Sari 22 (2004) 63 - 76

Indexing Databases on Malay World Studies: The Critical Role of Knowledge Workers

DING CHOO MING

ABSTRAK

Pembangunan pangkalan data, enjin gelintar serta antara muka sudah tidak lagi menjadi monopoli syarikat yang menawarkan perkhidmatan maklumat atas talian atau penerbit pangkalan data yang besar. Bermanfaatkan kuasa teknologi, lebih-lebih lagi bercontohkan syarikat pangkalan data atas talian, ATMA sudah membangunkan pangkalan datanya sendiri sejak tahun 1999 untuk memenuhi kehendak para sarjana dan penyelidik yang berkehendakkan bahan dan maklumat tentang Dunia Melayu. Sehingga kini, ATMA sudah membina tiga pangkalan data yang boleh digunakan iaitu Rencana tentang pengajian Melayu dalam pangkalan data PADAT, Peribahasa Melayu dan Pantun Baba yang boleh dicapai di http://www.atma.ukm.my. Dengan itu, sarjana dan penyelidik yang berkenaan dari seluruh dunia sudah boleh mendapat maklumat dan bahan tentang dunia Melayu untuk pertama kalinya di Internet. Memandangkan proses pencarian semakin banyak dilakukan pengguna sendiri, maka keberkesanan, ketepatan dan kesesuaian hasil pencarian sudah menjadi isu penting. Sehubungan itu, kemudahan dan kualiti adalah wajib kepada kewujudan dan masa depan pangkalan data yang berkenaan. Namun, pencarian atas talian boleh gagal dengan halangan yang kecil sahaja, termasuk jalinan itu menjadi terputus dengan tiba-tiba, prosidur log-in diubah, kata laluan menjadi tidak sah, penyata pencarian menjadi terlalu panjang, buku panduan tidak memberi bantuan yang diperlukan, atau kekurangan bidang pencarian. Masalah-masalah pencarian itu akan menjadi semakin besar dan rumit bila beberapa sistem atas talian itu digabungkan. Memandangkan kesilapan boleh berlaku di peringkat pembangunan pangkalan data itu, maka dedikasi, perhatian dan sokongan institusi adalah dianggap penting untuk membolehkan pengguna sendiri boleh melakukan pencarian mereka dengan baik dan memuaskan.

Kata kunci: Pangkalan data, KWIC, KWOC, pengajian Dunia Melayu, PADAT, projek pengindeksan, proses pengindeksan, peraturan pengindeksan, analisa kandungan, teknik pencarian, kesesuaian

ABSTRACT

The construction of databases and the designing search engines and interfaces are no longer the monopoly of major online services or database

publishers. Harnessing the power of such technology, ATMA has been constructing in-house databases on Malay world studies since April 1999, following the example of major online databases in offering scholars and researchers a collection of materials that can meet their needs. We now have three fully functioning databases, namely Single Articles on Malay World Studies, Malay Proverbs and Pantun, Syair and Dondang Sayang of Malaysian Baba. All of them are accessible at http://www.atma.ukm.my. Now, scholars of the Malay World from all over the world are able to enjoy customised information and document retrieval for the first time. With the search process being increasingly carried out by the end-users themselves, the efficiency, precision and relevance of search results become a major question. Simplicity and quality become necessary to a database's survival. However, online searches can stumble on the simplest of obstacles: the connection suddenly breaks, the log-in procedure changed, the password invalid, the search statement too long, the manual does not give the help one needs, or the lack of certain search fields leaves users helpless, etc. These problems accumulate and complicate matters when several online systems are accessed simultaneously. Since errors can occur at each stage in database building, thus, dedication, care, institutional support and many other critical factors are vital if information workers are to perform well, and if they are to deliver services satisfactorily.

