Cooperation in the Building of the South Asia Resources Database: Past Issues and Future Challenges

Ian Dawes and Maggie Exon
Curtin University of Technology

This paper describes a project to create a database containing the holdings in Australian libraries of materials relevant to South Asia. In some respects its genesis and development broke new ground in Australia since the database is the result of active cooperation between academics and librarians in the practical processes of creating such a tool for scholarly research. Essentially, the database was to have two primary functions: firstly, to provide scholars with a reliable information source describing South Asian materials in Australian libraries; and secondly, to act as a survey mechanism to determine the strengths and weaknesses of the collection.

The database contains over one hundred thousand records recording approximately three hundred thousand holdings from libraries around Australia. That the database can be this size is largely due to the existence of the Australian Bibliographic Network (ABN) which was the source of a large proportion of the base records for the database.

This paper will therefore examine the potentials and the problems in carrying out a project which involves cooperation between a large centralised national database provider and a consortium of people with a specific subject interest.

Background

The project which produced the database had as its sponsors a group of academics and librarians concerned with a perceived decline in the research value of the Australian national collection of South Asian library materials. Australia has had a long and distinguished research interest in South Asia which was at its height between the late 1950s and the late 1970s. This research interest resulted in the development of a fine national collection of South Asian materials, suitable for the level of academic research of the period. The collection, under the patronage of influential scholars, flourished in several university libraries throughout Australia.

However, just as interest in South Asia began to wane in the early 1980s, so did the collection building process. Collections which no longer had the support of influential scholars or the commitment of funding from universities lost momentum. The impact of this on the South Asian collection building process was twofold: firstly, acquisition of new materials declined markedly; and secondly, the cataloguing of existing materials became an area of low priority making the identification of, and access to, South Asian materials extremely limited.

It was in an attempt to redress the problem of access which led a group of Australian scholars and librarians to cooperate by forming a consortium so they could apply for funding for the establishment of a database of South Asian materials held in Australian libraries. Both the scholars and librarians not only wanted a tool for the listing of both the rich cache of forgotten South Asian materials as well as new materials, but also a tool capable of providing access to these materials anywhere in Australia or the world.

In 1995 the ARC (Australian Research Council) agreed to support a proposal from a consortium of Australian universities and national institutions to improve Australian holdings of South Asia library materials and to improve access to South Asian library materials. $AUD200,000 was supplied to the project from the Australian Government; contributions from consortium members meant that over $AUD340,000 was devoted to the project. In 1996 the project successfully applied for and received a second round of funding which, with members contributions, totalled $AUD225,000.

Once funding for the project had been guaranteed, a workshop was held where scholars and librarians from the consortium members, along with representatives of the National Library of Australia, examined strategies designed to achieve the major aims of the project; primarily to establish a database. Those attending the workshop agreed that the database should be known as the eeSouth Asian Resources DatabaseAE or SARD and should be developed using the following general guidelines:

1. Be designed to record a variety of types of bibliographic and access information.

2. Have a sophisticated record structure(s) which could deal with all kinds of materials involved in sufficient detail for efficient retrieval. The database should be able to deal with archival material and include bibliographic information supplied by non-library professionals, such as scholarAEs descriptions of their own personal collections.

3. Have the ability for on-going record enhancement when necessary including the addition of non-bibliographic type information, condition reports, access reports and multiple location reports.

4. Be available via the World Wide Web (WWW).

5. Be able to display introductory information which recognises all contributors to the database including the Australian Research Council, consortium members and designers.

6. Have a user information feedback function via e-mail.

7. Be maintained by librarians to ensure quality of the database and integrity of the data is maintained.

What also came out of the workshop was a broad agreement to seek the cooperation of libraries with major South Asian holdings to participate in a national survey of South Asian materials. The survey was to take the form of a request that targeted libraries supply details of their South Asian holdings. This information was to take the form of electronic catalogue records of South Asian materials in a format which could be loaded to the project database. If no electronic record existed, cooperation was sought to allow project field officers to create an electronic record of any South Asian material held by the library. Field officers were to use laptop computers loaded with appropriate software to create a catalogue record which could then be loaded to the database.

