Virginia Disc 2: Preparing, Presenting and Retrieving MARC Standard Library Data on a CDROM

By Raul B. Quizon and Edward A. Fox

TR 90-5

Virginia Disc 2: Preparing, Presenting and Retrieving MARC Standard Library Data on a CDROM

CR Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems - human factors; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - indexing methods; H.3.2 [**Information Storage and Retrieval**]: Information Storage - record classification; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - retrieval models, search process;

General Terms: Experimentation, Performance

Additional Keywords and Phrases: bibliographic records, CDROM, CD-ROM, compact disc read-only memory, library catalogs, Micro-VTLS, Personal Libarian, retrieval effectiveness, serials holdings

VIRGINIA DISC 2 :
PREPARING, PRESENTING AND RETRIEVING MARC STANDARD LIBRARY DATA ON A CDROM

by

Raul B. Quizon

Report submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTERS

in

Information Systems

APPROVED:

Edward A. Fox

E. A. Fox, Chair

Deborah Hix

D. S. Hix

Marilyn L. Norstedt

M. L. Norstedt

October 1989

Blacksburg, Virginia

VIRGINIA DISC 2 : PREPARING, PRESENTING AND RETRIEVING MARC STANDARD LIBRARY DATA ON A CDROM

by

Raul B. Quizon

Committee Chair: Dr. Edward A. Fox
Computer Science

Most of today's commercial information storage and retrieval systems are still based on Boolean queries and inverted files despite the recent advances in information retrieval research. This report discusses the creation of the CDROM publication, named Virginia Disc 2, with the intention of comparing two retrieval systems: Micro-VTLS and Personal Librarian. Micro-VTLS belongs to the class of conventional Boolean systems with inverted index files. Personal Librarian uses newer concepts learned from the Syracuse Information Retrieval Experiment (SIRE). The motivation for comparing the two retrieval approaches is to demonstrate the superiority of newer concepts like term weighting, ranking, and similarity measurements employed by Personal Librarian.

Virginia Disc 2 uses MARC, a widely accepted standard format for communicating library information. The automation of bibliographic information is a prevalent application in libraries and a CDROM implementation using new information retrieval concepts could generate interest in adopting these new approaches.

This report discusses a comparison of Micro-VTLS and Personal Librarian in the areas of retrieval speed, recall and precision. The preliminary findings suggest the superiority of Personal Librarian. This report also reviews the process of preparing the Virginia Disc 2 CDROM. It discusses the considerations, the preparation effort and the final product of storing MARC library data on a CDROM.

ACKNOWLEDGEMENTS

The Virginia Disc 2 project benefited substantially from the lessons learned in Virginia Disc 1 which was developed primarily by graduate students Whay Lee, Beth Weaver, Prahabkar Koushik, Amjad Daoud and Jim Johnson. These persons have provided me assistance and friendship over the past two years.

I am indebted to Marilyn Norstedt for the quality time and expert advice she has given to the project. There is no substitute to a seasoned librarian when it comes to developing any software which handles the MARC standard for communicating library information. Thanks to Dr. Deborah Hix for her comments in improving the Virginia Disc 2 user interface as well as her willingness to join my panel.

I would like to thank Deveron Milne of VTLS,Inc. for his contribution to the Micro-VTLS database version and for answering my numerous phone calls. Mark Arey and Phil Daniels of AT&T provided extensive support to our premastering efforts using their Meridian Data System. Jeff Doner of Nimbus Information Systems also provided unqualified support for all three discs of the Virginia Disc series.

My personal gratitude goes to Dr. Anton Siochi who has made me a convert of the importance of a well thought out human computer interface and word processing on the Mac II. His editing and advice helped shape this manuscript considerably. Salamat po.

Finally, none of these could be possible without the ceaseless organizational efforts of Dr. Fox. Dr. Fox's prodigious work habits redound down to his student workers. Publishing a CDROM in an academic environment could only have been possible under his leadership. Thank you for the opportunity, the countless hours of excellent editing and the guidance in seeing the Virginia Disc 2 project to completion.

TABLE OF CONTENTS

LIST OF DIAGRAMS

APPENDIX DIAGRAMS

LIST OF TABLES

CHAPTER I
INTRODUCTION

> Without communication there can be no society, and without some form of recorded knowledge and a means for the preservation of that record there can be no enduring culture [Shera, 1980].

The need for information retrieval systems occurs predominantly in libraries and other institutions responsible for the preservation of recorded knowledge such as books, magazines, films, and musical scores. The medium for storing and disseminating recorded knowledge varies with technology. Libraries in the US adopted microforms in the 1960s. In the early 1970s, the application of computers to library systems led to computer-based catalog files and magnetic tapes. In the late 1970s and early 1980s, libraries embraced microcomputers and computer networks. Lately, the advent of high capacity optical devices like the CDROM augurs a new generation of devices for storing and distributing information.

This new generation of optical devices needs innovative applications, better access software, and a new role definition between information publishers, providers and users. Retaining old approaches like transporting a library database from a magnetic memory based media to a CDROM may not be enough. Newer concepts suggested by information retrieval (IR) research should be tested.

This report discusses the creation of a CDROM publication, named Virginia Disc 2 which contains two retrieval systems, Micro-VTLS and Personal Librarian, for comparison. Micro-VTLS belongs to the class of conventional Boolean systems with inverted index files. Personal Librarian uses newer concepts learned from the Syracuse Information Retrieval Experiment (SIRE). The two implementations use standard MARC (MAchine

Readable Cataloging) records, a widely accepted communications format for distributing library catalog information. Both packages hold bibliographic information for the public's access and so the individual PL or Micro-VTLS implementations will sometimes be referred to as CDPACs (CDROM Public Access Catalogs).

The motivation for comparing the two retrieval approaches is to demonstrate the benefits of newer concepts like term weighting, ranking, and similarity measurements employed by Personal Librarian. For years now, commercial mainframe-based retrieval systems have adopted very little of the promising new approaches uncovered by information retrieval research. A CDROM implementation with such innovations could generate interest in adopting new information retrieval approches.

The report answers two questions :

1. How does a SIRE-like system, Personal Librarian, compare with a conventional Boolean based system like Micro-VTLS in a CDPAC application ?

2. How does one create a CDROM Public Access Catalog ?

Discussion begins, in section 1.1, with an introduction to IR research. Further introduction is given in section 1.2, which gives an overview of the Virginia Disc Series of CDROMs, to set the context for describing Virginia Disc 2.

Chapter 2 compares conventional Boolean systems and SIRE-like systems. Because of the paucity of literature on the actual implementation of PL, it will be assumed that PL performs like SIRE, from which it was derived.

Chapter 3 discusses the construction of Virginia Disc 2. Much of the Virginia Disc 2 project involved data preparation or reorganizing the MARC format data to a suitable form for

publishing. Chapter 4 explains the MARC standard because it is critical to understanding what data are in Virginia Disc 2.

Chapters 5 and 6 describe the specific processing procedures, and other facts about Micro-VTLS and Personal Librarian, respectively. Chapter 7 shows the data related to a preliminary comparison of the two systems and the last chapter concludes and summarizes this report.

1.1 INFORMATION RETRIEVAL RESEARCH

The following section introduces some information retrieval (IR) concepts. The aim is to show the disparity between what IR research has accomplished and the state of today's commercial storage and retrieval systems.

1.1.1 Functional View of IR

Writers have different views regarding the scope of IR systems. Maron [1977] states that the purpose of an IR system is to accept subject requests, in some appropriate form, and to search and select all and only the relevant information items from many stored documents. This raises issues about the meaning of relevance and whether 100% relevance is the ultimate objective. Belkin, Oddy and Brooks [1982] expand the task by saying that an IR system should help the searcher choose from an unknown set of documents under a possibly anomalous state of knowledge. Wilson [1978] argues that IR systems provide information rather than answers to questions. The literature on IR clearly agrees on distinguishing between question-answering systems (data retrieval systems) and document retrieval systems [Blair and Maron, 1984] [Minker, 1978]. Data retrieval systems (i.e., database management systems) give exact responses to queries. Document retrieval

systems respond with no guarantee that the retrieved documents will satisfy the user's requirements. This report shall deal exclusively with document retrieval systems.

Salton and McGill's [1983] functional diagram of information storage and retrieval (ISR) systems, shown in Figure 1, depicts a system with two phases: indexing and retrieval. Indexing is the phase of analyzing the incoming documents and expressing the information content in the language of the indexing system. The retrieval phase matches a user's query with the indexed documents to yield the relevant documents. The figure also incorporates Van Rijsbergen's [1979] functional model of retrieval which includes a feedback mechanism.

Some interesting research questions appear on the diagram to indicate the state of flux in IR research. The functional model shown in Figure 1 can account for the variety of approaches researchers have proposed through the years. What it cannot account for is the diversity of views on what should be considered a relevant document. A taxonomy of research approaches in IR can be inferred from the types of answers researchers have given to this very important question.

