J. W. JOLLIFFE
Keeper of Catalogues
Bodleian Library
Oxford, England

Project LOC: Centralized Processing of Local Collections

Project LOC arose from the convergence of two factors: a growing awareness in the great libraries of Great Britain of the potential utility of computers in dealing with their large-scale processing problems, and a very long-standing need for the provision of information about the collections of early books in the college libraries of the Universities of Oxford and Cambridge.

These college libraries are variously cataloged, both in physical form of the catalog and in degree of competence with which the records have been created. Some libraries have published their catalogs, others have barely listed their holdings. The central libraries, the Bodleian at Oxford and the University Library at Cambridge, have no responsibility for the college libraries and no authority in matters concerning them. In both universities, however, there have been movements, over the past 300 years, to produce union catalogs of the collections in the universities as a whole.

One way of bringing this about, an old one in fact, is by the notion of borrowed cataloging; i.e., by using some existing catalog as a basis for cataloging style and arrangement, adding locations of duplicate copies and records in the same style for works not listed in the base catalog. The novelty, in the case of Project LOC, lies in the use of a computer version of the base catalog and in data processing techniques for updating and expanding this version.

The base catalog chosen was that of the British Museum. Three principal reasons led to this choice: its availability in published form, the known richness of the Museum's collection of early books, and the desire to bring

into a practical working relationship the Museum Library and the two next largest libraries in the country.

There was a fourth agent in this: the Andrew W. Mellon Foundation, or the Old Dominion Foundation as it was known when the project began. This foundation had already sponsored the conference at Brasenose College in 1966, at which plans and experiences were exchanged and discussed by a group of librarians and library data processing experts from both sides of the Atlantic. In the following year, it sponsored the visit of Foster Palmer of Harvard University and Lawrence Buckland of Inforonics Inc. to the three major libraries to discuss individually with them their plans for library automation. These two reported privately to the foundation and to the three libraries their reactions to and their views on the progress envisaged.

The chief problems were two: a lack of experience in library data processing, and a lack of personnel able to be involved in it. For this reason a joint approach, a joint venture, was encouraged. An investigation into factors affecting the creation of the long-desired union catalog of early books was to be carried out. Freed by the generosity of the foundation from the consequences of waste of funds, it was possible to try out, on a fairly large scale, different methods of data collection and matching of records; the project was able to make mistakes and explore blind alleys without committing any further expenditure to the erroneous procedures it might adopt.

The use of a computer to manipulate a union list, to generate subcatalogs for the libraries whose collections were to be included, to provide indexes of various sorts to the list, and to provide statistical information about the chronological, topographical and linguistic characteristics of the works contained in the list, seemed inevitable. The main list itself might be constructed in other forms and by other means, but only with a computer file would it be possible to provide the other indexes and information easily and relatively cheaply.

So the project was concerned with two main problems: how to collect and process the information about the collections and what methods to adopt to avoid repetitious cataloging of duplicates. Two physical methods of recording and two levels of completeness of recording were tried out.

I was appointed director of the project, working with a committee composed of representatives of the British Museum, the Bodleian Library and the Cambridge University Library, together with the Harvard University Librarian, Douglas Bryant. The employment of staff to collect, punch and proofread the information about the college and other libraries devolved upon the two university libraries; in each of these, the day-to-day organization and supervision of work was the responsibility of a senior permanent member

of staff. The concern at the British Museum was with the conversion of the Museum records into machine-processable form, the provision of rules and procedures for the local teams, the correction of the machine files, the specifications for the computer programs and the evaluation of results.

