AUTOMATING A RETROSPECTIVE CANADIAN UNION CATALOG: A PROPOSAL (UNE PROPOSITION POUR CREER UN CATALOGUE COLLECTIF CANADIEN DES DOSSIERS-MACHINE RETROSPECTIFS)

William J. Cameron
School of Library and Information Science
The University of Western Ontario, London N6A 5B9

ABSTRACT

The paper describes methods of linking by machine, the machine-readable data base of the HPB project, the cataloging records of the *National Union Catalog: pre-1956 Imprints*, and selected bibliographical tools. One result, in the form of a machine-readable register of Canadian locations, can be selectively expanded into a substitute in machine-readable form for the present Canadian Union Catalog (CANUC). The theory and practice underlying the creation of the HPB main file and its highly specialized collocation files are described and possibilities for integrating the project into evolving systems for universal bibliographical control of retrospective materials suggested. (On explique dans ces pages des méthodes pour lier par ordinateur les données lisibles dans la machine du projet HPB, les dossiers du catalogue collectif *National Union Catalog: pre-1956 Imprints*, et les renseignements bibliographiques pris dans des bibliographies choisies. Un des resultats, en forme d'un registre de sigles des bibliothèques possédantes, est capable de devenir peu à peu un index établi par ordinateur pour développer un Catalogue collectif automatisé de dossiers retrospectifs qui remplacera les tiroirs de fiches de la Bibliothèque nationale. On explique aussi la théorie et la pratique du projet HPB, c'est à dire le fichier principal (main file) et les fichiers d'arrangement spécialisés (collocation files), avec ses possibilités de liaison avec les systèmes de contrôle bibliographique universelle qui sont en train de se développer actuellement.)

THE HPB PROJECT

The HPB (= Hand Printed Books) project is an attempt to gain the maximum amount of bibliographical control (by means of a minimum amount of bibliographical data) over the universe of letterpress-printed books of the hand-printing era (i.e. up to the end of the 18th century). The machine-readable data base consists of a main file of catalog entries and a multiplicity of so-called collocation files.

AUTOMATED CANUC/CCC AUTOMATISE

The singular feature of the main file is that each entry (at present there are about 30,000 of them) represents a new way of defining a unit of the bibliographical universe, and the major purpose of the collocation files is to organize these units into a variety of combinations that will be useful to small groups of scholars in the humanities. The unit is called a "bibliographically distinct volume" and may be roughly defined from the point of view of the volume's producer as "a book, pamphlet, or other printed item conceived of and printed as a single physical entity" or from the point of view of a modern bibliographer as "a volume consisting of a title page or its equivalent followed by a complete series of signatures and pagination". This unit may be contrasted with sub-units such as a bibliographically distinct "number" (the analogy with a periodical or serial seems obvious) or with the unit of a "book" which seems impossible to define unambiguously. But more importantly, the unit can be contrasted with a *rhetorically* distinct volume. This term attempts to describe a text with a beginning, middle, and end which, with preliminaries and appendices, can be (but frequently is not) contained in a bibliographically distinct volume or a series of such volumes. The unit may also be contrasted with a *physically* distinct volume (where the physical binding may coincide with a bibliographical unit, but more frequently with the rhetorical unit or subdivisions of it).

The singular feature among the cataloging principles for each entry is that equal emphasis is put upon the recording of data for the authors of the (intellectual) *work*, and the authors (printers, publishers, booksellers, etc.) of the physical *book*. Precise and concise data are recorded for the primary author, collaborators, translators, editors, etc. through letter by letter transcription of names on the title page, and distribution of the names into fields for primary, secondary, tertiary and quaternary author, with deduced and conventionalized relationships indicated (e.g. "and" for collaborators, "trans", "ed", "illus", "set by" for others). Precise and concise data for "physical" authors are ensured by transcribing letter by letter (including punctuation whether ambiguous or not) all imprint information concerning personal names and the words that express their relationship (e.g. By x for y and sold by z).

The amount and form of data in each field is the minimum required for *identification* (not description). Editions, issues, or states not distinguished by this minimal data are distinguished in a "points" field by providing some highly selected points of difference between the otherwise very similar items in accordance with bookcollectors' and bibliographers' traditions (modified by machine requirements).

