LOCAL MOUNTING OF THE BIOSIS DATABASE: TRIALS AND TRIBULATIONS

Joseph G. Wible
Hopkins Marine Station
of Stanford University
Pacific Grove, CA 93950-3094
email: wible@krypton.stanford.edu
phone: 408-655-6228
fax: 408-373-7859

ABSTRACT: A discussion of the procedures undertaken by Stanford University Libraries to mount Biological Abstracts and Biological Abstracts/RRM on the campus online computer.

There are many steps involved in local mounting of large bibliographic databases such as BIOSIS. There is license negotiation, disk storage allocation, search engine selection, etc. This article begins after all these steps have been completed, and you have a sample tape of the raw data and need to make decisions on how the data are going to be indexed and displayed. It will provide you with some sense of the nuts and bolts decisions that have to be made. Large academic institutions that are considering mounting the BIOSIS database should find the information of immediate relevance. But even if you are not mounting BIOSIS locally, you should find some of this information relevant in your role as a BIOSIS searcher. The more one knows about the guts of how a database is structured, the better one is able to retrieve the desired information from that database.

When looking at the BIOSIS tape data, it is easy to recognize that the original database design has not changed much since it was originally set up to produce the print product "Biological Abstracts". For example, there are two title fields. One title field has the author's original version of the title in upper and lower case letters. This is used to print the citation and abstract in "Biological Abstracts". Then there is a second title field where the title is in all upper case letters, some words have been "normalized", and a series of supplemental keywords have been added. In the past BIOSIS made a lot of changes to titles to standardize terms, but today about the only changes made are done to deal with non-standard characters (eg. Greek letter, math symbol, etc.). Therefore less than 5% of the titles have any changes made, except for being put into upper case and having keywords added. This second version of the title is used to create the KWIC index for the print product.

Given the large size of the BIOSIS datafile, Stanford did not want to store two versions of every article title. As part of the pre-load processing of each tape, the supplemental keywords are stripped off and placed in their own field. The two titles are then compared and, if they are the same, only the upper/lower case version is retained. If they differ, both versions are kept to take advantage of the "normalized" words in the indexes.

Online searchers of BIOSIS are used to being able to limit their results by document type. What online searchers may not realize is that BIOSIS does not provide a document type field on its data tapes. Instead, database vendors such as DIALOG must use algorithms to create data for the document type field. BIOSIS provides advice on how generate this data. In most cases, the first word of the supplemental keywords indicates the document type of non-journal article records. For example "book" is the first keyword in book records. However, "book" is assigned only to records for whole book. It is not assigned as a keyword to the records for individual articles within the book. Stanford needed to assign a "book" document type to the individual articles within a book so that display format would make sense. One uses different fields and formats when citing a section from a book then when citing a journal article. Therefore Stanford modified the algorithm suggested by BIOSIS. Any record that has an ISBN or has "book" as a keyword is given a document type of "books".

Another field online searchers have come to expect that is not explicitly supplied by BIOSIS is the supertaxonomic groups such as invertebrates or mammals. This field must be created as part of the dataload through an algorithm run against the BIOSIS supplied biosystematic codes.

Another glitch in the dataload arose when trying to deal with journal titles. The data tapes include only the journal abbreviation. Stanford wanted to display full journal title, so we purchased an authority tape from BIOSIS. What was originally received was the tape used to produce the printed "BIOSIS Serial Sources". However, this tape could not easily be used to expand the abbreviated titles to their full title. BIOSIS uses CODENs as their authority for linking records. But if one matched the CODEN from an individual citation record against the authority tape, there was not a single match. Instead one could get three or four matches as the CODEN matched to various "See" records that tracked title changes over time. As with the data tapes, the authority tape was derived from a system designed for producing a printed product and not an online searchable database. Eventually Stanford was able to get a tape where a CODEN search matched to a single record containing the full serial title. It is interesting to note the tape was labeled as a tape created for BRS. Obviously Stanford wasn't the first one to face this problem.

Data tapes from BIOSIS come twice a month. Each delivery contains two separate tapes, one for records going into "Biological Abstracts" and a second for records going into "Biological Abstracts/RRM". This is another artifact of the printed index. As part of the data loads, Stanford adds a field to record whether it is a BA or an RRM record. This is done so a searcher of the database can limit the results to one set or the other.

There were numerous other decisions to make. For example, what fields should go into the "subject" index? The decision was made to include words from the title, abstract, keywords, and biosystematic codes. Not included were words from the concept code field since they tend to be too broad and generously applied to a high percentage of records. Which fields should be indexed? The online searcher's ideal would be to index every field, but this must be tempered by the realities of disk space and computing capacity. Some things such as language, year of publication, and supertaxonomic groups were set up as limits that could be used against an existing search result rather than as index terms. What fields should receive both word and

phrase indexing? Here the needs of the searcher prevailed and the title field and the publication field received both word and phrase indexing. The importance of indexing the author's address field was also recognized. It is great to be able to search for articles published by Hopkins Marine Station.

Once all these dataload decisions were made, there still remained the very important, but time consuming task of writing all the help screens for the online file. Around 40 help screens specific to the BIOSIS database were created.

In defense of BIOSIS, it is important to point out that many of the problems highlighted here are being addressed. Of course, now that Stanford has found solutions to each of the challenges presented by the BIOSIS data tapes, each time BIOSIS makes improvements Stanford must rewrite its tape load program to adjust to the new database structure. But none of this is unique to BIOSIS. There have been similar stories told by colleagues involved in mounting other bibliographic databases.

Thanks are extended to the other members of the Stanford team who worked hard to design the BIOSIS database load - Grace Baysinger, Bo Parker, and Roberta Lucier.