Putting the horse before the cart: rapid access to data banks by the ‘SIGNPOSTS’* method

Audrey M. Adams

Document Image Processing (DIP) provides an efficient computer storage of information. Efficient retrieval of information is equally important and requires good indexing. This is a semantics problem only human indexers can solve. Data banks even more than books require good indexers to provide efficient access to information. ‘SIGNPOSTS’ provides a matrix for establishing a fully articulated index for databases.

Document image processing (DIP) not only offers the ‘paperless office’ as a feasible possibility but provides efficient storage of information for businesses, libraries, research institutes, indeed anyone faced with a large and growing information storage problem. American businesses alone generate more than 1 trillion pages of paper documents a year.1

The latest personal-computer-based DIP systems appear to offer a cost-effective means of storing information. But computer storage is only half the problem solved; equally important is efficient retrieval of information. How is relevant information to be recalled on demand by various users wanting differing access to the same data? The problem is a semantic one and the solution lies in an effective indexing system.

Problems of information retrieval

The problem is not new. It exists with printed information and has been met there individually, by back-of-book indexes, cumulatively, by subject guides to periodical literature, by abstract services and the Library of Congress and British Library subject catalogues with varying degrees of precision. All of these manual systems face practical limitations. Back-of-book indexes are restricted to space allocated by editors. Library catalogues and abstracting services handle large amounts of information and are based on a compromise. In theory, they could index to any level of specificity (precision), but space and time impose limitations. These systems have based their indexes or subject categories on ‘implied greatest use’, that is, to answer a restricted range of preconceived questions. If a query does not fall within this range, it can be answered only by a prolonged search through the text of books and periodicals. For example, a subject classifying system may list an article, among others, under ‘religion’. The article itself may also contain references to many items not of religious significance. How does the reader pick out the relevant ones, e.g. the Roman gold ring engraved with a ‘chi-rho’ and the patera engraved with an ‘alpha/omega’? Subject classifying helps only to a certain level, depending upon the subject knowledge of the user. For those users without detailed subject knowledge, they are a signpost to nowhere. An information retrieval system is efficient only when it enables users with no or varying subject knowledge to retrieve all the applicable information quickly.

The principles of indexing information are the same for a manual or a computer system, but the computer makes precision indexing and fast retrieval a practical possibility. Precision or specificity and rapidity of access to information should be the essential criteria for indexing. Allied with precision is cross-referencing. These are the means of extending rapid access to information to as many users as possible. In other words, precision and cross-references provide the different routes to the same information. Technically speaking, these two components constitute a fully articulated rapid access index.

The problem of information retrieval from computerized databases has been recognized, and various solutions tried. Before full-text storage was feasible for many users, abstracts and bibliographic details were recorded with controlled keywords the means of access. Most of these systems were based on some form of Preferred Terms (PT), Broad Terms (BT), Narrow Terms (NT) and Related Terms (RT), that is, manual systems applied to computerized information.

PT restricts the use of keywords and employs the device of ‘Use for’ (UF) to terms. As applied to listings for medieval abbeys’ hospitals, the indexer would assign ‘hospital’ as the PT and UF to the synonyms of hospital:

* ‘SIGNPOSTS’ © 1981 A.M. Adams, All rights reserved.

The Indexer Vol. 18 No. 1 April 1992
As the other terms are not allowed as index search terms, the user is faced with two difficulties. First he has to consult a manual, or on or off screen, telling him which word he is allowed to search on. Secondly, when he gets his bibliographic printout, the articles retrieved may not use the word ‘hospital’ but ‘almshouse’, ‘bedchamber’ or any of the 11 synonyms, and he may not recognize these words as meaning ‘hospital’. Therefore he has been given a signpost to nowhere. He has to return to the manual to understand his references.