COLLOCATION FILES

A collocation file may exist in one of three kinds: (a) hand-produced, (b) machine-readable, and (c) machine-produced. The simplest collocation file would be a list of HPB main file addresses written down manually in the form:

AUTOMATED CANUC/CCC AUTOMATISE

a
b
c
d
e

where data elements a, b, c, d, and e are HPB addresses. (These are in the form of six-character sets NNNNLL or LLNNNN, where L = a letter and N a number.) The purpose of this hand-produced file is to record the results of bringing into contiguity HPB entries that are separated in the main file. This file, if it is a useful record for a scholar, may be made machine-readable in the form:

```
0010 a/b/c/
0020 d/e/
```

where the numbers represent machine-generated addresses for each line record. (They enable text-editing to be done on-line where necessary.)

From the machine-readable file can be generated the final form of a collocation file - a reproduction of the hand-produced file in which the data strings (representing HPB addresses) have been converted by machine to display the full HPB main file data (or specified fields in specified display sequence or form).

Elaborations of each kind of collocation file are possible. In the hand-produced file, the data elements can be natural-language headings, sub-headings, comments, notes, etc., or conventionalized bibliographical data, references to a published bibliography or library catalog, cross-references to other parts of the file, etc. Corrections, additions, deletions, substitutions, etc. may be made to the hand-written collocation file as a scholar increases his bibliographical control over the colligation of the materials for which the bibliographical data in the file are a surrogate. At some point in time, the file will be worth converting to machine-readable form so that an elaborated machine-produced substitute can be made available for further editing and elaboration. Any hand-recorded corrections entered on the printout can later be incorporated into the machine-readable collocation file by on-line text-editing.

Collocation files can thus be used to generate complex bibliographies for particular purposes, with a great deal of flexibility for modification and updating as new information becomes available. But they can also be used for much more general purposes. The *quintessential* information already available in a very elaborate published bibliography can be recorded as a collocation file which thus becomes an index to that bibliography. This is especially useful if the published bibliography orders its entries on a principle at variance with the user's need for consultation. For instance, bibliographies in chronological order may need title, author, printer, or subject indexes. Corrections and expansions of the original bibliography and other updating information can also be incorporated into the collocation file, thus extending the usefulness of the published work.

Formalization of some of the natural-language connective comments in a collocation file can result in an optimizing of machine manipulability as well as compactness of storage. For instance the field delimiter (/) which is also used as a printout instruction (Begin a new line in the printout) can be varied by using the following symbols:

= (= is the same as)
+ (= and)
> (= please refer to)
< (= the data that follow are locations)

Thus, the kind of collocation file developed to index a standard bibliography might be rendered thus (in machine-readable form):

```
0010    a=b/c=d/e=f/g=h/i=j/
```

or (in hand-produced form):

a=b
c=d
e=f
g=h
i=j

where a,c,e,g, and i are references to a standard bibliography, and b,d,f, h, and j are HPB numbers. If the bibliography does not distinguish bibliographically distinct volumes, or if it provides different entries for editions, issues, or states (recorded as "points" in the HPB entry), the index to the standard bibliography might be rendered thus:

```
0010    a=b+c/d=e/Edn A/f=e/Edn B/g>a/
```

where a, d, f, g are references to a standard bibliography; b,c,e, are HPB numbers; Edn A and Edn B are references to the "points" field of e. (The implication of g>a is that g is subsumed under a, as explained by the collocation of b and c.)

Another kind of formalization makes it possible to extend the HPB data bases into evolving systems for universal bibliographical control. If bibliographical references can be formalized sufficiently to be machine-recognizable like the HPB address, the cataloging data in the main file can be automatically substituted for the cataloging data in the bibliography or catalog to which the machine-recognizable references refer. For instance, the nine-character address system in *The National Union Catalog: pre-1956 Imprints* (with its ad hoc variations) can be developed into a 12-character address system that is not only machine-recognizable, but capable of expansion and correction. A printed catalog entry in this publication (or any other with such a formalizable address system) can be made machine-readable by means of a switching entry in the collocation file of the form

a=b

where a = a machine-recognizable reference, and b an HPB address (or string of addresses in the form x✦y/z) or an address of some other machine-readable catalog.