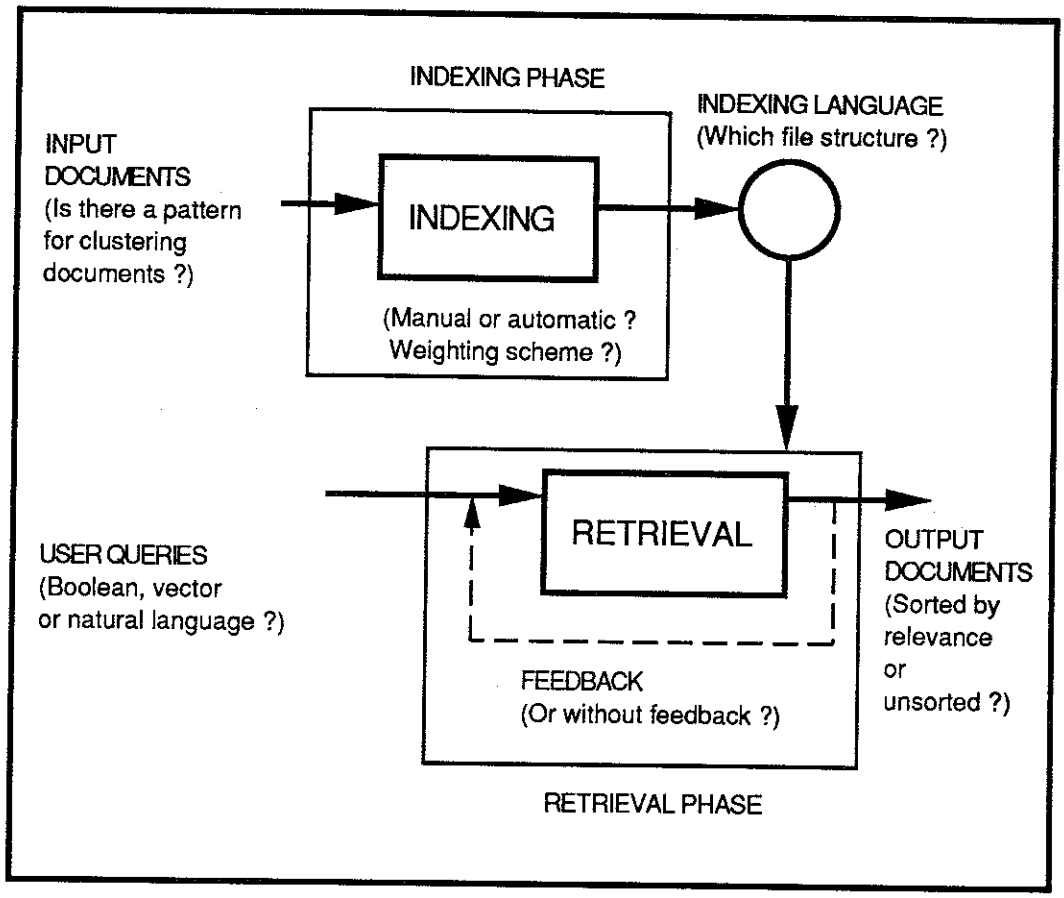


Figure 1. A Functional View of Information Storage and Retrieval. Some interesting research questions are suggested in parentheses.

1.1.2 Information Retrieval Approaches

Maron [1977] asserts that so long as a document is "about" the concept described by the query, the document is deemed relevant. A ubiquitous implementation of Maron's concept involves the conventional Boolean query. Boolean queries are keywords or terms connected together by the Boolean operators AND, OR and NOT. A typical search query would look like :

public AND (catalog OR access)

The above query would divide the indexed documents into two groups: those that satisfy the query and those that do not.

The concept of "aboutness" deals with grades of relevance rather than absolute relevance. This fuzzy relevance notion leads to the application of fuzzy-set theory to retrieval [Bookstein, 1985]. Waller and Kraft [1979] apply the concept of weighting, a method of emphasizing some of the user's query terms over the rest of the query terms. The fuzzy-set method rates the indexed documents beyond just deciding relevance or non-relevance. Applying fuzzy-set logic permits one to formulate the traditional Boolean query while allowing more expressiveness because some keywords can be stressed by weighting. Obtaining output documents, ranked according to perceived relevance, is another clear benefit. The user saves time by searching first for the items that are most likely to be relevant. Bookstein [1985] cautions against some weaknesses of fuzzy-set retrieval including its inheritance of the doubts cast upon the Boolean framework.

Another notion of relevance involves a probabilistic view. Relevance is associated with the probability of being relevant given the word occurrences in the document collection and the

word list of the query [Bookstein, 1985]. An approximation to the probability of relevance can be computed from the word distribution of the document database and the terms present in the query [Bookstein, 1985]. Statistics about word distribution are gathered during the indexing phase. The probabilistic model is usually implemented as a decision theoretic process. Costs are assigned to the retrieval of a non-relevant record and the non-retrieval of a relevant record.The decision to choose can thus follow a loss minimization rule.

The vector space model measures relevance by a distance measure. If documents were broken down to their component terms and weights assigned to rate the relative importance of each term, we can build a matrix representation of the documents, as seen in Figure 2. The d_{ij} take on values in the range [0, 1] as a measure for the importance of term T_j in representing document D_i .

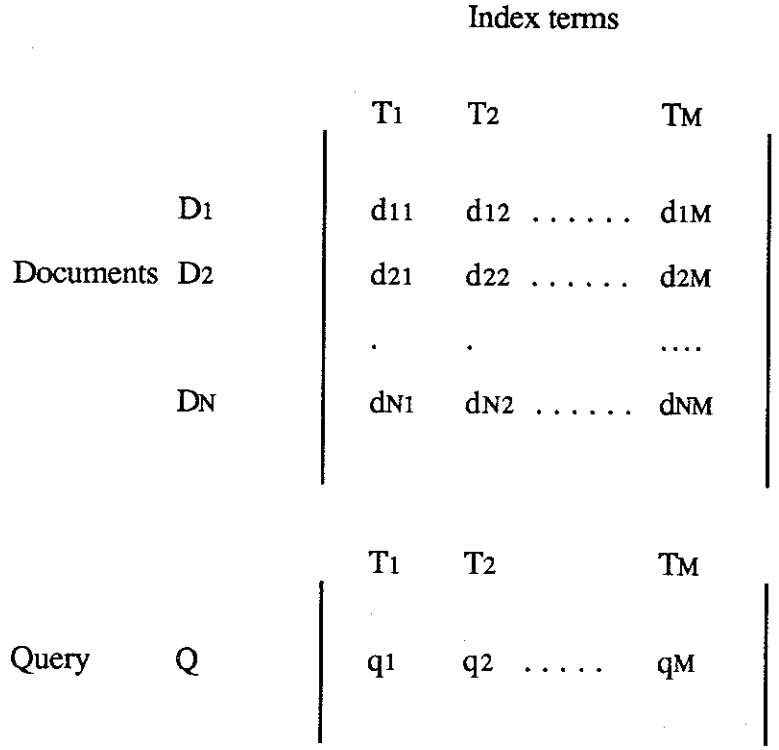


Figure 2. A Vector Space Model of N Documents and a Query Q.

Documents and queries in a vector space model represent points in an M dimensional Euclidian space. Relevance can be thought of as inversely proportional to the distance measurement between the query and a document. The distance from query Q to document D_i approximates the relevance of query Q to document D_i. There are various accepted relevance metrics like the cosine measure. Intuitively, the cosine measure (shown in Figure 3) approximates the angular distance between a query and a document in M dimensional space.

$$\text{COSINE } (D_i, Q) = \frac{\sum_{j=1}^{M} (d_{ij} q_j)}{\sqrt{\sum_{j=1}^{M} d_{ij}^2 \sum_{j=1}^{M} q_j^2}}$$

Figure 3. The Cosine Similarity Equation

The problem with the probabilistic and vector space approaches is that they generally do not incorporate Boolean operators. The P-norm extension solves this problem by providing a facility for softening Boolean operators [Salton,Fox,Wu, 1983], interpreting ANDs and ORs as somewhere between strict two valued operators and non-Boolean term independent operators.

Relevant documents can also be located using the clustering approach; similar documents can be stored and later retrieved together. Clustering in IR is a means of extending the range of query-to-document matches to help formalize document to document relationships. Clustering is used in tandem with other retrieval approaches and tries to take advantage of the common situation that similar documents are needed together.

Other methods determine relevance by the user's direct feedback. Browsing and hypertext systems present the user with fast and powerful front-end manipulation tools to navigate through documents and let them decide upon document relevance during a sequential interactive process.

1.1.3 The State of Current Retrieval Systems

Despite the wealth of ideas gained through IR research, few commercial systems incorporate the new IR approaches into their own products. Some have suggested that search service vendors have had economic motives for sustaining the gap in theory and practice [Hildreth, 1985] [Radecki, 1988].

