4 Creation of Okapi source file

This chapter describes how the Okapi catalogue file is created and organised. The data originate from MARC exchange tapes. Field selection is done on the Polytechnic DEC-10 and the remaining processing on the PLAN 4000 network. The Okapi record format is much simpler than MARC. The reasoning behind field selection is discussed, together with some of the problems. An example of a complete MARC record in Okapi format is given in Figure 4.1 at the end of the chapter.

Fuller details can be found in:

Appendix 1: MARC field selection criteria
Appendix 2: Subfields used from selected MARC fields
Appendix 3: Okapi record format
Appendix 4: Special characters used in the Okapi record

4.1 Machine-readable catalogue and choice of project test site

The Polytechnic of Central London has had a machine-readable catalogue since 1975. It is made available on microfiches which are updated monthly. This catalogue provided the source data for the Okapi project.

At the beginning of the project PCL's machine-readable catalogue was being maintained by the British Library's BLAISE/LOCAS service, and PCL was also a circulation member of SWALCAP. In June 1984 the catalogue was transferred from LOCAS to SWALCAP.

PCL's library occupies five sites in central London, from Marylebone to Holborn. The largest site is in Riding House Street and contains some 30,000 titles including Art, Business studies, Communication and Social Sciences. This site was chosen to provide the catalogue and test site for the project. It was chosen because it is the largest site, it has wide subject coverage, and it was also thought likely to provide a realistic range of user attitudes and aptitudes during the evaluation phase.

In April 1983 the project obtained a MARC exchange tape from LOCAS containing only the test site's stock — about a third of the Polytechnic's holdings — by selecting on the site code (held in a local field). An initial selection of fields was made from these records and a provisional source file was created for developing and testing the early versions of Okapi.

The following year it was decided to recreate the file from the machine-readable catalogue, in spite of the amount of work that this would entail. The reasons were threefold. Firstly, preliminary work on record displays had shown that the initial record format had been inconveniently oversimplified. Secondly, the acquisition of the PLAN hardware (Section 1.3) made it possible to handle a much larger file. Thirdly, the realistic testing and evaluation of the system required that the data be as up-to-date as possible.

Consequently, in April 1984 a second exchange tape was obtained from LOCAS immediately prior to the transfer of the catalogue to SWALCAP. This second tape (occupying three reels) contained PCL's entire monograph catalogue except for the few items which were not retrospectively converted — about 90,000 records.

PCL's catalogue probably contains a fairly typical assortment of pitfalls for the OPAC creator. Any catalogue created over a period of time will not be homogeneous or self-consistent, even when, as at PCL, one standard has consistently been followed (UK MARC, including the current edition of the Dewey Decimal Classification). The PCL file inevitably contains spelling and typing errors, and mistakes in the assignation of tags, subfield codes etc. There are also places where UK or Library of Congress (LC) policy changes result in more than one version of a name or title giving rise to several sequences where one would be preferred. Retrospective conversion using cheap labour and/or data from other institutions is another hazard. The use of controlled language for subject headings gives the possibility of some standardisation, but it is rare to find a large catalogue in which such a policy has been applied throughout. The PCL catalogue is typical in containing subject headings from several sources: LCSH, Precis, PCL. Nearly half the records have no subject headings at all, being from the UK Retrospective File, or EMMA since 1979.

To some extent an OPAC reduces the severity of most of these problems. If an author's name is misspelt an item is far more "lost" in a

conventional catalogue than in an OPAC. Because it can supplement orthodox catalogue entry points, for example by indexing text words, an OPAC provides far more points of access to each record, and, because it is interactive, there is no reason why it cannot be "helpful" and "intelligent" and make allowance for inaccuracy both in the data and from the user. But, of course, if someone has mistyped the spine label or misshelved the book then even the most friendly catalogue in the world will not enable the user to find it.

4.2 Selection of fields from the MARC record

The choice of fields and subfields from the MARC record was influenced by several factors. Each record had to have as many access points as possible. Okapi needed to be able to display records sensibly, but could not afford to waste disc space by holding unnecessary fields, or by duplicating information already held elsewhere in the record. The record format had to be much simpler than the MARC format in order to save code as well as space. The online catalogue had to give enough information to be acceptable to staff and students accustomed to the microfiche catalogue, so that at the very least Okapi could be evaluated effectively. It was also hoped to produce a system that could be adapted for other libraries eventually.