There are also difficulties for the indexer using this system. Who chooses the PT and what criteria are in play so that a new indexer can use the system and produce similar results? And how are new terms added? Is the PT an arbitrary assignment depending on the bias of the indexer which may change with the next indexer—not a system at all, but idiosyncrasy? Similar objections can be raised in the use of BT and NT. In other words, it is a time-consuming operation for the indexer to work out the ramifications, and requires a constantly changing manual to use. The enquirer trying to retrieve data may be confused rather than enlightened if he does not understand the indexers’ assignments of UF, NT, BT and RT.

Some software systems offer syntax-driven retrieval, that is, ‘proximity operators’. These allow the user to specify the word/term searched in relation to a second word/term within a specified proximity. They are sometimes combined with a ‘wildcard’ search facility. To retrieve any relevant data the searcher needs a considerable subject knowledge. These systems confuse syntax with semantics and are not able to connect or retrieve all the variant terms used to express a given concept.

Classification systems

Some classified systems like SHIC (Social History and Industrial Classification) try to deal with the semantic problem by constructing hierarchies based on assumed function and assigning a numerical code to keywords. They are of no use when function cannot be assigned to terms, so such terms wind up in meaningless ‘miscellaneous’ categories. As the hierarchy is rigidly assigned, the system cannot allow for hierarchies that may shift. For example, in recording archaeological evidence, a Roman hypocaust may be part of a villa or of a bath-house, or the hypocaust alone may remain and cannot be dealt with as it is not attached to a larger unit.

When the ‘780’ (music) classification of the Dewey Decimal system recently was altered to improve access for subjects, the result was a reclassification of music, thereby resulting in costly reclassification for users of Dewey. This is the latest in a series of costly reclassifications. Previous ones took place in Sociology (301–307), in the Political Process (324), in 308, 309, 311, 312, 396, 397 and others.

As these classified schemes are based on past and present knowledge and ways of looking at things, they cannot accommodate easily or flexibly future developments, discoveries, attitudes, interpretations, etc. Very often, therefore, classifications have to be arbitrary rather than specific. For example, in both the Library of Congress and the Dewey Decimal classification systems, ‘education’ is dealt with chronologically, i.e., categories assigned under preschool, primary, secondary, further education, etc. Where then to classify a book that deals with how to prepare a course on any subject across the age range?

Apart from classification difficulties for the cataloguers, any system using numbers as entry to data is subject to a high error factor. Telephone numbers were limited to seven digits as the highest number the average person could accurately remember. A ‘sickle’, using the SHIC classification scheme, is listed as 4.11.3421, requiring seven digits in a complicated decimal arrangement. Even then it is not word-specific, as ‘reap-hook’ and ‘scythe’ share the number. Even in expert and experienced hands, it is possible to transpose a number and get quite the wrong result if the transposed number is a valid one for the computer. Errors in numbering systems are difficult to spot. One such transposed number in the British Library’s cataloguing system (published as the British National Bibliography) misplaced the English composer Edward Elgar in ‘sexual aberrations’. (Long since corrected!)

Numeric or alphanumeric classification, however unsatisfactory in practice, is required to place books on shelves. Using it to retrieve information contained in books or data banks is to confuse location of materials with indexing of contents.

Document image processing

Document image processing (DIP) is seen as doing away with these unsatisfactory indexing methods. By storing full text online, indexing needs are said to be met by allowing every word to be a keyword. Each user then becomes his own indexer. This solution raises false hopes and ends in frustration for the user. His retrieval rate is low and slow, usually including false hits because he is restricted to the words he knows. Language varies not only historically and regionally but is constantly changing, e.g., ‘acquired immune deficiency syndrome’ was soon popularly shortened to an acronym, ‘AIDS’. But to a computer which can only match identical or root words, they appear to be two different diseases. Enquirers asking for ‘AIDS’ will miss all the references to ‘acquired immune deficiency syndrome’ and all its later synonyms, ‘human immune deficiency 1’ and ‘human immune deficiency 2’ and their acronyms ‘HIV1’ and ‘HIV2’. With technical subjects, e.g., botany, terms may be written in Latin or English and the English terms further complicated by regional varia-
tions. For some, the term is ‘pansy’, for others, ‘heart’s case’ or ‘Johnny jump-up’, etc. In the abbeys’ hospital example noted above, there are 11 synonyms for this word. Unless the user knows all the variants of a given term or concept, he cannot retrieve all the relevant data.