The immediate problem was one of selection: upon what data should the project work? The one restriction in scope which seemed permanently useful was that of attempting to cover only books published before the nineteenth century. Three catalogs had shown that unique items were hidden in the college libraries, and the history of the growth both of the large libraries and of the college libraries suggested that the proportion of unique items in the college libraries would decline rapidly if the nineteenth century were to be included. Certain libraries might have been excluded, but this would have removed the element of surveying the whole field, and might have led to false impressions about the scale and nature of the problem of producing a full union list. Similarly, restriction by language or further restriction by date of publication would have introduced a clear bias into any conclusions, since the problems associated with books of a particular age or in a particular language would not have been confronted. The only way in which a representative section of the collections could be studied without the considerable problems of devising and administering correct samples in each library was by choosing within a segment of the one type of file which was common to all libraries—the alphabetical author catalog. Whatever strictures might be applied to the quality of some of these catalogs, their very existence gave the project an opportunity to choose books which would represent quite faithfully the distribution and overlapping of the collections as a whole. A single letter was chosen; all headings beginning with letter O were to be examined and the shelfmarks of pre-1801 books were to be noted so that the books themselves could be used. Other letters which, on the face of it, would have supplied about the same number of books to be dealt with, would have introduced problems of bias towards or against particular languages or particular cataloging problems. The letter O had the advantage of including a voluminous classical author, Ovid, a fair amount of Greek and German, and even some Russian and Irish, although the proportion of Irish surnames in the final list was by no means as large as had been expected. Because of the diverse standards of the college catalogs, we had in mind to bypass them in a full-scale operation; ironically, we had to work through them in the pilot project.

For the comparison of two methods of recording, this selective approach was considered undesirable. A round-the-shelves operation would need to be as cheap and as quick as possible. The methods to be compared were a photo-

graphic one, recording the title page together with a form giving details of library, classmark and supplementary details of imprint and authorship taken from the colophon and elsewhere in the book, and a manual one, in which the required information was filled in by hand on a form. Two levels of detail in the form were also to be compared: was it possible to carry out matching against the base file with a lesser amount of information?

For this part of the project, in which estimates of the cost of continuous working for the methods and levels of detail were to be obtained, two complete college library collections were used. The early books of Hertford College, Oxford were at that time on deposit in the stacks of the Bodleian Library, while the college library building was undergoing redecoration and repair. This collection, in better surroundings than any college library, was used for comparisons between the costs of creating records to two levels of detail. The other library, at Cambridge, was that of Peterhouse, where the main part of the collection is housed in a single room. This library was used for comparing the costs of preparing machine records for the books using microfilm as a recording medium and of the similar operation using hand-written forms. There were approximately 3,500 books in the Hertford collection and approximately 4,400 in the Peterhouse one. For the investigation into overlapping in the whole range of libraries, some 22,000 records were ultimately used: approximately 7,000 from the British Museum, 2,600 from the Bodleian and 2,300 from the Cambridge University Library. All the other libraries together yielded just over 10,000 records for the O books.

The relevant section of the British Museum catalog was read, and papertape records made of all pre-1801 records. A similar scan of the Cambridge University Library's working catalog was made; the relevant entries were photocopied and half were sent to the British Museum for punching. The Bodleian Library's pre-1920 catalog was just beginning to be converted to machine-processable form; the O entries were taken out of sequence, and the pre-1801 records were then punched out on papertape for LOC processing.

This processing was carried out on the Cambridge University's Titan computer. The fact that the director and the processing facility were in different places had certain consequences for the course of the project; chiefly in the delay in turnaround for some stages, but on the positive side it meant that the programmer was left to her own devices in many respects and was able to work more quickly and more confidently than if she had had to explain and perhaps justify all the details of her decisions. Since she had been chosen for her knowledge of the machine and the multi-access system, besides her manifest ability as a programmer, it was better that she worked on her own without the necessity of educating the director in details of the system

which were, properly, none of his concern. A terminal was provided for her so that she could work at her home in a village four miles outside Cambridge.

In deciding what information should be recorded, it was necessary to look ahead to the ways in which matching between newly input records and the base file was going to be carried out. By choosing the Cambridge computer, the project was committed to a batch mode operation and, in consequence, algorithms were required which would yield a high probability of identity of records produced in different circumstances, without human intervention. The information on which these algorithms would work had, therefore, to be recorded.

However, since the matching was between or within two classes of records, those included in the base file and the library catalog records converted directly and those created by the project itself on an examination of the books, any matching between the classes was constrained to the set of information already available in the base file. Matching within the class of records produced by the project could take place on information chosen to make such matching easier. A further consideration in deciding upon the information to be recorded was the possible use of all this information in a catalog entry. Information could be recorded for the catalog entry which would not be used in matching and vice versa.