In other words it was necessary to juggle the conflicting aims of making records that were brief and simple, but nevertheless fairly comprehensive.

In common with other libraries which make use of the MARC format, PCL's input standard does not use all the available MARC fields, nor has the subset that it has used remained exactly the same over the years, because of changes in national library policy and local housekeeping practice, and the availability of resources for retrospective conversion. It is therefore inevitable that the selection of fields and subfields is to some extent institution-dependent and even somewhat arbitrary.

The Okapi record uses data from the following MARC fields [1]:

001	control number
008	information codes (language code only)
082	Dewey numbers
083	verbal feature headings derived from PRECIS field

100,110,111	author (initials only from forename subfields)
240 or 245	title (excluding statements of responsibility)
248	part numbers and titles
250	edition statements
260	publisher and date
440,490	series titles
505	contents note (only for records with analytical entries)
509	DC edition note
600,610,611	name subject headings
650,651	LC subject headings
700,710,711	added names

and local fields: accession numbers, site codes, spine label.

Appendix 1 gives selection criteria. Not all subfields are included. Appendix 2 gives full details of selected fields.

4.3 Discussion of MARC fields omitted and other problems

4.3.1 Statements of responsibility

After considerable discussion it was decided not to include statements of responsibility, $d, $e, $f from the 245, 248 (title) fields. Instead any added names (700, 710, 711) would be displayed as well as used as access points. The advantages of this policy are that there is no duplication of names in the record, which saves space, and there is no danger of the name searched for not being displayed. The disadvantage is that Okapi cannot display the actual contribution a particular person has made, i.e. translator, editor, illustrator, etc.

The alternative policy of including statements of responsibility and using added names for access only was rejected. It would certainly mean that the contribution was made clear, provided that the added name corresponded to a statement of responsibility. The disadvantages would be wasted space in the majority of records, and confusion in cases where the name searched for was not displayed because it had been held in a note field, rather than in a statement of responsibility.

The question of the value of statements of responsibility is raised in Section

9.4.5, where there is a discussion of a possible alternative method of storing the information they contain.

4.3.2 *MARC fields omitted*

The main MARC fields omitted in the current Okapi record are the 745 (added title entry) one of the CCR recommended fields [2], the 300 (physical description) field, which is not included in PCL's input standard, the 5xx (notes) fields, apart from 509 (DC edition) and 505 (contents), and the 9xx (reference) fields. The latter are particularly useful for names with more than one accepted spelling, e.g. Trotsky, and have been put to good use in Cambridge University's OPAC.

4.3.3 *Records with analytical entries*

The contents note (505) is included only for records with analytical entries. It is a compromise solution to the problem raised by such records. Ideally the record format should enable it to be displayed appropriately, according to which analytical level it was accessed by, but this would entail heavy overheads. However, it is not easy to simplify a MARC record with analytical entries without distortion or truncation. One way of handling them with a simple record format would be to create a separate record for each level. Lower level records could either be linked to the parent record, or contain the "in" statement as a note. It is cataloguing practice to restrict analytical levels to a maximum of three.

Fortunately records with analytical entries are fairly rare, and one can argue that their occurrence does not justify the overheads that would be involved if the chosen record format was designed to accomodate them perfectly.

The original subset comprised about a third of the PCL catalogue - approximately 30,000 titles This contained 57 records with analytical entries, i.e. less than 0.2%. An examination of these records revealed that in over half of them the contents were quite adequately given by the 505 (contents note) field, and that in over two-thirds of the remainder the 245 (title) field contained details of the complete contents. In only five cases could it be said that the full contents of the volume would not be described by including the 505 field in addition to those fields already selected.

It therefore seemed reasonable to design a simple record format with only one author field and only one title field, and to cater for records with

analytical entries by incorporating any 505 (contents note) MARC field into the record.