This semantic problem was described by Dubois in 1983 in an article on the ‘COFFEELINE Database’ where he noted that a researcher asking an unindexed database for ‘instant coffee’ missed all the more numerous references to the trade synonym ‘soluble coffee’. The more complex the data, the more references will be lost to the user who is his own indexer.

**Semantics**

In other words, full text online does not do away with the semantic problems involved in information retrieval. It brings them home with a vengeance. This is not a machine problem but one involving human intellect and diligence. Efficient access to data and a good retrieval rate depend upon a fully articulated index. Such an index can only be constructed by hard intellectual effort applied by someone with an analytic ability, one with subject knowledge and with knowledge of user requirements, known and projected. For example, some users of archaeological information may want to query at the broad subject level, i.e., any references to ‘Roman religion’ or ‘Roman metalworking’. Others may want specific information, ‘Roman gold rings’, others may be looking for use of decorative techniques, ‘Roman engravings’. Therefore, a record of a ‘Roman gold ring engraved with a chi-rho’ would be relevant to these four very different queries. A fully articulated index would meet all four requirements, that is, provide the varying routes to the same data. All four printouts would include bibliographic references of the ‘Roman gold ring engraved with a chi-rho’. Unless all the descriptors can be listed, cross-referenced and recalled, the indexing/retrieval system falsifies information by omission and also loses data.

There is a second and vital consideration in understanding and therefore allowing for users’ needs. A lifetime of reference work in various libraries and setting up of libraries and information systems made me realize that many people seeking information are so concentrated on the problem that they do not always recall the exact words or terms they need for access to data and the hoped-for solution. But if they are given all the aspects of a word or concept, they can readily identify the one(s) they want. It’s like a multiple-choice question. Given the possible answers, you may recognize the correct one, but would not have been able to provide the correct answer to a straight question. Good indexing provides valuable clues and sets the data detective on the route to a solution.

**The ‘SIGNPOSTS’ method**

‘SIGNPOSTS’ is the outcome of this experience. It is an efficient user-driven indexing/recording/cataloguing system for a computer database on any subject. ‘SIGNPOSTS’ can be used for full text online or for index or abstract services. It features a unique dictionary matrix, a recording matrix, an indexing/recording manual, a users’ manual, a printed thesaurus of subject-specific terms and a requirements specification for computer programming.

The ‘SIGNPOSTS’ dictionary matrix provides the means for constructing a fully articulated index that allows access at any level of subject knowledge. Its basic assumption is that the user knows nothing. To access data, users need only to be literate. They do not need computer skills, special jargon or specialist expertise. This easy accessibility for a range of users is provided for by linking common usage words to technical terms and for retrospective catalogues, historic to current terms and all relevant cross-references. It is also an ‘open’ system that allows for future addition and modification relative to new findings and knowledge. This is possible because the dictionary can accommodate any language (without transliteration), and cope with regional, historic (including obsolete), common usage and technical words.

Specifically, this dictionary provides for: keywords; modifiers, e.g., descriptors of type, colour, material, method of manufacture, name, condition, use, etc. It accommodates any number of descriptors required to record accurately. It also provides for linkage of parts to wholes: synonyms; related terms and generics (subject appellations). Therefore, it can provide information across the range from very specific level to broad subject level. As noted in the ‘Roman gold ring engraved with a chi-rho’, the keyword would be ‘ring’, the descriptors ‘gold’, ‘chi-rho’ and ‘engraved’ and the generic entries ‘religion’, ‘metalwork’, ‘jewellery’.