A study by Smit and Kochen [1988] related the results of interviewing 37 database vendors about innovations in their retrieval systems. Smit and Kochen conclude that :

- Developers lack knowledge about potential innovations in information retrieval systems.
- The database vendors would innovate if customers started switching to competitors or started using alternative delivery systems like CDROMs.

The following industry trends should motivate innovations in commercial IR systems.

1. Hildreth [1985] reports the enthusiastic acceptance of online public access catalogs (OPACs). OPACs (e.g. VTLS) are built to replace the card catalog and surveys show that users overwhelmingly (> 70 percent) prefer OPACs to the card catalog. OPACs allow greater direct online access for the public.

2. The substantial rise in CDROM publications threatens to take over market share from entrenched online retrieval services.

> "The advent of the CDROM and its interconnecting to the personal computer has raised questions as to whether searching will move away from timeshared services such as DIALOG in favor of local searching."
>
> - R. Summit, president of DIALOG Information Services [Summit, 1987]

3. Librarians themselves want better control over their retrieval software and are demanding more innovative retrieval systems [Janke, 1987].

End-user searching of online databases and CDROM publications will be commonplace in many libraries and universities. Public access catalogs on CDROMs called CDPACs will also be commonplace. This shall undoubtedly create a clamor among users for better retrieval software.

1.2 VIRGINIA DISC CDROM SERIES

Several attenders of the 1987 Microsoft CDROM conference met to discuss the benefits of locally (in Virginia) producing a CDROM product. The project would stimulate interest in this innovative medium as well as gain valuable experience for the participants. The CDROM product itself would directly benefit those institutions on the distribution list. For instance, local libraries would have CDROM discs with original Virginia catalog data which they could search. They could determine if CDROMs fit their libraries' needs.

Several industrial concerns stood to gain from a Virginia Disc publication. Nimbus Records, a producer of quality audio compact discs in Green County, Virginia, would gain a broader market by expanding their services into CDROM publication through its division, Nimbus Information Systems. An audio disc manufacturer could add the capability of producing CDROM data discs with only an incremental investment in equipment. Nimbus Records became the primary industrial supporter for the Virginia Disc series. Another

beneficiary, VTLS, Inc., a library automation company based in Blacksburg, Virginia, markets a microcomputer based library automation system, Micro-VTLS, which could be tested on a CDROM. By transferring a hard disc based system to a CDROM, new product concepts could be tested for performance and overall viability. Several other organizations, namely Personal Library Software Inc., Newman Library, and the Virginia State Library, allowed use of their software or data for test purposes.

The Virginia Disc Series includes at least three discs. Virginia Disc 1 holds a large number of software packages and databases from VPI&SU and a variety of other sources. Virginia Disc 3 contains data meant for the MacIntosh running A/UX while Virginia Disc 2 stores library bibliographic and holdings information from the Virginia State Library and the Virginia Tech University Library.

Two libraries, the Newman Library and the Virginia State Library, contributed a portion of their catalog files for the Virginia Disc 2 CDROM. The Newman Library provided 10,084 bibliographic records and their accompanying holdings records. The Virginia State Library and Archives contributed 5,598 bibliographic records classified under the topic of "Virginia History."

The Virginia Disc 2 project has the goals of :

- Making MARC serial records available on a CDROM so that duplicate conversion activities can be avoided among state-supported libraries.
- Helping increase understanding of how data from varying formats can best be stored and accessed on a CDROM.
- Providing insights regarding factors involved in local publishing of CDROMs.
- Documenting different types of search and indexing software as they relate to types of data and library settings.

The first goal raises some organizational, legal, and technical questions. Statewide cooperation among libraries with the idea of creating a union of their catalogs on a CDROM has precedents, e.g. statewide projects in Missouri and Illinois [Davis, Foster, Raithel, 1989]. Such projects involve considerable expense and organization. Organization is the first impediment. The quality of MARC records vary from one source to the next and coordination is needed to assure the quality of original cataloging among the participating state libraries. The second impediment is legal. Transferring MARC records purchased from a bibliographic utility to another library can incur legal problems. VTLS, Inc. encountered a similar obstacle in testing the company's networking software called VANILLA [Arneson, 1989]. Only original cataloging (i.e. cataloging produced by the participating libraries) can be transferred without any legal restraint. Libraries which do retrospective conversion, conversion of old books to the MARC standard, could contribute to a shared cataloging endeavor. The last obstacle is technical. The participating libraries contributing MARC records must coordinate the quality of their shared records.

The latter three goals were reasonably well achieved. The first goal was not achieved but was investigated for viability. The intent to make MARC records available on a CDROM was deemed a useful product by the Virginia Disc 2 developers but not as important as the latter three goals. The second and third goal mandated the acquisition of practical experience from the project. The lessons learned from the CDROM project appear in chapter 8. The fourth goal of Virginia Disc 2 coincides with our desire to compare Micro-VTLS and Personal Librarian, which is discussed in more detail in the next chapter.

CHAPTER II
COMPARISON OF BOOLEAN SYSTEMS AND SIRE

This chapter deals with a comparison of specifications for the two classes of retrieval systems in order to understand better their differences. Micro-VTLS belongs to the conventional Boolean class while Personal Librarian is related to the SIRE-type class. The SIRE system has a file scheme not unlike the inverted files of Boolean systems but its capabilities differ substantially (as shown in Table 1). Personal Librarian is described as an advanced version of SIRE [Fox and Koll, 1988] and it has promising features for accessing large bibliographic collections.

Two ratios, precision and recall, are accepted measures of a retrieval system's performance. Precision measures the ratio of relevant documents to the number of retrieved documents. Recall is the ratio of relevant documents retrieved to the total available relevant documents in a collection. The two ratios as realized by current systems exhibit an inverse relationship and the improvement of one usually leads to deterioration of the other. Good retrieval systems have high (close to 1.0) values of precision and recall.

2.1 CONVENTIONAL BOOLEAN SYSTEMS

2.1.1 Indexing

Conventional Boolean systems use Boolean queries to specify desired documents and inverted files to enable rapid query processing. The items stored in the inverted files may be terms from document titles, abstracts, or full-text (in the last case, leading to the description "full text retrieval system"). These can be automatically determined from the document text

[Salton and McGill, 1983]. Other items, often called keywords or descriptors or index terms, may be placed in inverted files as a result of manual indexing.

Conventional Boolean systems, like Micro-VTLS, often use manual indexing as opposed to automatic indexing. The subject classifications are prepared by expert classifiers in the Library of Congress (LC). Many users of public access catalogs do not know this (only 28% know this, based on [Steinberg and Metz, 1984]). A user's query might not match the LC's controlled vocabulary and this could result in very few retrievals. Of course using the LC subject headings for subject indexing allows a minor library to tap the immense cataloguing talents of an institution like the Library of Congress or of a major bibliographic utility. The availability of MARC records and the MARC standard greatly simplifies the task of indexing conventional Boolean systems. It perhaps is a factor in the longevity of conventional Boolean systems.

Automatic indexing relies on the premise that words in a document can be used to characterize the meaning of documents. The words or terms to represent a document must be carefully chosen by the indexing algorithm. Documents should be indexed only by those words that discriminate well between the meanings of documents. Salton and McGill [1983] describe ways of assigning weights, thus importance, to terms. Some metrics, like the inverse document frequency, can measure the importance of a word relative to other words in that document. Weights are an important aspect of advanced retrieval systems that rely on automatic indexing.

A study [Blair and Maron, 1985] criticizes automatic indexing systems for recognizing too few relevant items. Blair and Maron recovered as little as 20 % of the total relevant items in a collection and recommend using manual indexing instead. Salton [1986] defends automatic indexing.

2.1.2 Query Formulation

Exactly what are the limitations of a Boolean system ? While other retrieval approaches like p-norm or systems like SIRE also accept Boolean query expressions, conventional Boolean systems like Micro-VTLS only interpret the operators AND, OR and NOT strictly.

Bookstein [1985] lists several criticisms of Boolean systems :

- Boolean logic can perplex users who poorly understand the strict interpretation of a Boolean expression. For instance, a request like A or B or C or D responds the same way when A exists as when both A and B exist.
- Users desire some method of organizing the results according to importance (ranking).
- The users cannot stress certain key terms over others. In the same vein, index terms cannot be weighted over other terms when representing a document during indexing.
- There is no systematic way of modifying a request in response to the quality of the retrieved items (i.e. a feedback mechanism).
- The difficulty of formulating Boolean queries can produce awkward and lengthy queries.

2.2 SIRE-LIKE SYSTEMS

Personal Librarian (PL), like the SIRE system, belong to the vector space model approach of ISR. PL, being a commercial product, is not discussed in detail in the technical literature; thus this discussion will describe the SIRE system instead.