4.3.4 Other problems of creating a MARC subset

The statements of responsibility question highlights two of the main problems facing those who try to create a subset of fields from the MARC record.

By giving absolute priority to comprehensiveness the MARC format creates records that are extremely large for the amount of actual information they contain. This is due partly to the duplication, or near duplication, of data in the record, and partly to the rather profligate provision of tags, indicators, levels, and subfield codes which in combination allow for the individual identification of 26 million different types of information.

Another major problem arises from the nature of personal names, which is a fact of life that cannot be blamed on the MARC format but is certainly exacerbated by it. It is quite possible for the same name to appear in several different forms:

Example: 245 $e[translated from the French by Anne J. Cope]
700 $aCope$hAnne Jacqueline
245 $aAnne Cope, the story of a translator
505 $acontains a supplement by A. J. Cope

It is desirable but almost impossible automatically to recognise personal names as such wherever and however they appear so that they can be kept as a phrase, avoiding ridiculous but relatively harmless index entries for words like "Anne". It would also be useful to be able to recognise a surname whatever its code or context. The creation and use of the surname key is discussed in Section 5.4.1 and in Chapter 7. The automatic identification of types of terms has been discussed recently by Jones and Bell [3].

The problem also arises when the user is being prompted to enter the name of an author (see Chapter 7).

4.4 The Okapi record

After deciding which fields and subfields were to be selected from the MARC record it was necessary to decide how they were to be arranged in the Okapi record.

If considerations of space and time were not critical, there would be no reason to make a subset of the MARC record at all. The full MARC record could be examined by the indexing programs in order to create the index entries, and suitable subsets of any MARC record could be created at run-time to display to the user.

A second approach would make use of two versions of the MARC file: the original file and a displayable version. The index would be created from the original MARC file. Index creation is a batch process so time taken and space required are not critical. The much smaller displayable file would be all that was online during searching. It would contain images of each record ready to be displayed on the screen. They would require little or no processing before display, and no tags or other identification. It is usually thought desirable to be able to display two or three different versions of the record, so this approach might entail also storing brief records. The brief records then become a sort of index. There is some mention of this in Section 5.3.1.

The third approach is to design a compromise record that is more compact than the MARC record, but still contains sufficient structure to be able both to display records sensibly and to create appropriate index entries from them. This is the approach adopted for Okapi. A carefully designed record format can be manipulated online with the minimum amount of code; for example, if most of the punctuation is already embedded in the data, instructions to do this are not needed in the online program.

The Okapi record consists of a fixed length field directory followed by nine variable length fields:

Field directory
1 Author
2 Main title
3 Edition and publisher
4 Series and part titles
5 Added names

6 Class marks
7 Accession numbers
8 Codes and control number
9 Subject headings

Each field (including empty fields) is terminated by a "#", and the last (ninth) field is also followed by CRLF (carriage return line feed).

Since the field directory contains the field lengths, the field terminators are redundant, because the length can be used to determine the end of a field. Similarly the CRLF at the end of the record is also not strictly necessary. It is normal to include such redundant information as a security measure (a belt and braces approach) since it enables one to continue to use a file even if a record has been corrupted.

The record is of variable length padded with "+" characters to a multiple of four bytes (double-word). So the record length is rounded up to a whole number of double-words. This means that a record address can be expressed as the address of a double-word rather than as a byte address. In other words it is possible to have four times the addressable area without increasing the address length. This makes it possible to address a file of reasonable size using three-byte disc addresses.

Explanatory note: it is desirable for record addresses to be as short as possible since each posting in the index has to include the record address. The maximum integer two bytes can hold is 65,535, so only a quarter of a megabyte can be addressed (in double-words). Three-byte addresses can address 64 megabytes (in double-words).

The fields in the Okapi record are described in detail in Appendix 3. Appendix 4 describes the special characters used in the Okapi record.

4.5 Creation of Okapi file from MARC file

One drawback of the use of microcomputers, whether networked or not, for library applications is the dearth of magnetic tape equipment. Even when it is obtainable it is relatively expensive. MARC exchange tapes will often provide the source data for library applications and these tapes will usually have to be "read" on a mini or mainframe. Since this applied to Okapi it made sense to do some initial processing of the data, in particular field selection, on the mainframe so as to reduce the volume of data before

transferring it to the micro network. The creation of the Okapi file from the MARC file therefore falls into three main stages.