The requirements specification provides a matrix for linking singular and plural forms of a word. Most software makes provision for root searching which establishes the link for many words. But there are variations in language usage and a search on the root of ‘mouse’ will not retrieve its plural ‘mice’. A search on roots is possible in the ‘SIGNPOSTS’ query system and will show all possibilities, allowing the enquirer to select only those terms that really apply to his query, thus eliminating false hits. For example, an enquirer could enter ‘St. Mary’ and be offered:

- St. Mary-at-Hill Church
- St. Mary Bothaw Church
- St. Mary Graces Abbey
- St. Mary Spital Priory
- St. Mary the Virgin Little Ilford Church
- St. Mary Woolnoth Church
- St. Mary’s Church
- St. Mary’s Dock
- St. Mary’s Nunnery

When the user identifies his ‘St. Mary’ reference, that one alone is applied to the database for searching out references.

The dictionary matrix provides a flexible and expandable means of setting up keywords. that is, they can be
This allows the user to browse around and see how terms (synonyms and generic terms) are slotted into the dictionary. Also any applicable descriptors may include red, pink (colours); Prince Henry (variety name) and flower seed (generic term). It also provides a conceptual map that can be added to incrementally to make sense of information.

'SIGNPOSTS' dictionary and thesaurus matrices differ from other indexing systems in that they are free of the constraints of 'Preferred Terms', 'Broad Terms' and 'Narrow Terms' and from the limitations of classified systems. There are no 'see' references to allow for as all words can become keywords and links made through the synonym provision. There are no 'see also' references as this function is exercised through the automatic provision of related terms from which the user can select.

At recording stage of abstract/index systems, 'SIGNPOSTS' is fast. There are no manuals of 'Preferred Terms' or classification schemes to consult and no time-consuming decisions as to what is a 'Preferred Term', what a 'Broad Term' and what a 'Narrow Term' each time a keyword is established.

At query stage, 'SIGNPOSTS' is the only system that provides a full context of terms, that is, tells the user more than he initially knew to ask for. It is an expert system usable by non-experts.

'SIGNPOSTS' also provides a printout of the thesaurus. This allows the user to browse around and see how terms are used. If he does not see a cross-reference that he thinks should be provided for, he can ask for its inclusion. The system is user-driven at all stages and can easily accommodate new relationships, interpretations, etc.

An example of 'SIGNPOSTS' in action will demonstrate the precision and possible speed and efficiency of this index-driven data retrieval system. As applied to a seed catalogue, the dictionary will contain keywords describing flowers. The flower's name is the keyword. The descriptors may include red, pink (colours); Prince Henry (variety name) and flower seed (generic term). Also any applicable synonyms are noted, e.g., 'pansy' and its synonyms 'heart's ease' and 'Johnny jump-up'. Keyword, descriptors, synonyms and generic terms are slotted into the dictionary matrix.

The 'SIGNPOSTS' recording matrix similarly provides a pattern for entering characteristic details, e.g., flower (plant type); height; spread; plant spacing; prolific bloomer; uniform (habit); border, hanging basket, pot, tub, underplanting, window box (where to plant); Feb.–Mar. (sowing season); seed (increase); June–Oct. (blooming season); bright sun (any special requirements); half-hardy biennial (growing type) and fragrant (special characteristics). It also contains provision for a detailed description (abstract). Any other details can be added, e.g., price per packet, order number and inventory control categories.

A potential seed buyer could ask, for example, for 'blue, fragrant, biennial flowers blooming June–Sept. with a height of 6"–9", suitable for a border'. He would receive a printout listing the specified flowers and their prices. This would include all shades of blue, e.g., pale blue, sky blue, sapphire blue as 'blue' was the request. If 'sapphire blue' had been specified, it alone would be selected. Details of each relevant flower will appear in the abstract enabling the buyer to be fully informed and able to select just the flowers he really wants from those available. Experienced gardeners could order by Latin names and query any new varieties.