One of the aims of the project was to attempt to determine the likely differences in cost between a single-pass method of recording (in which all the information that might be required would be acquired at the shelves, whether or not all of this would be used to establish matches) and a two-pass method (in which only the information necessary to establish a match would be recorded on the first occasion, while a return to the shelves would be necessary for works found not to match against the growing base file). The work with the Hertford College books showed that recording all the information by hand took roughly twice as long as recording a set of information for matching alone (date, title, author). This, however, gives only one factor in the equation for determining the superiority of a one-pass or two-pass method; the other is the proportion of duplication. With the O books, all required information was set down to avoid the practical necessity of a second pass, but matching was tried on subsets of this information.

The "cataloging sheet" on which this information was written approached most nearly, in details of description, the completeness of current cataloging codes: there was provision for a library symbol, a classmark, a title (split if necessary into three parts), an author name, date, language, edition, number of volumes, place of publication, and a publisher's or printer's name. In addition, there was a fingerprint. The labor force for this recording

operation was composed of "intelligent but untrained" people. In a low-cost operation, the information to be recorded was reduced to that which people untrained in library work might most easily learn to recognize; thus collation statements, format statements, indications of editors or translator were omitted.

Most of this information was contained in the records in the base file—language is the exception among the descriptive items. Matching between these records needed to be based on these items of information. The shorter records contained only the date as it stood on the book, a short title, the author's name and the fingerprint.

It might seem naive, given the known difficulties of establishing an "author heading," for "author's name" to be included in the set of information which "untrained" people were judged capable of recognizing. The project compensated for this by paying as little attention as possible to what they had recorded as "author" when attempting to match records. This meant that for the shorter records only date and title were available for matching: the fingerprint was a separate exercise to be described later.

It was plain that a straightforward literal comparison of the strings of characters representing the titles would lead in the majority of cases to a match not being found. Moreover, if one catalog had included an epithet or a name at the beginning of the title, while another had omitted it, the two records might be widely separated in a file sorted by title. A comparison of each record with every other might have been tolerable in the project, but was unthinkable in a full-scale operation. Much thought was therefore devoted to the problem of organizing the file so that there was a good probability that two records that should be compared would be, while as many futile comparisons as possible would be avoided.

The method adopted was to create a "keyed-title" record. Various normalizing procedures were carried out on the title before keys were generated; all letters were changed to upper case, punctuation and diacritics were discarded. Up to three keys were created for each title; all words shorter than four characters were discarded; the first remaining word was taken as one key; and of the rest those which sorted first and last alphabetically. The effect of this was that with titles of average length, there was a better than 0.5 probability that two differing versions of the same title will produce at least one identical key.

This virtually solved the problem of file organization. The matching problem itself remained. In the event, four matching procedures were used: a comparison of titles, which was the one for which the file organization just mentioned was intended; a comparison of search codes constructed by algo-

rithm from the records; a comparison of fingerprints recorded from the books; and, as a necessary step both to an evaluation of the effectiveness of these procedures and to the preparation of a specimen union list, visual comparison of records and human evaluation of the information they contained.

I will describe the first three methods and the way in which their relative effectiveness was determined. I have already mentioned the way in which records for title comparison were prepared—by using single words as keys. This word was not the sole element in the key—the other was the date, and this leads to a little complication.

When recording directly from books, the Project LOC staff was required to set down the title page date as it stood, whether in Roman or Arabic numerals; it is possible to derive an Arabic numeral from a Roman, but not in all cases of imprint dates is it possible to do the reverse. However, all the three large libraries, like most other libraries, had already normalized these dates as Arabic in their catalogs. The project used three kinds of date in its machine records: a text date taken directly from a book, a catalog date as given in a converted catalog entry, and a search date derived from either of those two. The search date was used as the second element in the keys for the title comparisons, thus permitting comparison of catalog records and of records created during the project.