2.2.1 SIRE Indexing

The SIRE system has the normal inverted file processing facilities as in Boolean systems but SIRE also incorporates term weighting during the indexing phase

[Noreault,Koll,McGill, 1977]. SIRE does automatic indexing rather than manual indexing. Some of the added indexing features of SIRE include term weighting and stemming. Term weighting improves retrieval by giving emphasis to words which better distinguish between documents. Some words are more effective than others in resolving the ambiguity of a document's contents; they are identified through statistical analysis. Root matching or stemming serves to conflate the morphological equivalents or duplicates of terms.

2.2.2 SIRE Retrieval

SIRE's retrieval phase consists of two stages. First, a Boolean query is processed in the normal manner. With the documents that qualify, SIRE applies the COSINE (refer back to Figure 3) or other similarity computation to rank the relevant documents.

Ranking or the ordering of the output documents according to presumed relevance allows the user to look at the most important documents first and decide when to stop looking down the list. This is an obvious advantage when one encounters a list too long to browse. Noreault, Koll and McGill [1977] claim they are able to move the relevant documents 35% of the way from random scattering to the top of a relevant list.

2.3 FEATURES OF PERSONAL LIBRARIAN

Personal Librarian contains a lot of features as seen in Table 1. Personal Librarian is shown to have a number of features not present in Micro-VTLS.

Table 1. A Comparison of Micro-VTLS and PL Features

FEATURE	Micro-VTLS	Personal Librarian
1. Field search	Can search any of 12 access keys	The user has access to all the fields
2. Field range search	Cannot handle numeric fields	Yes
3. Boolean operators	Yes	Yes
4. Scan more than one field in search	Yes	Yes
5. Wildcard search	No	Yes
6. Help screens	Yes	Yes
7. Adjacency or proximity search	No	Yes
8. Ranked output	No	Yes
9. Natural language queries	No	Implied OR between terms
10. Using a document to find similar documents	No	Yes
11. Using an old query as a keyword	No	Yes

CHAPTER III
CDROM PUBLICATION

Chapter 3 introduces the CDROM publishing process and then discusses the Virginia Disc 2 publication process.

3.1 PUBLICATIONS ON CDROM

The number of CDROM publications has risen dramatically in the past three years. Over 300 products were available as of 1988 in CDROM format . Many of these products, like BiblioFile and WILSONDISC, rely on reformatting MARC standard bibliographic records and presenting the data with a proprietary or commercial storage and retrieval package. The dominance of the MARC bibliographic standard in online cataloging services and tape distribution presents the challenge of preparing, presenting, indexing and retrieving MARC records in a manner suitable for CDROM publishing. The Virginia Disc 2 project discussed in this report faced such a challenge.

BiblioFile, the acknowledged first CDROM publication, predated the CDROM standards which energized the industry. Martin's [1986] account of BiblioFile's development cycle describes the time before the arrival of standards and a variety of support products for CDROM publishing. The High Sierra Group Proposal [Einberger,1986], now replaced by ISO 9660, a common volume and file format for CDROMs, vitalized the industry. CDROM publishers then had a logical format on which to base their applications; a format independent of the CDROM drive manufacturer. The ISO 9660 standard was used in preparing Virginia Disc 2.

The seminal 1986 publication by Microsoft Press, CD ROM The New Papyrus, popularized the virtues of the new technology : a thousand fold increase in storage capacity, cheap CDROM players, cheap discs (1¢ per Megabyte), new multimedia possibilities. Several hundred more articles about CDROMs published over the succeeding two years are listed in [Barnes, A. CDROM Compact Disc -- Read only memory. 371 Selected Citations. July,1988], and attest to the popularity of this new publishing medium.

3.2 STANDARDS

Optical publishers should be cognizant of the standards prevailing in the industry and the emerging standards for optical technology. The physical dimensions are governed by the Philips/Sony "Red Book" which specifies the audio recordings on compact discs and the "Yellow Book" which discusses data storage. Because of this, almost any CDROM can work on any player. ISO 9660 defines the logical format, a volume and file format standard for read-only optical media [Lynch, 1987].

A proposed NISO standard for a common command language for search systems has been circulating since 1986. The Z39.58 proposed standard declares the vocabulary, syntax, and operational meaning of commands for use in an online interactive system [Crawford, 1989]. The adoption of such an interface standard would provide motivation for altering the command interface of Micro-VTLS, PL, or any other bibliographic publication on CDROM. Even if the standard is never approved, it behooves people interested in CDROM bibliographic publishing to use the proposal at least as a comparative basis of design. Library patrons and librarians have expressed the difficulty of learning a new command interface for each new CDROM bibliographic publication.

3.3 CDROM MANUFACTURING PROCESS

The Virginia Disc 2 publication process consists of the data preparation phase and the production phase (see Figure 4). The data preparation phase encompasses the file filtering, indexing and integration work done on an IBM PS/2 Model 80 while production concerns the premastering, mastering and replication steps done using special equipment from Meridian Data, Inc. The data preparation phase usually takes the longest time. Once finalized however, the production phase proceeds relatively mechanically.

After data preparation, the files (data and programs) are transported via nine track tapes to the pre-mastering facility. Submission standards tend to vary among premastering facilities and the disc manufacturer needs to test if the tapes can be loaded properly. Premastering involves adding error detection and correcting bytes to the data. Premastering facilities usually employ a system where the performance of the CDROM can be simulated. This allows the developer to assess the applications' response time without actually pressing a disc.

Mastering involves the creation of a physical master disc. A glass CDROM master is produced by the action of a laser burning pits into a photoresist surface. Later, this glass master is used to produce copies of the disc.

Replication is done by obtaining a metal stamper from the glass master. Using the metal stamper as a mold, polycarbonate resin is poured over the stamper and the plastic is pressed with the stamper's indentations. A reflective coating of aluminum is then applied to the pits on the plastic disc. Finally, a protective layer is applied over the reflective layer and the replicated CDROM is labelled and delivered [Armstrong, 1986].

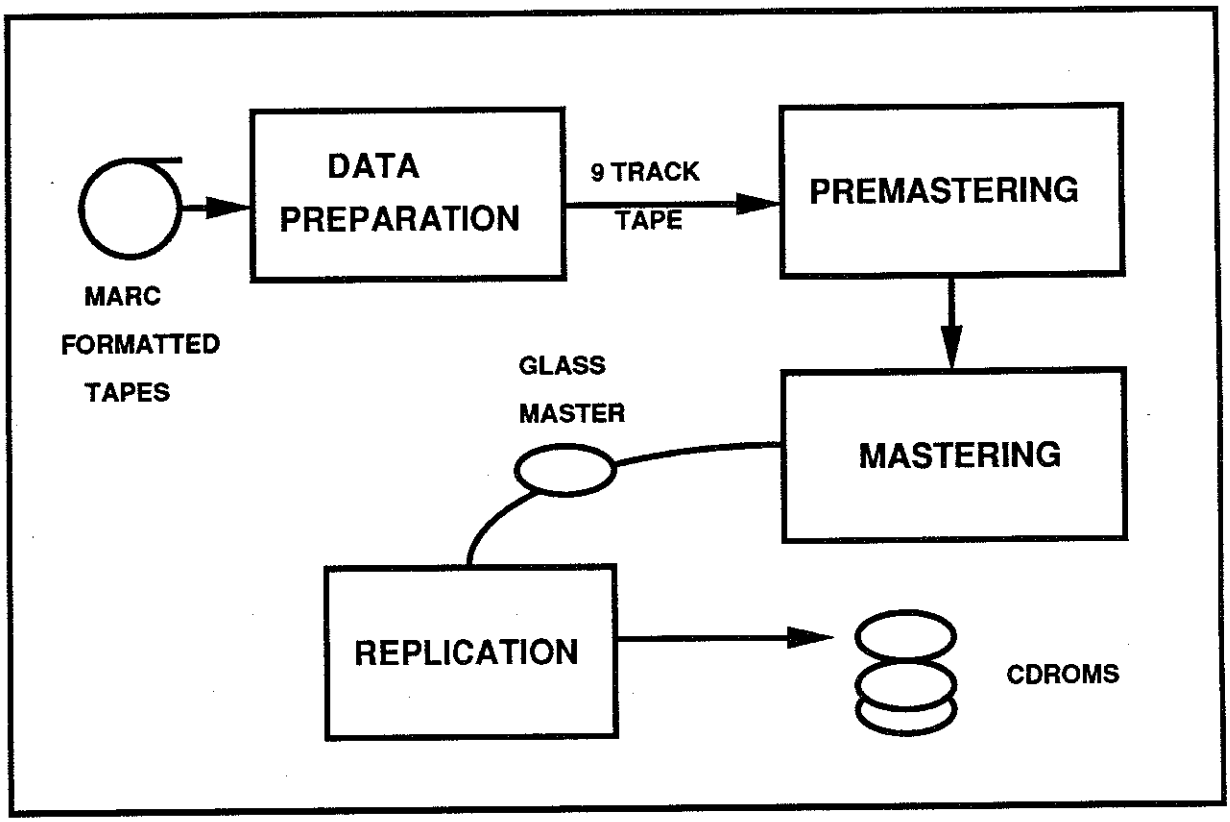


Figure 4. The CDROM Production Process

3.4 VIRGINIA DISC 2 PUBLICATION

Appendix B lists the participants in the Virginia Disc 2 project and Figure 5 portrays the publication process.