(i) The creation of the basic nine-field Okapi record from the MARC record on the Polytechnic's DEC-10 computer.
(ii) The transfer of the data from the DEC-10 to the PLAN network.
(iii) The final creation of the complete Okapi file on the PLAN network.

The next process, indexing, is described in Chapter 5.

4.5.1 Creation of Okapi record from MARC record

Although various programs already exist to strip selected fields from MARC exchange tapes it was decided to write a special program for Okapi in order to retain flexibility and so that other processing could be done at the same time.

The program is written in COBOL and runs on one of PCL's mainframes (a DEC-10) prior to transferring the records to the PLAN network. The program is driven by a table of required tags and subfield codes.

The MARC tags, indicators, subfield codes, level and repeat numbers are not retained in the Okapi record. The information they provide is used during the field selection process and where necessary converted into a different form in the Okapi record. The information is used in the MARC record to:

— identify the type of data
— show the relation of data to other data in the record
— give the number of non-filing characters
— indicate what punctuation should be used.

In the Okapi record:

(a) Data types are mostly implicit, e.g. Field Six can only contain class marks, Field Seven can only contain accession numbers. Elsewhere the data type is indicated by a special character (see Appendix 4).

(b) The relation of one piece of data to another is implied by the ordering of the data within the field, for example in the part and series title field.

(c) The non-filing characters, usually a leading article, are demarcated by special characters (see Appendix 4).

(d) Suitable punctuation (based on [1, Appendix G]) is added to the data as the Okapi record is being constructed including, for legible display, a full stop at the end of text fields that do not already end with another punctuation mark.

4.5.2 Transfer of data to PLAN network

The next stage was to transfer the data from the DEC-10 to the PLAN network. Since the two configurations have no common medium (disc or tape) it was necessary to find a way to make them "speak to each other". The chosen method was to make one of the Apple microcomputers behave as if it were a DEC-10 terminal. It could then be used to transfer data from the DEC to the PLAN either directly, down a line, or indirectly via floppy discs or a small hard disc.

To make an Apple behave like a DEC-10 terminal a special program was written for a stand-alone Apple II with a serial interface card, (Computech's serial communications interface, the Diplomat card). Using this program the Apple initially behaves like an ordinary DEC-10 terminal, at speeds up to 9600 baud. It can then be commanded to receive a file of catalogue data from the DEC. This file can go straight down a fast line to the network's Winchester disc, or be written to floppy discs, which are easily transported to an Apple on the PLAN network and can thus be copied onto the network's Winchester disc.

4.5.3 Final Okapi file creation on PLAN

The final stage in creating the Okapi file is carried out on the PLAN network. The codes in Field Eight (codes and control number field) which were transmitted as ASCII, for reliability, are converted to binary, for compactness. A field directory is created for each record. Each record is padded out to a multiple of four bytes long. The records are concatenated into one multi-volume CP/M file, and a separate record directory file is created. This directory file contains a four byte address for

each record in the source file. It permits direct access to a source file record by a notional record number. The directory file is used during maintenance and as a safeguard. It is not used by the online search program which accesses the source file via the indexes (which contain disc addresses).

4.6 File size, mean record length and other statistics

The Okapi file contains about 90,000 records and occupies nearly 19 megabytes of disc space. The mean record length is 214 bytes. The longest record is 916 bytes. Data are selected from 24 MARC tags plus local fields. The simplified record format and lack of duplicated data achieve a massive reduction in file size: the Okapi file occupies only 26% of the space required for the MARC file. However, PCL holds an unfiltered file: i.e. fields additional to the PCL input standard have not been deleted. If PCL had maintained a filtered file the reduction would be much less dramatic.

The separate record directory file, mentioned in Section 4.5.3, occupies 352 kilobytes.

4.7 Order of records in the file and on screen

At present the sequence of records in the Okapi source file is the same as the exchange tape, i.e. the records are in control number order. This is also the sequence in which records exactly meeting the search criteria are displayed.