Key words: Databases, KWIC, KWOC, Malay World studies, PADAT, indexing projects, indexing process, indexing rules, content analysis, searching techniques relevance

BACKGROUND

Databases are growing in numbers. As their quality improves, one of the most significant contributions computer science make in documentation is with the introduction of search techniques, such as keyword indexes (keyword-in-context or KWIC and keyword-out-of-context or KWOC). Since April 1999, we at ATMA have developed computer-based databases using the above-mentioned state-of-the-art technology. The enormous and ever-increasing wealth of literature on the Malay world published all over the world over the last few decades has made it necessary documentation of these materials to be done digitally, based on computer. It is hoped that the experience and knowledge that we have accumulated in developing this pioneering project will be applicable to others with ambitions of designing their own systems. So far, lack of vision, personnel, funds and facilities are factors accounting for the lack of Malay World Studies databases. The low status enjoyed by Malay World Studies had meant that Malaysian researchers have had to make use of databases developed outside the Malay world and were only marginally relevant to the subject. By and large,

the idea of constructing ATMA's databases arose from the great and intense frustration we all had experienced when retrieving materials (Ding 2000 & 2002; Ding & Supyan 2000 & 2002; Shamsul et al. 2002). All our databases are oriented towards Malay World Studies, with materials coming in all formats, languages and disciplines. They are integrated, using relevant fields for storing and searching. Ultimately, any query about any of our fields will smoothly produce a body of relevant information.

Among the challenging tasks encountered in constructing databases is indexing, another part of information management. It is a process of analysing the contents of a document and expressing them in simple yet concise language. This intellectual operation inrepresenting the text in the form of a list of possible descriptors is to facilitate information retrieval. To do that, information workers must first gain an understanding of the work process. The problem is that the structure of a subject may not be obvious or easily comprehensible, and the ways of expression and the choice of concepts may be unique and even esoteric. Words used in the documents may be ambiguous, vague and metaphorical. Any case, the index constitutes an orderly guide to the conceptual contents and physical location of documents or records. It employs a set of tags or descriptors which earmark the sources of information for which the users are searching, and systematically leads them to the relevant documents.

Indexing can be simply described as a process of creating surrogates for documents by summarising their contents. The representations lead the way to the complete text. There are many ways of indexing. We at ATMA index articles in PADAT by using author, subject, title, keyword and source. This choice of options gives flexibility and functions for a broad spectrum of subjects, vital to help users to retrieve the documents themselves. In short, index is the necessary finding tool in helping users locating the documents. The problem may be bigger than one would expect, since many users at times are not sure about what they are looking for. Their ideas may be simple or complex, superficial or profound, and their topics may not be clearly formulated even as they come to search for materials.

DATABASES AND INDEXING PROJECTS IN MALAYSIA

Among the successful computerised indexing projects in Malaysia and Singapore is the indexing of medical periodicals, called *BIBLIOMED*, which has a history of more than 50 years. Relevant articles related to medicine and health science, published locally and overseas, are indexed by the National University of Singapore Medical Library in accordance with the Medical Library Subject Headings of the National Library of Medicine (MESH). BIBLIOMED has been published by SEAMIC since 1973. Another equally successful indexing project is *PERIND* at NUS (National University of Singapore). Between 1960 and 1981,

periodical articles in the humanities and social sciences relating to Singapore, Malaysia, Brunei and Asean were indexed manually in card form. In May 1982, the *PERIND* database was set up, with record input by the staff of the Humanities/ Social Sciences/ Management Reference Department of the NUS Central Library (for humanities and social sciences), the Medical Library (for medicine and related subjects) and Law Library (for law). All relevant materials have been input using the MINISIS computer system. The NUS Library also participates in the Agricultural Information Bank for Asia (AIBA) network and contributes to the compilation of *AGRINDEX*. A total of 39 local periodicals are scanned regularly for inclusion in *AGRINDEX* and *AGRIASIA* databases. The scope covers agriculture and its related fields, including fisheries, food and veterinary science.

The Malay World Studies article database at ATMA (*http://www.atma.ukm.my*) is the first of its kind in the world. We glean old and new materials in digital and printed form, from journals, books, seminars and proceedings, theses in English, Malay/Indonesian, Chinese, Dutch, German and Japanese, among others. The use of such diverse sources is to ensure adequate coverage on as many subjects as possible. This multidisciplinary approach provides a range that compares favourably with information retrieval systems such as those developed at KITLV (Leiden) or with PERIND in Singapore. Our initial target is an arbitrary number of 50,000 articles to be collected within a five-year period, 1999-2004. We assume this figure to be both manageable and sufficiently large for all trends on any topic to be obvious. In collecting the materials, one major concern is to avoid duplication. This could be minimized with prompt processing between their receipt and their inclusion in the database. According to our calculations, only about 10,000 articles can be processed every year, using one full-time staff doing data entry, editing, cleaning and updating. A rough estimation tells us that about 2-3,000 new research articles on the Malay World are published every year.