The project began with an offloading of MARC data onto magnetic tapes by the computer departments of the Newman Library and the Virginia State Library. The MARC format is a widely accepted standard so the donor libraries already had the programs to produce and accept MARC records. Newman Library provided six 9-track tapes comprised of 3 tapes with bibliographic information and 3 tapes with MARC holdings information (see Table 2). The Newman Library contributed 10,084 bibliographic records with the accompanying holdings records from its serial collection. The Virginia State Library gave one tape containing 5,598 MARC records of documents about Virginia's history.

The tapes were processed in parallel by VTLS, Inc. and VPI&SU. Of the total of four databases created, VTLS,Inc. prepared two databases with Micro-VTLS while VPI&SU built the two Personal Librarian databases. More processing details are discussed under the chapters about Personal Librarian and Micro-VTLS.

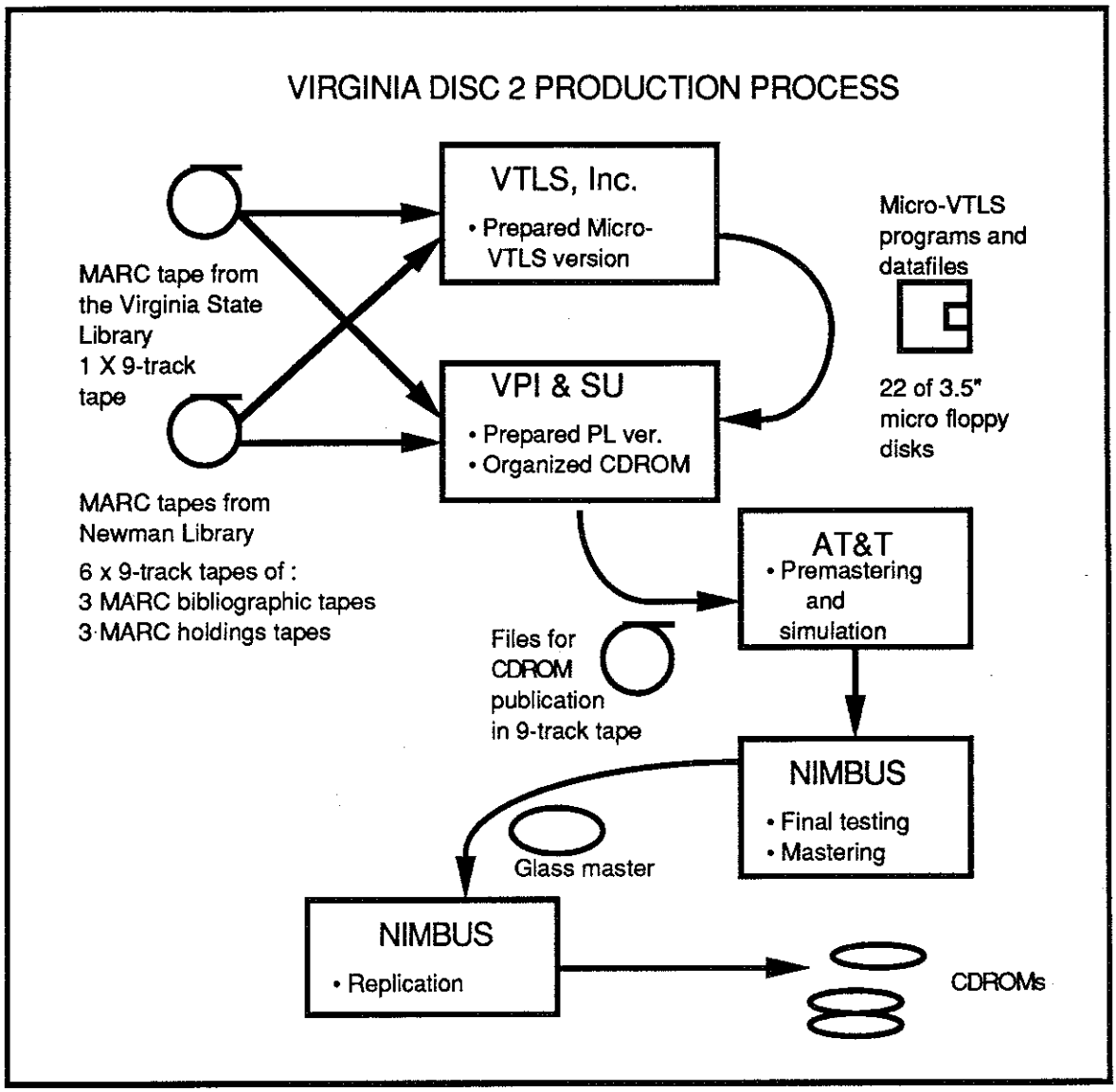


Figure 5. Virginia Disc 2 Production Process

Table 2. Attributes of Input Tapes to Virginia Disc 2

	Newman Library		Virginia State Library	
Contents	Serials collection of Newman Library		Documents about Virginia's History	
Record types	Bibliographic	Holdings	Bibliographic	Holdings
Number of tapes	3	3	1	none
No. of MARC records	10,084	10,084	5,598	none
File sizes	21.63 Mb	22.05 Mb	11.56 Mb	none

Of concern among the participants was the timeliness of the holdings data, which because of delays, were rendered inaccurate. Still, the data served their initial purpose for comparing the capabilities of the two retrieval systems. Originally, the plan was to process the four databases or OPACs in parallel between VTLS,Inc. and VPI&SU. Data preparation was to be complete by September, 1988. A loss of data due to misloading the MARC tapes in July, 1988 and the slipped schedule of Virginia Disc 1 led to the delay of VPI&SU's Personal Librarian version. VTLS,Inc., in the meantime, had completed an early Micro-VTLS version by September, 1988 while the VPI&SU team was still realizing that up to 35% of its Newman Library and Virginia State Library had been truncated. VTLS, Inc. went ahead and mailed the Micro-VTLS OPAC of the Newman library data to Nimbus Information Systems so their large discs could be used for indexing. By November, 1988, the Virginia Disc 2 project languished due to the unavailability of accurate data for the VPI&SU team and the temporary halt by VTLS, Inc. pending further developments. Suggestions were made at VPI&SU to rid the data sets of truncated records and salvage what was left. An almost complete set of data was eventually obtained.

On December, 1988, copies of the MARC data tapes from VTLS, Inc. were made available to the VPI&SU team. By December, 1988, only one of the four databases was complete and the project was late by three months. The data preparation process is discussed in more detail in Chapters 4 and 5.

The VPI&SU team worked under the disadvantage of not being well versed with the MARC standard nor with the proposed MARC holdings format. The learning curve for studying MARC records caused delays. Books are written about MARC but MARC is an evolving standard with many variants attributable to the type of material represented, the bibliographic utility source, and the country standard used.

The first half of 1989 was productive for VPI&SU even with a two month long hardware problem in May-June, 1989. In July,1989 the VPI&SU team waited for VTLS, Inc. Newer versions of Micro-VTLS necessitated a reindexing of the Newman library data and indexing the still undone Virginia State Library data. VTLS, Inc. transported the new Micro-VTLS version in 3.5" micro floppy disks to VPI&SU. VPI&SU prepared the software to integrate all the four databases in a cohesive CDROM publication. The menu-based interface developed at VPI&SU provides facilities to install/deinstall the software, access any of the four databases, and display online help. By the middle of August, all four databases were ready.

A recurrent problem arose in the mode of communication files between VTLS, Inc., VPI&SU, and the production facilities at AT&T and Nimbus Information Systems. File copies on 9-track tapes sometimes were unreadable at the destination facility. The development system at VPI&SU lacked a directly connected 9-track tape drive. For groups who wish to coordinate during the data preparation phase of a CDROM publication, the usefulness of a common device and format for transporting data (e.g. 1.44Mb micro-floppy, streaming tape cartridge, ethernet lines) cannot be overstated.

The help of several persons were solicitied to improve the user interface. Ms. Norstedt, head of cataloging for the Newman Library, helped interpret the MARC codes properly and gave ideas about OPACs and the patrons who use them. Dr Fox, of the computer science department provided suggestions for improving the usability of the interface and helped edit the screen displays. Dr. Hix, also from the computer science department, suggested ways to make the installation process more tolerable to the user and improved the help screens. Over thirty minor alterations were done to improve the usability of Virginia Disc 2. In some

cases, the irritants were inherent in the packages used (e.g. Personal Librarian's command screen). No alterations could be made to any of the commercial packages used.