When two keys were found to be the same, the titles were then examined. As before, words of fewer than four letters were ignored. Now, it is not possible simply to count the number of words to be found in both titles: "Articles concerning the surrender of Oxford" and "Discussions at Oxford concerning the 39 Articles" both contain "Oxford," "Articles" and "concerning," but in reverse order. The order of words in titles was thus an important element. One other effect had to be allowed for: the differing truncations which might result from different catalogers. "A petition . . . presented to the . . . House of Commons" and "A petition humbly presented to the honourable House of Commons" are probably renderings of the same title. If we strip them and number the longer words with their original positions, we have

PETITION(2) PRESENTED(3) HOUSE(6) COMMONS(7) and
PETITION(2) PRESENTED(4) HOUSE(8) COMMONS(10)

for the lists of common words. Both the number of words in each list compared with the number in its present title, and the span of the longest string of common words were taken into account to produce an index number which could range from just over zero to 1.00. This number was printed out together with the full titles for subsequent evaluation by eye.

The second comparison method used search codes. These consisted of alphanumeric strings of fourteen characters, taken from the date, title, an

edition statement, if any, the place of publication and the author's name. These codes were sorted into a single sequence and the sorted list was scanned for identical items. The procedure of creating and comparing them was much faster than for the title word method, and no complexities of file handling were presented.

These two methods were used for comparing catalog records with each other and with the records created by the project for college books. The third method was used only for comparisons between college book records, since it used information not available in the catalogs. This information was a "fingerprint"; i.e., three groups of six characters, each group taken from a different page of a book: the recto after the title page (or the first recto if there was no title page), the third recto after this and the fifth after that. The characters were the last two on each of three lines: the last on the page, two lines up and two lines up again. It had been thought possible that such a character string drawn, as it were, at random from the text might function as a unique identifier of the book. The three "pages" were kept distinct and matching took place on the fingerprint and the text of the imprint date. The fingerprint pages were rotated to bring each to the head of the key in turn, and all of these three keys were sorted. The numbers 1, 2 and 3 were added into the sort key to avoid spurious matches on different pages. The sorted list was scanned, and any adjacent items with at least one page of the fingerprint in common were printed out. A full match was one in which the fingerprints were identical throughout and the text dates were the same.

In each case, of course, the sort item included an identifier for the full parent record so that verification of matches by visual comparison of records could be carried out.

All three types of matching were performed on the files containing the two complete college libraries, Hertford and Peterhouse. Although the overlap between the collections was small, each library contained some duplicate copies which had been recorded separately. The advantage of using these small closed sets was that it was not difficult to establish a list of duplicate items against which the lists produced by the different matching methods could be set. The first two types of matching were also performed on the combined files containing O books, while fingerprint matching was carried out on the combined files of college O books.

In the case of the Hertford College and Peterhouse comparisons, the three systems gave the results seen in table 1.

These figures show clearly that the keyed title method was unsatisfactory because of the large volume of spurious matches. In a full-scale operation each such match would need visual verification, and, in the majority of cases,

Matching Method	*Matches*	*Spurious*	*Percentage Matched*
Search code	40	2	22.2
Fingerprint	164	3*	95.2†
Keyed title	138	297	76.6

Table 1. Results of the Three Matching Methods
*In each case, the items were variant issues of the same edition.
†This is calculated to a base of 172, since in 8 cases, one copy had not had its fingerprint recorded.

rejection. Introducing a cut-off point into the computed index number would result in the loss of some genuine matches, since, in both the spurious and the genuine matches, the value ranged between 0.01 and 1.00.

The search code seems equally fallible; in a batch-mode operation, the matches must be as definite as possible. No variations can be admitted which might remove the obstructions to a match of similar search codes, for each portion of the bibliographical information which contributes a character or two to the search code records details in which differences are significant. Nevertheless, in view of superior results with similar codes at OCLC, a careful review of all O book search code matches was carried out. In these, the percentage of matches made rises above 50, although spurious matches also rise to 2 percent. This is still not adequate for a full-scale operation.