The ultimate service of ATMA's database, like any other resource centre or indexing service, is retrieval. As mentioned earlier, information retrieval is a complex task that requires analyses of how documents are to be requested and what indexing parameters are to be used. How well information retrieval works depend on the following functions:

1. The acquisitions of documents,
2. Content analysis,
3. Content representation,
4. The coding of content indicators,
5. The creation of a document file,
6. The creation of operational search strategies, and
7. The physical dissemination of the retrieved results.

The emphasis in this article is on indexing. Any document not indexed (by a given parameter field) will not be retrieved (by that parameter). Thus, the determination of relevant indexing fields needs to be done from Day One. This step requires conceptualising and selecting fields within which the documents can be stored in and requested from. Even after this has been done, there is no guarantee that no other problem in document retrieval will arise. Indexing requires the intellectually taxing task of summarising of often complicated thoughts. Thus, it is not uncommon that information workers are accused of missing central points in the documents and, even worse, of adding irrelevant ideas. Such mistakes often stem from insufficient knowledge about new developments on the part of the information workers, or from the fact that they have not been active researchers. Similarly, the use of biased index terms can lead to biased search results. This problem is difficult to solve, since there are always discrepancies between the preferences and opinions of the author and the information worker. No matter where the failing lies, all parties involved must feel responsibility towards this matter. The data must be accurate and reliable since computer-based searching is based on the exact matching of words or phrases. If only accurate data allows retrieval, then an error will definitely lead to locating irrelevant documents or no hit at all. Thus, accuracy, completeness, consistency and timeliness are vital. They are explained as below:

a. accuracy – may be defined as being true to the original form of the data,
b. completeness – may be loosely defined as covering all materials claimed by the providers of the service,
c. consistency – may be defined as uniform application of a standard set of rules,
d. timeliness – is the time lag between the publication of the primary versus the secondary materials.

INDEXING PROCESS AND RULES

Indexing is information management to ease retrieval. Any information retrieval system is worth its name only if information and documents therein can be satisfactorily traced, located and accessed. As mentioned earlier, indexing is brief representation description of a document, usually in common language. It is intricate work involving a careful analysis of data and aims for succinctness in style. B. C. Vickery (1968): defines it "*as deriving from a document a set of words that serves as a condensed representation of it. This representation may be used to identify the document, to provide access points in literature search, to indicate its content, or as a substitute for the documents*". A database, resource centre or indexing service of high quality is the cumulative result of a series of good information management decisions. The knowledge workers, librarians or subject specialists must make certain that all the right terms are selected, and that superfluous terms are excluded. Superfluous terms waste the user's time

and lead him/her to unwanted information. What is worst, leaving out central terms will keep information hidden. Information workers have to make educated guesses about users need and how they will react to the chosen index entries. When indexing a document, information workers have to imagine the different terms a selection of users with varying aims will use to look for the same document. At this point, information workers should even ask themselves whether or not the users would be satisfied with the hits they get in using those terms. Are those terms really functional? If not, how can they be complemented by the search process itself?

Since indexing is partly a science and partly an art, the rules to follow are usually merely guidelines. The cardinal indexing rule is "index all important concepts". But what is important? Is it the frequency of its repetition? Instinctively, we do seem to think that something is repeated often because it is considered important, while peripheral ideas get only scant mention. The good first step in indexing is to involve professionally trained and experienced subject experts. They should be able to decide on the level of indexing and whether the documents are worth indexing in depth and in what way. Such a decision may however be an expert opinion as well as a value judgement, influenced by the objectives and policies of the indexing agency, and dictated by the information needs of the clientele. There are at least three points involved here:

1. Expertise judgement of document contents,
2. The appropriateness of descriptive terms,
3. The general goals of the indexing policy of the database producers.

The above are expected of an experienced information worker doing indexing according to a number of basic rules, which are:

a. Content analysis,
b. The assigning of content indicators,
c. The adding of location indicators,
d. The assembling of resulting entries,
e. The choosing of the physical form in which the index is to be displayed.