For the seed company, there are provisions for inventory control and customer data. The data on each seed package is encoded to note stock amounts and instructions on whether to reorder and when to reorder, i.e., automatic inventory control. There are also provisions on customer information. All user aspects of a seed catalogue are cross-referred in the data bank. The manager of the seed company can quickly discover how many customers he has in England, Wales, etc., what accounts and amounts are overdue and which customers are on mailing lists for catalogues and/or other information, which customers order only flowers, etc. or any other details he needs to analyse in order to plan future developments of his business.

The medical application of 'SIGNPOSTS' allows for any type of file, e.g., publications (monographs and periodicals), specimens, X-rays, audio/visual sources, diagnostic synopses. Again, the dictionary matrix will provide the access words and links, so that an enquirer could ask for any references to 'whiplash' and a search would pick up the synonyms 'neck, sprained' and 'vertebrae, cervical, sprained' and be referred to all the relationships noted in the various files, e.g., 'symptoms', 'treatment', etc. He could then select 'treatment' and be given all relevant file sources. Citations from publications are given in correct bibliographic style.

'SIGNPOSTS' is not just a theoretical system for data recording and retrieval. Its first application to provide an annotated, indexed bibliography of the archaeology of London, entitled 'Bibliography of Greater London's Archaeology' is nearing completion at the Museum of London.

Application to archaeology

Archaeology is a study of residuals, the remaining material manifestations of previous cultures. It is primarily site-oriented. It seeks to answer these questions about our ancestors, who was where, when, doing what with what? In terms of being able to accurately identify archaeological evidence, where and when can more often be assigned than doing what with what (the how and why). Archaeological evidence is like a jigsaw. All pieces are needed to complete a picture, but all pieces do not necessarily turn up at the same time, therefore all records need to be on the database. As such records go back to the 18th century, they contain historic words not in current use. Also, 19th-century antiquarians tended to record evidence of Roman and medieval features and artefacts in Latin.

Consultations with archaeologists of the Museum's Department of Urban Archaeology (DUA) and museum
The Indexer Vol. 18 No. 1 April 1992

PUTTING THE HORSE BEFORE THE CART

They also requested that priority be given to indexing of periodical literature, as it is the richest and least accessible source of archaeological information. Some of these periodicals have no indexes, some have individual volume back-of-book indexes. Only a few have partly cumulative indexes and all have varying terms and varying degrees of precision. These periodicals date from the 18th century and could not be full text online, so it was decided to index the articles to produce the bibliography. Current users were given as museum curators and archaeologists and potential users as historians, other researchers, planning authorities and the general public.

'SIGNPOSTS' was tailor-made for all these requirements and potential users. As its basic assumption is that the user knows nothing, the beginner to archaeology can access data. The experienced archaeologist, though skilled in archaeological methods, may be unfamiliar with information retrieval methods or the variations of English language usage over the centuries, or have no Latin. But 'SIGNPOSTS' provides him with the links between the 20th-century 'basement' and 'cellar' and the 19th-century use of 'vault'. It allows him to ask for 'needlecases' and be provided with their Latin synonym 'aciaria'. It alerts him to alternative usages over the centuries, or have no Latin. But 'SIGNPOSTS' allows for any language, as it is the richest and least accessible source of archaeological information. Some of these periodicals have no indexes, some have individual volume back-of-book indexes. Only a few have partly cumulative indexes and all have varying terms and varying degrees of precision. These periodicals date from the 18th century and could not be full text online, so it was decided to index the articles to produce the bibliography. Current users were given as museum curators and archaeologists and potential users as historians, other researchers, planning authorities and the general public.