The success of gaining the cooperation and goodwill of libraries throughout Australia provided the project with a foundation of support and credibility which extended well beyond that of the original consortium members. Because the project set out to establish a database detailing the eenationalAE collection it was required, by definition, to include all significant collections within the entire country. Without this, not only would the databaseAEs effectiveness as a research tool be diminished but it would have also impeded the toolAEs usefulness for surveying the national collection. In other words, without the cooperation of all targeted libraries the goal to provide scholars with a reliable research tool would have been seriously compromised.

However, this goal also brought with it a number of potential problems. Foremost amongst these was the issue of persuading targeted libraries to share the ownership of their catalogue records with the project. The creation of catalogue records can be costly and represents a considerable investment by libraries. Other than the desire to be contributors to a project aiming to improve the flow of scholarly information, there was no compelling reason why libraries should have agreed to assist the development of the database in such a fundamental way. Without exception, all targeted libraries agreed to cooperate with the project in undertaking the survey and to make their records available in a format (unspanned MARC) which could be loaded to the database. If no record was available libraries agreed to allow field officers to create records describing their collections.

The project also achieved two other significant agreements which were to prove fundamental to the success of the project. The first of these agreements was reached between the project and the Library and Information Service at Curtin University of Technology. This agreement saw Curtin University guarantee technical and staff support for a period of five years. This provided the project with the necessary stability to make staff appointments as well as providing the technical infrastructure necessary to host the project. The other significant agreement was with the National Library of Australia in the form of a Memorandum of Understanding, which would see a supply of relevant records from the Australian Bibliographic Network (ABN) directly to the project database.

Cooperation with ABN

As has been made clear, much of the information in the database comes from ABN. The ABN system provides, in effect, a union catalogue of holdings in Australian libraries. The database can be interrogated and, once a particular publication has been identified, a list of holding libraries displayed. It might therefore be asked why the consortium should have wished to develop a separate union catalogue.

There is a number of reasons why ABN could not provide a satisfactory service to South Asian scholars at the time the project was developed. Some of these will be partially at least overcome when ABN installs its new system (to be called Kinetica) early in 1999, but full details of how the new service will operate have yet been announced.

1.ABN is essentially only available to information professionals, not to the public. To search the database it is necessary at present to be a registered user and to pay fees. The assumption is that the database is to be used as a source of cataloguing data and as an aid to providing a user-pays inter-library loan system. Therefore fees are quite appropriate. The purpose of the project is, by contrast, to provide a union catalogue of South Asian materials in Australian libraries designed primarily for use by scholars and students.

2. There are a number of libraries, including some special libraries of particular interest in the South Asian studies field, which are not members of ABN. They feel that the fees involved would not be compensated by the services offered by ABN. In particular, they do not feel that they would be able to obtain from ABN catalogue records for enough of the highly specialised material which they acquire. The project wished to improve access to these collections by incorporating records of their holdings.

3. A number of important libraries which are members of ABN have significant collections which are not recorded on ABN. They are uncatalogued, are only listed in manual catalogues or are only listed in electronic catalogues which cannot produce output suitable for ABN use. The chances of this material ever being processed or re-processed to produce records acceptable to ABN are very slight, because of the resources needed. By obtaining funding, the project could contribute to making records for this material available in electronic form, both on the database and in the catalogues of the originating collections.

4. ABN, despite the fact that it has made some moves towards the recording of archival materials, contains for the most part only published materials. The project wished to capture information about such archival collections, especially those which had never been described and therefore were largely unknown. The Erulkar Collection at the University of Western Australia is a good example of this.

5. The current searching system available on ABN, even were it available to scholars, is not easy to learn or use. This point will be developed further.

The issue of public access to the Kinetica database has not been resolved. The system to be known as World 1 which had been proposed by the National Libraries of Australia and New Zealand, would have seen the development of a wide range of subsidiary databases on particular topics or for particular collections which certainly would have been available. Unfortunately software problems have meant that the National Library is now concentrating on upgrading the software for the ABN database itself and for its library users. It is not yet clear how public access to the database may change.