After knowing the rules, one must learn to apply them. Most scholarly papers follow certain patterns, giving information workers an important short cut in understanding their contents or their overall intent. For instance, it is common for papers to have an introduction, a summary and a conclusion. The introduction provides the structure and the internal relationship between the parts, while the summary and conclusion comes at the end as a reiteration of the content. Indeed, references are generally conceded to reflect the subject matter of the papers. Consideration should also be given to the chosen quotes. Authors tend to cite writers who have written on the same topic, both to gain support and to give readers an opportunity to go to closely related works.

Whatever the case, extreme care must be taken to see to it that information is recorded correctly, for the obvious reason that incorrect entries may render the documents inaccessible. Care must be taken at every stage – whether in data collection, data entry or data searching – to minimise errors. Often enough, a vital entry term gets misspelled and an important concept gets overlooked. Database providers may fail to carry out regular updates. Mistakes may be due to earlier factors. The authors may have their facts wrong, the publishers may have careless in all sorts of ways, and important work may have been totally missed by the information collectors. The reliability of the hardware and software, and the telecommunication infrastructure are always a worry. At the end of the line, we have the problem of bad searching skills on the part of the end users. This is hard to remedy, and the consequences may be unfortunate. The irritation an unsuccessful search causes tends to be blamed on the providers and the information workers.

In content analysis, spot reading is sometimes sufficient for information workers to understand what concepts are dealt with in a document. However, some documents may have to be read thoroughly before the information worker can feel confident about having identified the subject contents. The amount of time used in content analysis will depend on the nature and experience of the information workers. A piece of advice from A. C. Foskett (1982) is: "*Scanning a text to decide what it is about is the key operation in indexing, yet it is the least discussed and the least reducible to rule*". Scanning the texts is necessary not only in identifying subject concepts, but also in deciding on the amount of information to be presented. The depth of indexing or exhaustivity, or simply the number of topics that are to be covered in the index always influences the descriptor choice used. In a document that covers, say, five topics, if all of them are represented, then the indexing of that document is said to be complete. This means we have indexed in depth. Clearly the deeper the indexing, the more terms will be employed. However, some documents do not have clear topics, which mean that deep (and even exhaustive) indexing will not produce many terms. No one can be quite sure about how the optimal level of exhaustivity is to be decided. The depth of indexing will determine how well a retrieval system pulls out documents that are possibly related to a subject. However, one may ask if extremely deep indexing will necessarily retrieve a high proportion of the relevant documents in a collection? As more and more documents are retrieved, the risk of getting extraneous material rises. Thus, when we aim for exhaustivity, we must bear in mind that at some point, efficiency and accuracy will start to fall. However, in many cases, users are more worried about missing something than being inconvenienced by irrelevant material. One major cause for falling precision is that documents are not indexed to the same depth. Wordy articles tend to be indexed deeper than the thin ones. In any case, depth indexing demands more of the information worker. Though the depth and specificity of indexing are also dependent on the numbers of terms allowed per document, the selectivity of

terms is also partly based on the policies and the biases of the indexing agency. This means that indexing is not decided solely by the words that appear in the documents. Since the emphasis at ATMA is on Malay World Studies, information workers employed by ATMA are instructed to prioritise references to that subject and to ignore concepts unrelated or peripheral to it.

Indexing at ATMA is moving away from traditional pre-coordinate indexing to post-coordinate indexing. When uniterm post-coordinate indexing was first proposed, the idea was to index terms taken from the title and abstract and to allow users to form term combination to fit their individual needs. That way, one avoids an elaborate cross-referencing of complex authority lists, including the Library of Congress Subject Headings. We now provide advanced searches where users are allowed to combine author and title. Though post-coordinate indexing is not perfect, it has obvious advantages for both the database producer and the end-user. Among the reasons why we favour title indexing is that the title tends to indicate subject content, and usually holds important clues to the finer points of the document. A good title is, by its very nature, succinct. There is increasing evidence to suggest that titles are becoming more informative, especially in scholarly works. A good title can very well indicate the needed level of depth indexing. Due to the above reasons, some title indexes are functional surrogates for the documents. Among the drawbacks is the fact that they are always limited in the amount of information they can convey. Usually only the main themes can be summarised in the titles, leaving out some specific aspects. Furthermore, some subjects simply cannot be adequately specified by a short title. Some titles can be badly formulated and misleading, while others may be vague, either because they are too generalised or because the author may have been concentrating too much on making the title catchy. In the worst cases, and this is more common in the humanities, we get titles that are not relevant to the content of the document. Despite its manifold weaknesses, title indexing has important points that make it suitable to computer databases and the production of printed KWIC indexes. A large number of titles can be processed quickly and cheaply, and because of the use of natural language, it is relatively straightforward. ATMA attempts to remedy the weaknesses of post-coordinate indexing by adding keywords taken from the abstracts and the texts, which do express the preferences of the authors. These additional terms make our indexing more exhaustive and specific, and eliminate the problem of uninformative or misleading titles. All the keywords are displayed as headings: one entry for each significant keyword.