Partial matching in the fingerprint was another matter. Miscounting the pages or the absence of a leaf in a given copy can easily lead to discrepancies in the characters recorded for different pages. Again, since the fingerprint has no meaning, it is liable to be degraded in recording or punching, and significant differences, e.g., comma for full stop, can easily be overlooked at the proofreading stage. The figure of 95.2 percent is for matching on one or more pages; that for two or more pages is 74 percent; matching on all three pages was effective in 46 percent of cases. In the majority of cases where a page did not match, the difference lay in a single character and this could be attributed very often to misreadings of handwriting, to shift errors or to mispunching of adjacent keys.

The fingerprint therefore seems to be the best basis on which to perform batch-mode matching. The particular form adopted in Project LOC could be changed, say, to four groups of four characters or even five groups of three, while partial matching could be performed either on the basis of all groups except one, or of all characters except, say, two and those not adjacent in the same group. Shortening both the fingerprint itself and its component

groups should lead to higher accuracy in transmission through the various stages of recording and entering into the machine file, with a resulting improvement in matching efficiency.

In September 1972, a conference was held at Brasenose College, Oxford, to review the results of Project LOC. A provisional version of the report was circulated to participants, who were invited from libraries in the United Kingdom, the United States and Canada. At this conference, various aspects of the project were summarized by members of the LOC executive committee, and full discussions were held. I think it fair to say that no essential point of what had been done, what had not been done, and what might be done was left unexplored. After the conference, many of the participants responded to a request to submit observations in writing.

The LOC committee was quite clear about what had been done and why; in the more doubtful area of future progress, it was greatly helped by the reactions of those who viewed the project from outside. These centered on two main issues and on several lesser ones.

The first main issue raised was that of the scope of a full-scale operation. Should it be to produce a union catalog of pre-1801 books in the libraries of Oxford and Cambridge? Should it be only an eighteenth-century English union catalog as a preliminary to an eighteenth-century *short title catalog*? Should it include early books in other British libraries? Or in the major American research libraries? My view is that it should be confined to the original problem area, the libraries of Oxford and Cambridge, but an essential element in planning and implementing this scheme is its extensibility. The procedures, record and file structures must be designed both to permit acceptance of data from sources outside the range of libraries to be dealt with and to permit the transmission of the file or subsets of the file to other libraries and institutions. The LOC Project fulfilled neither of these objectives for two reasons: the first and simplest was the extra degree of planning and programming that would be required; the other was that nothing as tentative as the pilot project should be permitted to encroach by example and availability on the question of standards.

The other main issue raised by the participants of the review conference was that of the mode of computing activity that had been used and might be used again—that is to say, off-line batch mode processing on a large central computer mainly employed on other tasks. It was suggested that various factors such as staff training, methods of matching, and process control would be changed, and for the better, by interactive computing. With this view I have considerable sympathy, especially because of the continuing fall in cost both of minicomputers and of random access mass storage devices. Indeed, the

processing facility for a full-scale union catalog must, I think, be a dedicated minicomputer system for on-line data entry and correction and for the first stage of matching, with the main full files being maintained on a bureau machine for batch access.

Another point which needs attention is the quality of the staff to be used and the nature of their training. Project LOC used people without library training and gave them a minimum of training on the job. This want of training arose from the short time period allowed for working to collect and encode the data. The former feature, using people without library training, was a matter of policy, and insofar as such people are in greater supply than those with library training, it should still be a feature of a full-scale operation. Not that there should be no one engaged who has library training; the editorial and administrative posts need qualified and experienced librarians to fill them.

Finally, let us consider the size of the problem. At the beginning of the project there was no firm knowledge of the number of early books to be found in the two universities. Adams's catalog of foreign sixteenth-century books in Cambridge gave some evidence both as to overlapping and as to the distibution of copies between the central university library and the other libraries. However, what was true for the sixteenth century might be less so for the latter centuries. By comparing the distributions within centuries and in Oxford and Cambridge both separately and together, it was possible to narrow the limits of error in our extrapolations.

The final estimate is of some 1,600,000 copies in the libraries of Oxford, Cambridge and the British Museum, representing some 780,00 distinct editions. Of these, some 160,000 editions are to be found only in the college and departmental libraries of Oxford and Cambridge. So large a figure justifies our concern to find the means of recording and disseminating information about these collections.