the terms and the degree to which this might lead to ambiguity.

It was stated earlier that the main function of roles would seem to be the removal of a certain type of ambiguity beyond the power of simple links to remove. Links, it should be remembered, simply assert that some relation exists between two or more terms - between a and b, say, rather than a and c. A role states what the relation is.

From the preliminary analyses, it became apparent that the roles developed by Western Reserve University, American Institute of Chemical Engineers and by DuPont were not appropriate, designed as they were for subject areas (chemistry and applied chemistry) in which the appearance of the same term (e.g. a material) in significantly different roles was a fairly frequent experience. The Cranfield investigation of the W.R.U. index had already shown some of the drawbacks associated with roles, even in those fields for which they seemed particularly appropriate, and our analyses confirmed these. A major problem was the fact that when a term plays one particular role in an indexed document, it does not necessarily make that document irrelevant to a question in which the same term features in a different role: for example, a document on the 'Use of mufflers to control the sound of jets' suggests the roles: Muffler (Agent of operation), Control (Operation), Sound (Product), Jet (Cause). If a question were now asked on the 'Effect of mufflers on jets' the role of Muffler might well be designated as Influencing factor (cause) and that of Jet as Thing affected; the relevance of the document to this question would be obscured by the different roles assigned to the otherwise matching key terms. Such a situation implies that several quite different roles might be acceptable in searching; but this comes dangerously near to negating the whole idea of roles. Such, in fact, was the situation in the W.R.U. test, when the roles were frequently ignored in W.R.U. search programmes, presumably because of awareness of the danger of demanding too close a match.

An alternative plan which seemed to be suggested by the kind of example given above was that roles should not be assigned purely according to the relations between the terms in the document or question concerned (their syntagmatic relations), but also according to a wider picture of the relations between the terms in a subject area (their paradigmatic relations). It is generally recognized that the organization of terms into facets is closely related to the provision of role indicators. In special faceted classifications it is not uncommon to find a term appearing in more than one facet, the difference being due to the difference in role played; e.g., in a classification for pharmaceuticals, the same substance might appear as a Product, a Substance Extracted, an Agent of a reaction, or as an Agent of an operation. In such a system, where prior analysis of the terms of vocabulary has been undertaken, the more enduring relations (that a problem or by-product in the propagation of jets is the production of noise, which demands control, and that mufflers are one agent of control) would be recognized and the fortuitous alteration of roles suggested in the question would not have been allowed to obscure the situation.

Consequently, a set of roles was developed along these lines so that they closely reflected the categories which would be distinguished in facet analysis of the field. But trials of these (i.e., examination of a number of indexing descriptions in order to see whether ambiguities would be removed by the roles) showed that they left untouched what was probably the commonest problem of ambiguity in the vocabulary

being developed. This was the problem of knowing what terms were qualified by what adjectives. The example of 'High aspect ratio wings at low angles of attack' has already been cited. Without some interlocking device (links or roles) this document description would respond to a question on 'Low aspect ratio wings', or 'High angles of attack'. Similarly, an index description 'Transient aerodynamic heating of external loads' might respond to a question on 'Transient loads' or 'External heating', and 'Compressible flow over an adiabatic wall' might respond to a question on 'Adiabatic flow'. However, the situation is typically one which is solved by links rather than roles, since the type of relation is not in dispute, only which of the two (or more) terms are to be linked.

If such modifiers were to be regarded as roles the only solution which would remove all ambiguity would be to give them the same role as the term they modified, and then it would work only on the assumption that links were used as well. This was therefore tried (using such roles as Specifier of agent, Specifier of structure), but only at the cost of duplicating the linking function already performed by partitioning and interfixing.

One of the few situations, in the literature indexed, in which a concept (rarely a single term) might feature in a role which appears to differ significantly from its usual one is that in which a body of data is used as an agent in the investigation of some other phenomenon (to which it therefore takes second place in the document concerned). For example, in B1204, Two-dimensional airfoil section data are used as a means of studying high subsonic speed characteristics of swept wings. But closer examination throws doubt on the necessity for using a role even here; had the section data been referred to in some other way (as a parameter influencing aerodynamic behaviour, say) its essential relation to the primary subject (swept wings) would not appear to have been altered.

It was stated earlier that in transferring concepts to themes (the second step in indexing) the exact relationship was sometimes indicated for clarity. The relative infrequency with which the need to do this arose (and Document 1590 was not typical here) was itself a warning that the field of high speed aerodynamics was not likely to be very fruitful in the evaluation of roles as a device. This impression was reinforced by our preliminary analyses. Since, then, any measure of roles was likely to reflect an uncongenial and unresponsive test environment, the very considerable effort involved in developing a set of appropriate roles, applying them and measuring the impact, seemed of doubtful worth. For this reason the preparations for the evaluation of roles as a device were temporarily abandoned.