In any case, there are a number of other considerations which make a specialised database attractive. The database can be linked to other web sites of interest to South Asian scholars. We have established a feedback service in which users of the database are encouraged to contact us by email with corrections and additions to the database. We have access to the encouragement, support and, we hope, funds to ensure that the database is maintained into the future and does not just end up as a good idea of its time. This was a unique opportunity to create a database which made a link between the enthusiasm and specialised knowledge of scholars and the quality cataloguing of an established cataloguing institution.

Retrieval considerations

The ABN database is now very large. It is known that retrieval effectiveness degrades the larger a database becomes. In particular is this true of the ABN database which at present does not have an adequate Boolean search engines enabling easy narrowing and refinement of search strategies. Search results are characterised by poor precision and therefore very large hit lists are common. To split such a large database into clear content-based smaller databases seems a reasonable response to this problem. Area studies provides a good way of doing this. Users can research such matters as computing or family studies by searching on these generic terms, knowing that all the retrieved material will apply to South Asia.

Another problem is that the ABN system does not allow searches over the whole record, but only in nominated fields. This affects recall, especially when the searching is being undertaken by non-librarians, who often find it difficult to nominate the most suitable fields for a search. We felt that any database made available to scholars should have a sophisticated search engine which was yet easy to use and precise. It should allow the easy construction of complex Boolean searches. However, there should also be a simple search of the whole record which ensured that even those new to the database could obtain satisfactory results.

Another consideration was that we wanted to provide a database which allowed users to browse through indexes to the data to identify likely search terms and pick up such searching problems as variation in the spelling of transliterated names.

Technical considerations

Given the fact that we had decided that it was appropriate to run a separate database, we had to decide what software to use. The most important requirement of the software would be that it should be able to import and export records in a structure which was acceptable to ABN, since a great deal of downloading from ABN and uploading of some of our new records to ABN would be necessary. Most bibliographic records of published material are stored and traded in the well-established MARC format and the database would have to be MARC compatible.

An obvious solution would be to run the database through a standard library software package. However, we were anxious to get away from using a package which used data structures which were based on library materials and in particularly the field structure of the MARC format. We hoped to add materials of many different kinds to the database, including archival materials, ephemera, even realia. Experience had shown that the information captured about such materials had to be distorted to fit it into library cataloguing code requirements. In any case, the database was to be run separately from any existing library catalogue and library software was too expensive and had many functions which we did not need.

We therefore needed a software package which had the following characteristics:

1. MARC-compatible without the data structures being MARC based;

2. easily exportable to the world-wide web;

3. capable of dealing with a variety of data structures within a single database;

4. not subject to commonly-found limitations in size of fields, repetition of fields within a single record and indexing capabilities;

5. provide a flexible, fast and simple search engine.

Recording this list of requirements implies that we drew up the list and then identified a package from a list of possible contenders. This would be distorting the fact that we had an obvious candidate in mind from the very beginning. This was Inmagic DB/Textworks, a well-established database development package which had always been aimed at the library and text retrieval market rather than business applications. It has a number of features which make it particularly suitable for this market.

1. It allows for virtually unlimited data in any field. It therefore offers the possibility of the easy incorporation of the full-text of documents into the database. It also allows links to image fields which can contain an image version of a document.

2. It automatically creates whole-of-field and keyword indexes to any field and makes these available to those searching the database.

3. It allows any field in a record to be repeated, enabling it to deal easily with, for example, items which have more than one author or which have variant titles. In particular, this feature means that the software developers have been able to incorporate conversion of MARC records, these records being highly dependent on the use of repeating fields.

4. It automatically suppresses all unused fields in any record, allowing records with apparently entirely different structures to reside within the one file. There can be different data entry screens for different types of material. Reporting formats do not contain large amounts of unused space.

5. It has a fast search engine designed specifically for the retrieval of text.

6. It has a CGI-based web publishing capability.

7. It has very good editing and record matching capabilities, allowing the identification of duplicate records.

The Department of Information Studies at Curtin was already using this package for teaching purposes, and therefore an economical solution to the software problem was provided.