Are there problems in author indexing? Yes, there are many. Among them is the maintenance of consistency on the following points.

a. The number of names to be allowed per entry when a document has multiple authors,
b. The method of alphabetising,

c. The use of full name or initials,
d. The delineation of authors with common names.

So, carefully constructed guidelines are necessary for these points to be standardised.

What about subject indexing? To manage it well, our information workers have to be more than merely familiar with Malay culture. They have to be experts in it. After all, indexing requires subject intimacy and immediacy of decision, and is therefore best done by confident subject specialists. Nevertheless, a lot of difficulties involve in interpreting the subjects. Among them are:

1. It is never easy to understand exactly what the author/s mean,
2. The understanding of the content may vary with time and change of context,
3. The author may at times be unsure and unclear about his/her own intentions,
4. Authors may misinterpret and misunderstand earlier works from which they quote or from which they see themselves to be evolving from,
5. The selection of subjects tends to be strongly influenced by the policies of the indexing agency.

Having outlined some of the problems, it is then up to the training, experience, expertise, commitment and dedication of information workers to perform in such a way that adequate retrieval is ensured. The kind and level of indexing required will vary according to the background and needs of the end-users. On average, research assistants at ATMA spend fifteen minutes on each document, looking through the title, abstract and the body of text, and deciding on relevant terms. Linguistically, the problems they encounter depend on the multiplicity of languages used in the materials, not to mention variations in spelling and terminology between bahasa Melayu and bahasa Indonesia, for example. Among the linguistic problems encountered in natural language indexing are:

a. Semantics – variant word forms, antonyms, and the possible use of truncation,
b. Homographs – terms with more than one meaning,
c. Unclear hierarchical and other relationships – the lack of cross-references and the use of terms in contexts that do not give sufficient information for proper interpretation to be possible.

We favour natural language over, for example, the Library of Congress Subject Headings, mainly for reasons of economics and speed. This means that retrieval of documents relies heavily on the searcher's ability and ingenuity. Keyword searching complicates the process in several ways. Searching is more difficult and uncertain with synonymous terms, misspellings and general-specific term relationships. Without cross-references and general-specific term

relationships, users have to search through each of the possible approaches to a subject one at a time or make calculated guesses about the terms that might be found in the documents. One of the criteria of a good index is that it strongly focuses on entry terms that express the needs of the users and connects the index language to their way of thinking. However, uncontrolled keyword indexing places a burden on the searcher. Other problems beyond the control of information workers are that data can be corrupted by file transfer and it is never obvious the amount of data lost in the process.

Keyword indexing is basically a revolt against the cumbersome pre-coordinating indexing commonly used in the past. However, its simplicity makes it difficult to overcome problems arising from the vagaries of natural language: words drop out of vogue and the concepts they use are gradually replaced by new terms. The lack of consistency in terminology use in documents, especially over a period of time, affects the efficiency of keyword indexing. In so far as authors are experts in their fields, one must hope that the words they use in titles, abstracts and texts should be accurate, current and correct. Keyword indexing in natural language has the following positive and negative outcomes, here listed randomly:

1. The number of potential indexing words can be very much greater in longer documents,
2. Due to the lack of vocabulary control, a great variety of index words are likely to appear that have not necessarily been selected for their representativeness,
3. Indexing can be achieved at a detailed level, with many terms per document, with very little effort,
4. One cannot search on words with variant spelling, despite the fact that there is an ability to search on word stems,
5. Under some circumstances, natural language indexing may reflect more closely the terms used by the searchers.