<u>AUTOMATING CANUC</u>

These brief remarks about the HPB project may require further explanation or exemplification, but they should be sufficient to permit us to put forward a tentative sketch of how the 13,000,000 or so catalog cards in the retrospective portion of the Canadian Union Catalog of Books in the National Library could be converted to machine-readable form.

First, the consolidation of these cards into a vast (3,000,000?) file of master-cards with full listings of Canadian locations entered on the master-card could be continued, but as the Library of Congress has now reached the half-way mark in publication of *The National Union Catalog: pre-1956 Imprints* it would seem that duplicate cards could very well be matched against entries in that publication rather than with the putative master card. When a match occurs, a machine-readable record could be added to a "CANUC register of locations" in the form

a<b

where a is a machine-recognizable NUC address and b is a Canadian location. Requests for locations of such items can be made in the form of NUC addresses, which can be determined by the requesting library or by the searchers at the National Library.

A time-and-motion study has established that about 70-80% of attempts to match typical catalog cards from a contributing library result in a trouble-free match, and that such matching can be done at the rate of about one per minute. The troublesome examples can be set aside for special treatment. Some will prove to be "not in NUC", some will be found in unexpected parts of the filing order, some will be impossible to match because of inadequate data either in the contributing library's catalog data or in NUC. The first category can be dealt with by allocating a supplementary NUC number to ensure that it will appear in the appropriate filing position in the CANUC Register but not conflict with already-allocated NUC addresses. The second category can be dealt with by providing a "see" reference in the CANUC Register in the form

$$a>b$$

where a is a supplementary NUC number and b an already allocated NUC number (or vice-versa). Thus the CANUC Register can become a device for overcoming the maddeningly inconsistent or difficult-to-use filing system of *NUC: pre-1956 Imprints*.

The first and third categories can be used to begin the task of providing adequate cataloging data in machine-readable form that will eventually be substituted for the cataloging data in *NUC: pre-1956 Imprints*. This publication frequently displays more than one master-card for the same bibliographical entity, but "see" or "see also" references can be incorporated into the CANUC Register when doubts about the implied differences have been resolved. However, the first occurrence in the NUC filing order of a multiplicity of entries for the same item can be linked with a machine-readable entry of a less ambiguous form (it should not be surprizing that I suggest the tried-and-tested HPB format for pre-1801 books) so that the CANUC Register (a specialized kind of collocation file) can be elaborated into a selective Index to a bank of machine-readable catalog entries. Only about 30,000 pre-1800 entries are available in the HPB format. But even fewer pre-1900 items are available in MARC II format. It would seem at least worthwhile considering both data bases, one for pre-1801 and the other for 19th century books for incorporation into such an Index. The CANUC Register would be expanded thus:

$$a=b<c$$

where a = an NUC address or supplementary address

b = HPB address or MARC address
c = string of locations

b may be in the form x+y/z where x and y are HPB "bibliographically distinct volumes" and z is a natural-language comment on the previous data.

In accordance with these ideas, a feasibility study for an HPB/NUC Index is at present being developed as part of the HPB project. The collocation file which will result, with its accompanying retrieval system, is being built up by merging three simpler collocation files. The resulting file may be graphically represented as follows:

AUTOMATED CANUC/CCC AUTOMATISE

A.	Merged HPB/NUC Index file $a=b/c<d$
B.	NUC/HPB matching file $a=b/c$
C.	NUC locations register $a<d$
D.	HPB locations register $b<d$

where a = NUC address or supplementary address
b = an HPB address or string of addresses in the form $x+y+z$
where x, y, and z are an HPB address
c = editorial comment on previous data
d = a string of location symbols

CONCLUSION: COST-EFFECTIVENESS

The technical feasibility of this proposal should be demonstrable very soon. Its cost-effectiveness in relation to such proposals as MARC/ RECON will remain to be demonstrated. The tentative estimates of cost already available would suggest that it compares very favourably indeed.