'SIGNPOSTS' was tailor-made for all these requirements and potential users. As its basic assumption is that the user knows nothing, the beginner to archaeology can access data. The experienced archaeologist, though skilled in archaeological methods, may be unfamiliar with information retrieval methods or the variations of English language usage over the centuries, or have no Latin. But 'SIGNPOSTS' provides him with the links between the 20th-century 'basement' and 'cellar' and the 19th-century use of 'vault'. It allows him to ask for 'needlecases' and be provided with their Latin synonym 'aciaria'. It alerts him to alternative interpretations, e.g., 'detached kitchens' might be 'dovecotes'.

Archaeological evidence is nearly always fragmentary. The recently discovered evidence for the Shakespearian Globe Theatre can be listed precisely as to what remains: wall foundations, made of chalk block and timber stakes; brick foundations; a metallic surface and a layer of crushed hazelnuts (also found at the Rose Theatre site) whose purpose is not known. All these remains are recorded with the modifier 'Globe Theatre'. To list only 'Globe Theatre' is to imply a standing building waiting only for the cast and audience to begin a performance, i.e., to falsify the information.

The specificity of terms in 'SIGNPOSTS' allows accurate archaeological recording: what remains are found, what they are made of, how they are made (lathe-turned), condition, colour, use and any given details as each such detail provides evidence of different aspects of past life. Any such detail left out of the record falsifies the evidence. Returning to the 'Roman gold ring engraved with a chi-rho', to record this as 'Roman ring' omits the evidence of religion (chi-rho) and the technical arts (engraving) and metalworking (gold). 'SIGNPOSTS' allows for any language that has been used to record archaeological evidence. To date, the dictionary for the 'Bibliography of Greater London's Archaeology, besides English, contains Celtic, French, German, Greek, Latin and Saxon words.

The recording matrix provides for the three categories required for producing an archaeological bibliography, namely, publication data, location data and finds data (features and artefacts), that is, the archaeological evidence. Access for the enquirer is provided for search on various elements of each category, e.g., by author, by street name, or by artefact. The archaeologist looking for all data relating to a particular excavation can access this direct by entering the site code. Another enquirer may remember a particular author but not recall the relevant article she wrote. His search on author will turn up all articles and if the titles are not recognizable as the ones wanted, a search through the attached locations and finds provides further clues. Searchers can also apply a 'limiter', i.e., ask for 'illustrated' references only and eliminate the more numerous non-relevant text only references.

For listing of finds, there is provision for chronological definition, as 'when' is an important archaeological factor. The periods are the recognized ones applicable to the Greater London area: 'Palaeolithic', 'Mesolithic', 'Neolithic', 'Bronze Age', 'Iron Age', 'Roman', 'Migration and Early Medieval', 'Medieval' and 'Post Medieval'. There are also categories for 'Prehistoric, period unknown' and 'unknown'. These are required as authors writing over 300 years have varying degrees of knowledge and may have recognized an artefact as prehistoric but could not label by closer date. The chronological periods have date assignments BC and AD, for example, Roman = AD 43–450. The system allows for closer dating. Boudiccan fire-damaged pottery would be entered under Roman with a sub-date of 60–61. At retrieval stage, the enquirer can ask for pottery of this exact date and get only those labelled 60–61. But if another enquirer asks for Roman 1st century pottery, she will get those labelled 60–61 and any others in the date range from AD 43–99.6

Allowing for specificity of dates saves the researcher time. If Roman period (43–450) only were allowed, the researcher would have to wade through references ranging over 500 years to get the 1st century ones he actually wants. As noted above, ability to provide relevant data is the key to a good retrieval system.

As archaeology is site-oriented, the 'where' is a vital element in recording and interpreting archaeological evidence, therefore as detailed an entry as possible is desirable.
for location data. The minimum is a borough. Many 19th-century archaeological finds were described as found in ‘the Borough’ (Southwark) or ‘the City’ (London), building, bridge, park, etc. and/or street names and towns or villages are noted. Linked to the location element are the details of excavator/finder/observer; Ordnance Survey grid reference; the DUA code; the SMR number; where the finds are located: where the archive of plans and excavation notes is located and whether there are any illustrations, maps and/or plans. The SMR number provides the link to its database and hence is able to link the bibliographic record to the administrative record. This saves duplication of effort in recording archaeological data.