The human factor must always be considered. We do make inappropriate judgements, misinterpret ideas, have lapses in memory or concentration, and are guilty of omissions and inconsistencies in the indexing. However, many of the above problems are beyond the control of information workers. We are not looking for excuses. It is obvious that authors do vary their use of concepts in many different ways. This easily leads to different index terms being chosen for the similar documents. Inconsistencies will usually result in bad indexing, which in turn lowers the quality of information retrieval. Ineffective use of a resultant index may be due to the fact that no control is carried out of trivial words, inconsistent spellings, abbreviations, word variants and multiple word stems. Great editorial care must therefore be taken when searches are being done. One simple typing fault, spelling mistake or typographical error, say when “form”

gets written as "from", or "nuclear" as "unclear", will cause important documents to be left out of the hit list.

EVALUATION OF USERS AND SEARCHING TECHNIQUES

As mentioned earlier, ATMA's post-coordinating indexing system allows users to combine two or more single index terms to create a new class in advanced search. For example, putting the individual terms "*Seni bina*", "*Rumah*", "*Melayu*", "*Minangkabau*" will give rise to a new class "*Seni bina rumah Melayu di Minangkakau*". Users are free to combine the terms with Boolean operators to express their information need as closely as possible. The problem in post-coordinate searching is "false coordinating", i.e., we obtain the conceptual opposite of what was expected when the search request was formed. For example, a coordination of "*sepak bola*", "*pisang goreng*" will also retrieve documents on "*bola sepak*" and "*goreng pisang*". This problem can be minimised by stating a query as specifically as possible using the function of "exact match" in advance searching. Users should choose the correct terms and use all of them, taking advantage of any generic searching capabilities provided by the system. Frankly, most people are poor at searching, and lack an understanding of how to utilise the databases' full potential.

In any case, databases must be continuously evaluated in order to maintain quality assurance. A database cannot be considered adequate, if it has a poor index or no index at all. Poor indexing is tantamount to the sale of inferior and faulty products, like the sale of a book with spelling and typographical errors, or missing pages. The need to address these problems has become more acute with the increased usage of large databases, both online and on CD-ROM. An evaluation of database judges its effectiveness, efficiency and precision, i.e., how well it responds as a retrieval tool! A database cannot be easily determined to be good or bad, since many factors are involved. Attempts are made to define good or bad in terms of objectives. Does it fulfil its stated purposes, and are its scope and coverage adequate? Since we are talking about how well a database meets user requirements, studying user reaction and then examining the index for accuracy and consistency is a good start. A database can be evaluated either as a separate entity or in comparison with similar ones. The objective of the former is to rate the database in terms of the needs of the clientele: what are the subject areas covered, what are its stated purposes and costs, among many others things. When the latter approach is used, we compare relative quality and cost. To do that, we must have opinions about the other databases. In essence, evaluating a database amounts to evaluating the performance of information workers, and the technical and managerial staff.

Users usually have a range of objectives in mind when they approach a database. One fellow may want a grand sweep of a topic that pulls together

every bit of information written on it, another may only be in need of a general survey of a topic, a third person may want to verify a single fact, while a fourth may be in desperate need of some final bit of information, and may have the database as his last resort! Whichever the case, a database must be able to alert users to its possibilities, and its tools must be specific enough to be sufficiently precise in its retrieval of materials. With little effort and time, users must locate relevant materials without being flooded by non-relevant ones. This is the concrete goal of a good database. Relevance is the vital criterion in database evaluation, even if the notion is not without its problems. It should possibly be judged as a matter of degree. The reasons are as obvious as stated below:

1. The relevance of a given document may change along with changes in the document collection conceived as a stock of knowledge,
2. Two users with similar backgrounds may approach a system with exactly the same question, get the same answers, and yet be poles apart in relevance judgements,
3. Even the same person may vary his or her judgement at different times,
4. A given query is bound to find documents that the user had not meant to retrieve, not because of any fault in the system but because the user did not imagine that his (perhaps actually well-chosen) query would have that side-effect until after the answers had been returned.

Although there are manuals and instructions to aid them, users have to learn how to use ATMA's databases through trial and error. Since the onus of successful searching is dependent on the users, ongoing "editing" on their part is necessary to remove errors step by step and finally to bring the relevant documents together. There are a number of checks that need to be made in this corrective procedure:

a. Check spellings,
b. Check pronunciations,
c. Check typography,
d. Check for missing entries,
e. Check for unnecessary entries,
f. Check for headings.