The files were uploaded from a PS/2 to a VAX and written to 9-track tape for transport to the AT&T Document Development Organization facility in Winston-Salem, NC, which has a Meridian Data, Inc. premastering facility that can emulate a CDROM's performance. Premastering added in error correction codes and built the volume table of contents, the directory, and the files according to the ISO 9660 standard. Tapes with this data were sent to the mastering facility, Nimbus Information Systems, for replication.

Table 3 shows the attributes of Virginia Disc 2.

Around a hundred holdings and bibliographic records were lost because of the proprietary nature of the HP tape format in a problem that is discussed in section 6.2. Specifically, the Personal Librarian version of the Newman library data suffered from loss of data and this limits the comparability of the Micro-VTLS implementation with the PL version. The small fraction of loss of around 0.42 % should still allow a fairly accurate direct comparison of the two versions, however.

The whole publication process began with seven tapes containing MARC records. The wide availability of MARC records makes the Virginia Disc 2 publication process repeatable. A full discussion therefore of the MARC standards follows in the next chapter.

Table 3. Attributes of Virginia Disc 2

Data Source/Attribute	Virginia State Library	Newman Library
Size of MARC file input	Bibliographic =11.56 MB Holdings = none	Bibliographic = 21.63 Mb Holdings = 22.05 Mb
Number of records	5,598	10,084
Number of bad records	none	Bibliographic = 43 Holdings = 66
Number of unique words	15,233	33,224
Number of words read	276,898	493,394

Versions/Sizes	Virginia State Library Database Versions		Newman Library Database Versions	
	Micro-VTLS	PL	Micro-VTLS	PL
Total file sizes[1]	21.78 Mb	15.94 Mb	58.29 Mb	39.99 Mb
Database source file size[2]	14.98 Mb	2.87 Mb	44.58 Mb	5.63 Mb
Index file sizes[3]	6.36 Mb	2.26 Mb	13.40 Mb	4.20 Mb
Relative size of index[4]	0.42	0.79	0.30	0.75

Notes :

1. Measures the databases, index files, screen data files and all other support files

*2. Measures the files used as input for indexing: *.DBF files for Micro-VTLS and *.SRC files for PL*

*3. Measures the files created during the indexing phase: *.NTX files for Micro-VTLS and *.INV, *.DIC, ... etc. for PL*

4. Measures the ratio of index file sizes to database source file size

CHAPTER IV
THE MARC STANDARD

Library automation on a national scale has been viable only since the late 1960s, in part because of widespread acceptance of the MARC standard. MARC stands for MAchine Readable Cataloging, a communications format for representing library information in machine readable form. It was originally conceived as a format for distributing catalog information in magnetic tape form from the Library of Congress to its subscriber libraries. MARC has since grown in scope and acceptance and has profoundly affected the library automation industry by spawning institutions whose primary function is to create and distribute MARC records (i.e., bibliographic utilities).

4.1 MOTIVATION FOR STUDYING THE MARC STANDARD

Understanding of the standard by which librarians, bibliographic utilities and library automation companies communicate would help one appreciate the wide range of computerization possible in libraries. For instance, CDROM publishing of bibliographic data gives us the ability to republish the enormous amount of catalog information that is already available in machine readable form. By understanding the MARC standard, automation experts can design retrieval systems that take advantage of the consistency and conciseness of the standard.

Besides the volume of MARC records in circulation, the quality of captured information makes it an ideal data source for study. The production of MARC records is stringently checked before distribution by the bibliographic utilities. Moreover, the rules for cataloging

are minutely codified in a 600 page manual, the AACR2, the Anglo-American Cataloging Rules, ed 2.[ALA, 1978].

For Virginia Disc 2, the MARC format is important because we use the catalog record as a surrogate for the original document. For example, in the Personal Librarian implementation, automatic indexing operates on the MARC record, not on the full-text of the document that the MARC record represents.

Questions relevant to the project are :

• Are the MARC fields rich enough in subject words to discriminate one document from the rest ?

• Can MARC records from different sources easily combine to create a union of catalogs for a CDROM ?

• What are the types of MARC records and how can each MARC record type be used ?

• Which of the MARC fields are suitable for display and which fields are suitable for indexing ?

4.2 HISTORY OF THE MARC STANDARD

The Library of Congress (LC) in the early 1960's was the single largest creator of cataloging information in the US. At a propitious time when computers were making inroads into libraries, it was recognized that adopting a common standard for communicating cataloging information for books would eliminate a vast amount of transcription and re-cataloging.

From the start, the Library of Congress undertook a system of continuous consultation. The American Library Association (ALA), through its committee on Machine-Readable Form of Bibliographic Information (MARBI) acts as advisor to the LC on matters of changes to the widely accepted MARC standard. The Library of Congress regulates the standard today via the MARC Standards Office. The standard continually evolves. Representatives from MARBI, liaisons from the major bibliographic services, and representatives of other national libraries gather every 3 months to consider new proposals to revise or extend the standard.

4.3 BIBLIOGRAPHIC UTILITIES

The dominance of four main bibliographic utilities (OCLC, RLIN, WLN, UTLAS) today also influences the standard. Bibliographic utilities distribute and sell catalog information to libraries. The individual libraries can themselves sell original cataloging to these utilities as well.

The four top bibliographic utilities today all originated from projects undertaken in academic institutions during the library automation boom of the early 1970's. The largest, the Online Computer Library Center (OCLC), came from the Ohio College Library Center in 1967. OCLC provides over 3000 libraries with archive tapes and online cataloging.-- all of these for MARC-based systems. OCLC provides services through networks or agencies which link groups of libraries together. OCLC serves the Newman Library through such a facility and much of the serials data used in this CDROM project originated indirectly from OCLC. OCLC employs slight extensions to the MARC standard in order to allow online applications. The practice is not unknown to the industry and the OCLC records are known as OCLC MARC records. Likewise the Research Libraries Information Network (RLIN),

another major bibliographic utility, has RLIN MARC. Both are supersets of the standard MARC. OCLC MARC carries more information than the MARC standard. Both, on the basis of processing programs, are called "MARC-compatible". Besides OCLC and RLIN, the other major bibliographic utilities are the Washington Library Network (WLN) from Washington State University and the University of Toronto Library Automation Systems (UTLAS).

4.4 TYPES OF MARC STANDARDS

MARC refers to a generic term applicable to a host of national standards like UKMARC (Britain), CANMARC (Canada), USMARC and so on. In this document, MARC refers to USMARC or LC MARC, names interchangeable by virtue of the Library of Congress's active role in maintaining the standard. The outlook for the standardization of an international MARC format also appears promising. National agencies want to reach an international agreement for a universal MARC standard by 1993 [Norstedt, 1989].

4.4.1 Types of Material Represented

The success of MARC II, a standard for books, led to the development of standards for other types of library materials. A list of the different materials standards and the dates the standards became active follows :

1. Books (1968)
2. Serials (1970)

> "Printed or microform language material issued in successive parts bearing numerical or chronological designation and intended to be continued indefinitely" [ALA, 1978]

3. Visual Materials/Films (1971)
4. Archives and Manuscript Control (1973)

5. Maps (1970)
6. Music (1973)
7. Machine-Readable Data Files (1981)

Although the other MARC bibliographic formats are also important, the serial standard is analyzed in depth here because it is directly relevant to this CDROM project. The serial format includes periodicals, newspapers, annuals, journals, memoirs, proceedings, transactions of societies, and numbered monographic series. Data contained in Virginia Disc 2 include serial information from the Newman Library.

4.4.2 Holdings and Authority Information

Libraries normally maintain three categories of information. The seven bibliographic standards cited previously belong to category one. Bibliographic information contains information normally stored in a card catalog. Two more categories, holdings and authority information, evolved from the need to represent other information which libraries stored in non-standard ways.

4.4.2.1 Authorities Format

The MARC Authorities record (herein called authorities) provides information about names, subjects and/or series authority information [Library of Congress, 1976]. Some names or terms often recur in defining bibliographic records; in order to maintain consistency in the use, spelling and/or meaning of these words, a separate file about them has to be maintained. Previously, libraries maintained lists to assure the consistent entry of names and to avoid making repetitive decisions on the use of the names. By retaining a computer file on vital words or phrases, authority information could be used to :

- Provide references to other related headings,

- Maintain quality control in the bibliographic records,
- Automate changes to established headings, and
- Allow alternative spellings to the same word or name.

The use of authorities leads to a more consistent use of words, terms and spellings.

4.4.2.2 Holdings Format

The holdings format evolved out of the need to maintain serial bibliographic records. Holdings data describe the piece-wise collection of a library [Library of Congress, 1984]. Given a magazine for instance, the holdings record should enable a patron to determine the magazine copies that are accessible. The holdings records help solve the non-trivial problem of describing the complete item level holdings of a library in a concise manner.

Serial holdings data are extremely volatile because :

- New issues arrive periodically,
- The pieces of material could be removed to be bound or replaced by microfilm or another medium,
- Periodicals could be erratic in frequency and in consistency of issue.