The Ordnance Survey grid reference is essential to provide a fixed geographic location. Its necessity becomes apparent as street names sometimes provide only a current location. Basinghall and Fore Street have occupied varying locations in the City of London over time. A searcher asking for references to ‘Basinghall’ would be alerted to the differing locations by the differing grid references and avoid false hits by eliminating the ‘wrong’ Basinghalls. Further aid is provided to the searcher by the dating of the excavation/find/observation.

When the search is completed, the printout contains the user’s query and a double bibliographic listing. One listing is by journal name plus details of volume, date and paging, that is, the reading list. For finds data queries, the relevant word to be found in each article is noted. The searcher therefore does not have to remember a long list of synonyms but is given the appropriate one with each journal reference. He is given a signpost to somewhere. The other listing is by author and provides correct bibliographic citations. These provisions are timesavers for all users. For the novice to bibliography, the correct citations save him time in trying to work out the correct style for any articles he may wish to cite.

A recent search for ‘Roman lead coffins’ retrieved in a few minutes 72 references dating from 1814 to 1989 with correct bibliographic citations. It would have taken months to search through all the periodical sources to have the same retrieval rate. Their indexes are not specific enough to provide ‘Roman lead coffins’ at a glance, so any references to ‘Roman artefacts and discoveries’ and the vague but frequent 19th-century use of the term ‘antiquities’ would have to be followed up and a search made through the text to find the relevant terms.

‘SIGNPOSTS’ is also applicable to historic data: museum and library holdings; industrial and commercial inventories and parts catalogues; specialist databases (botanical); bibliographies and diagnostic databases (medical and maintenance/repair manuals).

Conclusion

Anyone setting up a data retrieval system has the problem of how and where to begin. They should begin with the users. What do they hope to retrieve from the database? If they have a current manual system, consultation with current users to establish what they are not retrieving but wish they could provides additional guidance. Transferring an unsatisfactory manual system on to a computer will not meet all users’ needs. Those with an unsatisfactory computer system of data retrieval similarly need to analyse users’ complaints and allow for their needs. The satisfactory retrieval system is one that meets current needs quickly and can add incrementally for new users with differing requirements or adjust for needs overlooked without having to restructure the whole system.

The analysis will provide the levels of specificity and cross-references required. Then the subject specialist (the indexer) is needed to establish the dictionary and data categories for recording any linked data required to provide full retrieval possibilities. The ‘SIGNPOSTS’ dictionary and recording matrices will provide timesaving methods to get the retrieval system speedily under way. However, it is not a substitute for the expert, but a manual and matrices to speed up the indexing process. The indexers are not done away with because there are indexing features in the computer program. Only the qualified indexer knows what makes data accessible to as many users as possible, that is, can establish the index terms. ‘SIGNPOSTS’ provides the means of quickly establishing the required fully articulated index from the terms inserted into the dictionary matrix by the indexer. As it can be added to at any time and at any level, it can also meet future retrieval needs. In other words, the indexers’ initial input ensures maximum retrieval with minimum effort and time on the part of the user.

A great deal of research time, effort and money has gone into producing efficient storage systems for information of all kinds in all its forms. This will not be justified until a matching effort is made to produce efficient data retrieval. As noted above, this involves semantic problems. ‘SIGNPOSTS’ proves that they are solvable.

Reference sources today tend to require labour-intensive searches to take advantage of automated technologies. Instead, ‘SIGNPOSTS’ offers a labour-intensive indexing system that automates rapid access to information by using the latest technologies. It puts the horse before the cart.