It had been hoped to display records in descending order of date of publication, (i.e. most recent first) and this was the main reason for isolating a single date of publication for each record (see Appendix 3).

The source file would be sorted by date and maintained in this sequence. This would complicate the batch procedure for updating the file since records to be updated can only be uniquely identified by control numbers.

The alternative of continuing to maintain the file in control number order, and sorting any set of records into date order immediately prior to their being displayed, is not feasible because of the time it would take if done online.

In the event lack of time prevented the reordering of the Okapi file. However, it should be noted that the sequence of records is of much less significance than in a conventional catalogue. In a known item search if the search key is correct then only the desired item will be displayed, so in this case ordering is (usually) irrelevant. If the search key is incorrect the order in which partial matches are displayed is significant. In subject searches ordering by date is more useful than ordering by *main entry*, but neither of these is of much utility. There is a further discussion of ordering in Section 9.4.5.

4.8 File updating and maintenance

From its conception Okapi has been a *catalogue* not a catalogu*ing* project. Catalogue maintenance and updating are outside the scope of the project; nevertheless, they are essential if Okapi is to be a viable system.

A simple minimal approach could be based on regular exchange tapes containing all new, changed and deleted records. These records would be converted and transferred to the PLAN and used to update Okapi's source file which would then be re-indexed, probably on a monthly basis.

References

1 *UK MARC Manual.* British Library. Bibliographic Services Division. 2nd ed. 1980.
2 **Seal A, Bryant P** and **Hall C.** *Full and short entry catalogues.* Centre for Catalogue Research. Bath University Library, 1982.
3 **Jones K P** and **Bell C L M.** The automatic extraction of words from texts especially for input into information retrieval systems based on inverted files. In: *Research and development in information retrieval.* Proceedings of the third joint BCS and ACM symposium. King's College, Cambridge, 2-6 July 1984.

MARC record excluding directory (non-ASCII chars. shown as \) [the record has been slightly modified for this example]

```
001    0905739000#
002    a   002400372#
008    761222s1976  en ah  We   00001  eng#
015 00 $aB7700853#
050 00 $aNC1280#
081 00 $a741.6$b01$b48$c18#
082 00 $a741.6$b01$b48$c18#
083 00 $aGraphic design. Symbols$bDictionaries#
087 00 $aX.419/3260$bWoolwich#
110 20 $aNowhere College#
245 12 $aA dictionary of graphic clich\es
       $ecompiled by Philip Thompson & Peter Davenport#
248 10 $h[ABC]#
260 00 $aLondon$d61 North Wharf Rd. W2 1LA$bPentagram Design#
260 01 $c[1976]#
300 00 $a[24]p$bill$ifacsims$c21cm$esd#
350 00 $a\1.00#
440 10 $aPentagram papers$v1#
500 01 $aFold. covers#
650 00 $aCommercial art$xDictionaries#
650 00 $aSigns and symbols in art$xDictionaries#
690 00 $z21030$agraphic design$zp1030$asymbols$z60030
       $adictionaries#
691 00 $a0773409#
692 00 $a0281034#
692 00 $a002399x#
700 11 $aThompson$hPhilip#
700 11 $aDavenport$hPeter#
957 00 $a760920c#
990 00 $aDICp#
998 00 $aG7776878REF#
999 00 $aG#
998 00 $a22.00293401RP#
999 00 $a2#
```

Okapi record

```
Directory |     17 36 18 07 24 17 15 14 70
Field 1   |     $Nowhere College#
      2   |     _A ^dictionary of graphic clich\bes.#
      3   |     @Pentagram design.#
      4   |     ^[ABC]. @[1976] ^Pentagram papers.#
      5   |      |Thompson P|Davenport P#
      6   |     741.6$01$48 DICp#
      7   |     22.00293401RP#
      8   |     4 226 0 0905739000# [NB only three bytes for the three codes]
      9   |     ^Graphic design. ^Symbols. ^Commercial art.
          |     ^Signs and symbols in art.#
```

Figure 4.1. Complete MARC to Okapi example