The USMARC holdings format was first proposed in 1982. A group which included the Virginia Tech Newman Library developed a format for holdings and locations data and tested the same on their respective collections. As a coding scheme, the MARC holdings format tackles a difficult problem and comes through with a well thought out design [Crawford, 1984]. The coding necessitates voluminous programming in order to manipulate records. Over a third of the author's programming time was spent in reformatting the holdings codes into a form suitable to display on the screen.

Holdings information may well be inappropriate for CDROMs because of its volatility. The very same reason for isolating holdings information from bibliographic information (rapid changes), generally makes the information unsuitable for publishing on a CDROM. Relatively static data, like catalog information, which can be useful even with infrequent updates (i.e. updates every three months) are suitable for CDROM publication.

4.5 MARC RECORD PROCESSING

This section describes the organization of a MARC record and shows the processing necessary to extract the information needed for Virginia Disc 2 (see Figure 6).

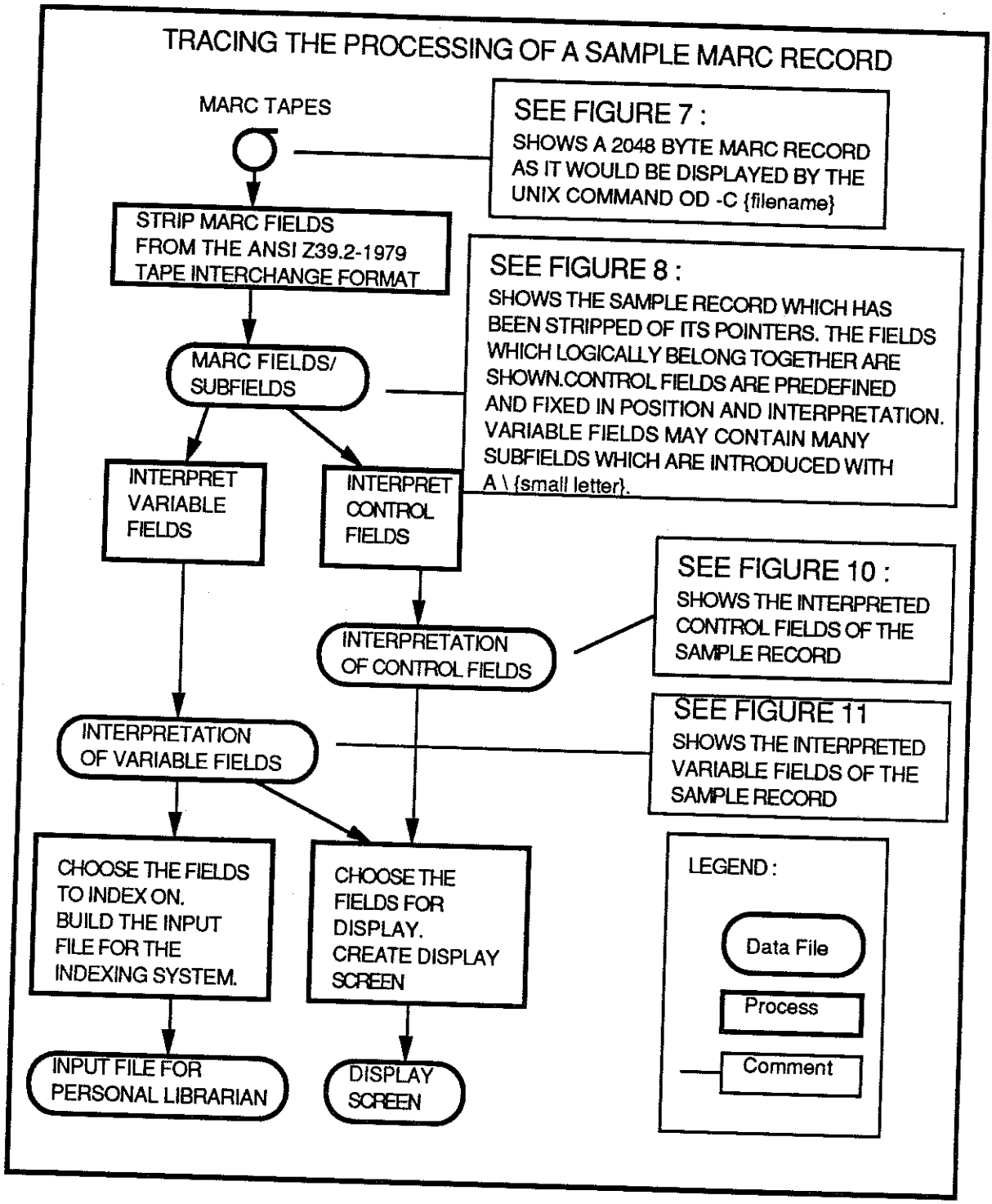


Figure 6. Trace of the Processing of a MARC Record.

Numerous books have been written about the structure of the MARC standard through its two decade history. The main bibliographic utilities distribute their own extended versions of the standard. The Library of Congress distributes the official version in its latest form. Serious software developers should consult the USMARC publications from the Library of Congress Cataloging Distribution Service. The abbreviated discussion below was based on an OCLC publication, Serials Format, 3rd ed. [OCLC, 1981]

For the Virginia Disc 2 project it was necessary to process OCLC MARC bibliographic serial records in order to :

1. Display bibliographic and holdings data on a CRT.
2. Prepare input files for the Personal Librarian indexing software.

The choice of which fields to display and which ones to index was based on several factors. The intention was to incorporate as much indexing capability as the fields could allow. Empirical observation, a technical paper that lightly touched on the topic [Markley and Calhoun, 1987], and comments from a seasoned cataloger [Norstedt, 1989], determined which fields to index with Personal Librarian. Over a hundred MARC records were examined to discover their contents and the frequency of use of the MARC fields. The following sections describe how the sample MARC records were processed to obtain the observations useful for deciding the suitability of the MARC fields for display and indexing (see Figure 6).

In the simplest terms, a MARC record is composed of fields which break down further into subfields. A field/subfield can be coded with a fixed set of choices or it can contain free form text (sentences and phrases) with a variable length. In order to process a MARC

record, one has to strip away the layers of pointers and other overhead in order to extract the essential fields and subfields.

The overall process of handling MARC data for Virginia Disc 2 is summarized in the subsections given below.

- **Read the tape correctly**

Major libraries in the US today have online linkage to OCLC and would have their MARC records transported by telephone lines. The Virginia Disc Two's MARC records came in the form of magnetic tapes. The lack of a blocking factor value on the label of the tapes from the Newman Library caused considerable grief to the author and considerable delay to the project. There is a danger that the first few short MARC record might appear complete -- with the wrong blocking factor. In the absence of a blocking factor, 2048 should be assumed. Pure Library of Congress MARC distribution tapes should have a blocking factor of 2048 bytes. The tape from the Virginia State Library was likewise read improperly the first time but experience gained from the Newman tapes resolved the problem readily. Figure 7 illustrates a typical MARC record structure.

OCTAL OFFSET																
0000000	0	0	5	3	3	n	a	s			2	2	0	0	1	8
0000020	1	I			4	5	0	0	0	0	1	0	0	1	9	0
0000040	0	0	0	0	0	0	8	0	0	4	1	0	0	0	1	9
0000060	0	3	5	0	0	1	5	0	0	0	6	0	0	4	0	0
0000100	0	1	8	0	0	0	7	5	0	4	9	0	0	0	9	0
0000120	0	0	9	3	0	9	0	0	0	1	3	0	0	1	0	2
0000140	2	4	5	0	0	1	3	0	0	1	1	5	2	6	0	0
0000160	0	4	0	0	0	1	2	8	3	0	0	0	0	3	5	0
0000200	0	1	6	8	3	6	2	0	0	2	6	0	0	2	0	3
0000220	6	9	0	0	0	4	0	0	0	2	2	9	7	1	0	0
0000240	0	2	9	0	0	2	6	9	7	8	0	0	0	5	4	0
0000260	0	2	9	8	036	o	c	m	0	3	9	7	2	8	6	8
0000300		8	6	1	1	1	8	036	8	2	0	4	0	1	c	1
0000320	9	7	7	9	9	9	9	a	t		q	r		p		
0000340						0	u	u	u	a	0	e	n	g		d
0000360	036			037	a	0	0	0	0	-	0	0	6	6	0	036
0000400			037	a	I	N	T	037	c	I	N	T	037	d	V	P
0000420	I	036			037	a	V	P	I	$	036			037	a	A
0000440	P	7	037	b	.	M	4	036	0	0	037	a	M	e	a	n
0000460	j	i	n	.	036	0	1	037	a	V	i	c	t	o	r	i
0000500	a	,	037	b	U	n	i	v	e	r	s	i	t	y		o
0000520	f		M	e	l	b	o	u	r	n	e	.	036			037
0000540	a				v	.	037	b	i	l	l	.	037	c	2	2
0000560		c	m	.		q	u	a	r	t	e	r	l	y	.	036
0000600	0		037	a	v	.		3	6	-				A	u	t
0000620	u	m	n		l	9	7	7	-	036		0	037	a	A	u
0000640	s	t	r	a	l	i	a	n		l	i	t	e	r	a	t
0000660	u	r	e	037	x	P	e	r	i	o	d	i	c	a	l	s
0000700	.	036	1	0	037	a	U	n	i	v	e	r	s	i	t	y
0000720		o	f		M	e	l	b	o	u	r	n	e	.	036	0
0000740	0	037	a		037	t	M	e	a	n	j	i	n		q	u
0000760	a	r	t	e	r	l	y	,	037	x	0	0	2	5	-	6
0001000	2	9	3	.	037	w	(	O	C	o	L	C	)	2	2	7
0001020	7	7	4	4	035	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0
0001040	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0
. . . .																
0004000	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0	\0

Figure 7. A Typical MARC Serial Record (2048 characters long)
Figure shows what a MARC record looks like when displayed with the UNIX "od -c" command.
Note: all the values are in ASCII except for 035, 036, and 037 which are in octal and \0 which is the octal zero or null.