References

4. The Museum of London was established by parliamentary statute in 1965 to collect, display and interpret items reflecting the history and archaeology of the nation’s capital. The Department of Urban Archaeology (DUA) was set up in 1973 to record and preserve where possible London’s archaeological heritage. At the suggestion of Michael Rhodes, Finds Research
Officer of the DUA, the Museum of London consulted with the Department of the Environment, the Greater London Council and the Council for British Archaeology, to establish a project to produce a computer-based indexed bibliography of London’s archaeology. The Bibliography is programmed using MIPS M-120/5 running Unix v.5 and operating system. The application (titled BAGL) is maintained in the Oracle relational database management system (v.6). Screens are programmed in Oracle SQL*Forms (v2.3) with modules also written in SQL*Plus (v.3) and SQL+Report (v.1). Access to the system is via SQLMenu (V.4). The programmer is David Evans, Registrar and Information Systems Officer of the Museum of London.

5. The Sites and Monuments Record of Greater London was set up under the Greater London Council. Upon the GLC’s demise, it was transferred to English Heritage.


Audrey M. Adams is a librarian and consultant on information services. She has set up, among others, a system for recording sound effects, a financial information system and a law library. Her ‘signposts’ indexing/recording/cataloguing system was applied under her direction to produce an annotated, indexed ‘Bibliography of Greater London’s Archaeology’ for the Museum of London.

Reverse alphabetical indexing

It’s a special case, but one familiar to anyone trying to learn the written forms of French verbs.

The grammar-book takes you through one verb of each type of infinitive-ending, and its index gives the location of each type. There are a number of snags: there is no agreed list of ending-types, most types have exceptions, one type is a dustbin category, and in any case infinitive-types cannot always be defined by their ending (there are two verbs ressortir of different type).

An exhaustive index of all infinitives has been assumed to be prohibitive. By way of compromise, index-makers try to give one entry for each ending-type plus one for the full infinitive of each exception. However, the learner can’t be sure what counts as an ending-type and doesn’t know which verbs are exceptional. The best tactic in these circumstances is to look up first the full form, in case it is an exception, and then one’s best guess at its ending-type (appendre? -dre? -ndre? -drel?). Index-makers also rely on the useful escape-clause ‘and its compounds’, which lets them off needing to cater for appendre at all (so that all four of those guesses would be wrong), but relies on the learner recognizing that inscrire counts as a ‘compound’ of écrire. The net result is much page-riffling, index-phobia, and student unrest.

But these indexes can be made much shorter, more convenient, and more comprehensive. The secret of these benefits is reverse alphabetization. In a sense this is obvious, since everything depends on the final string. Verbs which behave alike are often adjacent entries in a reverse-alphabetical sequence, which can therefore be collapsed to a single entry for their common part.

Thus, -cire would cover the preceding case even for students without even a smattering of etymology. Two entries, -dre and prendre would handle the exceptionality of prendre ‘and its compounds’—which in this case all happen to end with the same seven letters. Note that a default or longstop entry is shorter than that for any member of one of its exception-classes. This has the interesting consequence that such an index is to be scanned from the end backwards, the first non-incompatible form counting as a hit. Only compounds of (i.e. strings ending in) prendre would be caught at the entry which is second in print-order (but first in consultation-order). All other verbs in -endre would get past that entry and be caught by the default. In fact a default entry for -re would do, were it not for vaincre and its compounds; in the full index the first four entries in print order are re, cre, prendre, indre.

This procedure can be a little trying at first, for it runs counter to well-established habits (the first left-to-right alphabetizer also complained of brain-fag). Its advantages are: concision (29 entries cover all verbs in -re); more reliable coverage (including verbs not yet created, for these will end in -er and be caught by that entry); no need to guess where to look or what to look for (always as much of the full infinitive as possible, reading right-to-left); entries which are possible alternatives are adjacent (eliminating page-riffling); and the index can now be read by machine. Linguistically speaking, there is also the advantage of bypassing, and therefore undermining, the quaint notion that French must have conjugations on the Latin model.

P. J. Wexler

The Indexer Vol. 18 No. 1 April 1992