The database was set up with a very full and flexible field structure to enable all kinds of materials to be appropriately recorded. There is also full allowance made for the recording of details about access to materials and their condition, together with comments about future enhancement of the records.

Data capture and conversion

The records for the database come from a number of different sources and are of varying quality. The source of each record is recorded for future identification and enhancement if this becomes possible.

ABN data

Several carefully constructed searches were made on the ABN database to identify records of interest to South Asian scholars. The searches made use of both the fixed and variable field data available in ABN records. These records were sent by ftp to Curtin in unspanned

MARC format, a process which was remarkably successful considering the large size of the files. Little difficulty was experience in loading these records, except for the occasional faulty record.

The search strategies used by ABN have been recorded so that the searches can be repeated at intervals to pick up new data added to national database.

Data from other on-line catalogues

Information supplied to us in electronic form from other on-line catalogues has varied considerably in quality and format. Some of the records could not be produced in formats which would be acceptable to ABN so have never been uploaded to ABN. The records of one prominent university library were in a spanned MARC layout which could not be read by Inmagic and a conversion program had to be written by the project team to convert these records directly from their original format into Inmagic format. Records from another library, although to first glance normal unspanned MARC records, each differed by one character from that expected by Inmagic and failed to load until that one character was identified and changed.

Some of these records date from a very early period of library automation and do not reach todayAEs acceptable standards. Nevertheless they are useful additions to our database.

Records entered directly to Inmagic

Some card and fiche catalogues have been directly entered to our database. This has normally been a direct transcription of the record, whatever its quality. Some project officers were very concerned about the state of the records they were entering but a policy decision was made that it was more useful to have them rather than not.

Numerically the fewest but the most important were the records of original cataloguing, much of this of ephemeral and archival material. These records were added by the field officers directly to the database. At some time in the future these records will be uploaded to ABN, when the new system is established.

Challenges

Catalogue records for languages written in non-Roman scripts

Our holy grail is the same as for all those who create bibliographic databases; to be able to display all records containing non-Roman script in the original script. Unfortunately, this is not yet possible for our database, especially as most uses of the database are through the World-Wide Web.

We are therefore faced with the problems of transliteration and here our results are less than satisfactory. We received large numbers of existing records from the Library of Congress (LC) via ABN. These records use LC supported transliterations which are then encoded using the USMARC character set. This character set supports diacritics which cannot be displayed in the ISO Latin-1 character set (ISO 8859-1) used for the Web. The Inmagic software offers the possibility of removing these diacritics when adding the records to the database and this is the option we have chosen up to now. This means that the records are reasonably pronounceable and presumably most words can be recognised by a speaker of the language but the result is not a true representation of the language nor can the records be transliterated back into their original script.

We would like to improve on this situation but are uncertain about whether to pursue the path of better romanisation or attempt to move towards a database which can support other scripts. The wealth of different languages which are represented in the database and the plethora of international standards in this field mean that it is difficult for non-specialists in this area, such as ourselves, to make sensible decisions. It is good to come to a forum like this which should be a source of such expertise.

Ongoing maintenance and supply of records

Probably the greatest challenge to the project in the future is keeping the database relevant; that is, keeping it up to date and refreshed with records of new additions to the national collection of South Asian materials. This challenge is amplified by the fact that the project does not receive any on-going funding to maintain either staff or infrastructure.

It is particularly important that the relationship with ABN (Kinetica) is maintained. The project was fortunate in getting the main transfer of records before the recent problems with ABNs software upgrade. These now appear to have been resolved but the new ABN system will not have the increased flexibility which was hoped for from World 1. We need to ensure that, when the dust settles, all the agreements we had with the National Library as custodians of the national database, are still workable.

It is perhaps fortunate that the project was able to put into place agreements which will guarantee the database a home for the foreseeable future and it is also fortunate that the project can rely upon the enthusiasm of the consortium members who invested energy into developing the database. But, perhaps the databaseAEs greatest asset and the one which will go along way in guaranteeing the future relevancy of the database is the goodwill displayed by non-consortium libraries whose only gain was participate in what they saw was a worthy scholarly venture.