As mentioned earlier, ATMA's databases use both basic and advanced search techniques. In the former, users are allowed to browse all the citations retrieved and displayed, while in the latter, they can refine their searches by using Boolean operators in some sequences. Generally, browsing produces fuller retrievals, while selection retrievals are more narrowly relevant. The effectiveness of either search technique depends on the user's information searching skills, patience and the relative importance of relevance versus comprehensiveness in the retrieval's satisfaction of the query. Users can search terms of interest, linked

by AND and OR logic, and can also eliminate certain items by using NOT logic. To enable complicated searching, we are now developing searching by truncation and nestling.

CONCLUSION

Today, ATMA's databases are admittedly not as sophisticated as the major commercial retrieval systems available commercially. There is endless room for improvement. The excitement about online databases and the rising need for searches for materials to be nation-wide and international provides clear guidelines for how we are to develop. Errors are being minimised and the system is being made to be as comprehensive as possible. Originally a by-product of an IRPA research, our database project has come to accommodate the ever-increasing wealth of information on the Malay world published in all sorts of formats all over the world. We intend to make this portal a one-stop research gateway that will complement other databases related to the Malay World, such as the one developed in KITLV in Leiden. A search on a good database will surely lead to continued use. Efficiency and precision are always highly appreciated, and disappointments must be minimised if researchers are to feel the need to re-use our facilities. Maintenance is therefore of central importance, and the cost of continuous updating of technology and materials must be a price we are willing to pay. Even a high-powered and reputable database becomes obsolete if it is found wanting for too long. Any database that ever serves as a reference in any sense of the word should have an index of such quality that up-to-date information is always complete, and easily and quickly available.

Our fervent hope is to integrate all present and future databases developed at ATMA within our portal – *www.malaycivilization.com*. This will develop into a formidable gateway to information about the Malay World, enabling ATMA slowly to undertake national and international responsibility for the dissemination and retrieval of materials related to Malay World Studies. Through the use of our databases, scholars and researchers can expect to extract the maximum number of relevant documents in an acceptably short time. The alternative involves a needle-in-the-haystack task of tracking down all sorts of articles located throughout the world.

ACKNOWLEDGEMENT

This is the original paper presented at "Seminar Kebangsaan Pusat Sumber Elektronik" on 9-10 September 2002, at Universiti Teknologi MARA, Shah Alam.

REFERENCES

Ding Choo Ming. 2000. Access to Digital Information: Some Breakthroughs and Obstacles. *Journal of Librarianship and Information Science* 32(1): 26-32.

_____. 2002. Access to Materials in and on Malay Studies from Leiden to Bangi: a Model of Information Repackaging on Information Superhighway. Paper Presented at *ATMA-KITLV Colloquium: Dutch Scholarship and the Malay World: A Critical Assessment*, at UKM, 20-22 November.

_____. 2002a. Accessing the Malay world. *SEAREP Bulletin* 1:2 (April-May).

_____ & Supyan Hussin. 2000. Preservation of Culture and Knowledge Through Technology: the Experience of UKM. *UKM-UC Seminar on Technology in Development*, 25th January at UKM, Bangi.

_____ Supyan Hussin & Choo Wou Onn. 2002. Pembinaan Pangkalan Data di ATMA; Tapak Kecil dalam Langkah Besar Membaiki Aliran Maklumat Sehala dari Barat ke Timur. *Seminar Kebangsaan Bahasa dan Pemikiran Melayu, di Akademi Pengajian Melayu*, Universiti Malaya, 18-19 Jun.

Fidel, R. 1994. User-Oriented Indexing. *Journal of the American Society for Information Science* 45: 572-576.

Foskett, A. C. 1982. *The Subject Approach to Information*. London: Bingley.

Shamsul A. B., Rumaizah Mohamad &, Haslindawati Hamzah. 2002. Pengajian Alam Melayu di Pentas Global: Teknologi Maklumat dan Penstrukturan Ilmu di ATMA, UKM. *Seminar Kebangsaan Bahasa dan Pemikiran Melayu, di Akademi Pengajian Melayu*, Universiti Malaya, 18-19 Jun.

Vickery, B. C. 1968. Analysis of Information: 355-384 IN A. Kent and H. Lancour (ed). *Encyclopedia of Library and Information Sciences* Vol. 1. New York: Dekker.

Ding Choo Ming
Felo Penyelidik Kanan
Institut Alam dan Tamadun Melayu
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor Darul Ehsan
e-mail: chooming@pkrisc.cc.ukm.my