- **Extract the MARC fields from the MARC records in tape format**

MARC records conform to the American National Standard for Information Interchange on Magnetic Tape (ANSI Z39.2-1979). The ANSI standard describes a schema for storing a record composed of fields but does not interpret the meaning of the fields. Programs were written to strip away the overhead information and extract the fields and subfields which are defined by the MARC standard.

The definition of the fields and subfields forms the MARC standard. Fields are identified by three digit field codes, subfields are further distinguished with a character code. C programs were written to separate the fields and subfields from the ANSI Z39.2-1979 overhead bytes. Control fields have a fixed length and do not need a special character (octal 037) to delimit the subfields. Variable fields can have variable lengths so a delimiter (octal 037) is needed to separate those subfields. The example MARC serial record in Figure 7 should yield the fields and subfields in Figure 8.

Each field code is composed of a field code number and a two character indicator. The indicator helps to disambiguate some of the field codes when the cataloging comes from a variety of sources. A subfield is introduced by a 037 octal (indicated by a slash) and a subfield code. Even at this stage, inspection of the record will help uncover the best fields to index on.

FIELD CODES	CONTENTS OF FIXED FIELDS
001	ocm03972868 861118
008	820401c1g77ggggat qr p 0uuua0eng d

FIELD CODES and INDICATOR	CONTENTS OF VARIABLE FIELDS
035	\a 00000-00660
040	\a INT \c INT \d VPI
049	\a VPI$
090	\a AP7 \b .M4
245 00	\a Meanjin.
260 01	\a Victoria, \b University of Melbourne.
300	\a v. \b ill. \c 22cm. quarterly
362 0	\a v. 36- Autumn 1977-
690 0	\a Australian literature \x Periodicals.
710 10	\a University of Melbourne.
780 00	\a \t Meanjin quarterly, \x 0025-6293. \w (OCoLC)

Figure 8. A MARC Record with the Fields Extracted. Subfields are delimited by a backslash (\) and a single character subfield code (i.e. \a)

The MARC standards committee had taken care to make the field and subfield codes common across material types. Common codes like "Geographic Area Code" appears as field code 043, in the format for books, films, manuscripts, music and serials. An understanding of one material format leads to understanding the MARC format for the other type of materials. This minor detail saved a lot of programming time because nearly the same data preparation program was used for the Newman serials data as was used to prepare the Virginia State library data.

The MARC standard follows a field naming convention based on the first digit of the three digit field code. The OCLC Serials Format, 2nd ed [OCLC, 1981] lists the grouping for field codes which is repeated in Figure 9.

The goal of preparing MARC records for CDROM publications required the extraction of the fields and subfields. What remained was to determine which fields are suitable for display and which fields are appropriate for searching.

Tag Group	Function
0xx	Bibliographic control numbers and codes (including call numbers and classification numbers).
1xx	Main entry headings
2xx	Title, title paragraph
3xx	Physical description
4xx	Series statement
5xx	Notes
6xx	Subject added entries
7xx	Added entries and linking entries
8xx	Series added entries and variant forms of entry
9xx	Locally-used fields

Figure 9. General Grouping of MARC Field Codes
(Any type of material)

- **Interpret the fixed field code**

In some cases it is conceivably useful to index on fields like LANGUAGE and COUNTRY OF PUBLICATION (see Figure 10). A query like, "Give me French books on cooking.", may benefit from the keyword FRENCH. In the application of Virginia Disc 2, however, none of the fixed fields was ever used for indexing. An observation of the only two candidate fields, LANGUAGE and COUNTRY OF PUBLICATION showed that :

- The same information tends to appear in other indexed variable fields of the MARC record, and
- The dubious value for indexing supplied by these two fields does not merit the effort in coding the table of country codes and the table of language codes.

SUBFIELD NAMES	INTERPRETATION OF CONTROL FIELDS
CNTL: 003972868	Control number is 003972868
Rec stat: c	Record Status. 'c' for corrected or revised record
Entrd: 820401	Date the record was entered is unknown
Used: 861118	Date the record was last used is 86/11/18
Type: a	Type of record 'a' Language material
Bib lvl: s	Bibliographic level 's' for Serial
Govt pub :	Government publication code
Lang: eng	Language code 'eng' for English <<<POSSIBLE INDEX
Source: d	Cataloging source code 'd' for Non-LC cataloging
S/L ent: 0	Successive/Latest Entry Designator '0' Successive
Repr:	Reproduction ' ' for Not a Reproduction
Enc lvl:	Encoding level ' ' Full-level LC or NLM catalog
Conf pub: 0	Conference Publication '0' means Not Conference pub
Ctry: at	Country of Publication is Australia <<<POSSIBLE INDEX
Ser tp: p	Serial Type 'p' for periodical
Alphabt: a	Original alphabet of title ' ' not given
Indx: u	Index availability 'u' for unknown
Mod rec:	Modified record code ' ' for not modified
Phys med:	Physical medium designator
Cont:	Nature of Contents is not specified
Frequn: q	Frequency : This is a quarterly publication
Pub st: c	Publication status 'c' for currently published
Desc:	Descriptive cataloging form " " pre-AACR2
Cum ind: u	Cumulative index availability 'u' for unknown
Titl pag: u	Title page availability 'u' for unknown
ISDS:	ISDS center code " " for no ISDS center
Regulr: r	Regularity 'r' for a regular publication (not erratic)
Dates:1977-9999	Beginning-Ending date of subscription

Figure 10. Interpretation of Control Fields of a MARC Serial Record
Note : Blanks values usually have default values

- **Interpret the variable field codes**

The variable field codes seem the most promising for indexing and display. Their contents are illustrated by the example shown in Figure 11.

FIELD CODES	INTERPRETATION OF VARIABLE FIELDS
035	\a **00000-00660** System Control Number is 00000-00660
040	\a **INT \c INT \d VPI** Codes for the Cataloging Source
049	\a **VPI$** OCLC Extension. Serial is located in VPI
090	\a **AP7 \b .M4** OCLC Extension. Locally assigned call no
245 00	\a **Meanjin.** Title Statement
260 01	\a **Victoria, \b University of Melbourne.** Imprint Place of publication is Victoria Name of publisher/distributor is the Univ. of Melbourne
300	\a **v. \b ill. \c 22cm. quarterly** Physical Description : a printed serial, illustrated, 22cm ...
362 0	\a **v. 36- Autumn 1977-** The beginning of the publication data : the enumeration (v. 36) and chronology (Autumn, 1977)
690 0	\a **Australian literature \x Periodicals.** Library of Congress assigned topical or subject heading
710 10	\a **University of Melbourne.** Another possible form of access
780 00	\a **\t Meanjin quarterly,** Previous title **\x 0025-6293. \w (OCoLC)** ISSN is 0025-6293 There is no Control number

Figure 11. Field Content and Definitions of a MARC Serial Record.

- **Choose fields useful for display**

A MARC record contains a lot of overhead. Unnecessary fields must be eliminated so that the display screens are intuitive and uncluttered. For Virginia Disc 2, the choice of fields to display was influenced by :

- Examining the VTLS public access catalog system at the Newman library
- Consulting a book about bibliographic displays [Crawford 1986].

Another modification required for displaying the fields on a CRT is to automatically word wrap long lines. Fields which exceed 70 characters need to be cut at word boundaries and wrapped around to the next line. This preparation was done for the Personal Librarian implementation. The Micro-VTLS package allowed some fields like the title field (i.e. MARC field code 245) to be truncated and did not provide as much display space as the field required. The PL version screens are thus likely to be easier to read and are more complete.

- **Choose fields useful for indexing**

The choice of the fields to index depends on the application. An example about using the LANGUAGE field for indexing was mentioned previously. The choice of fields to index is affected by the application's needs and the quality of the MARC records.

The choice for fields to index in the Personal Librarian is shown in Table 4. For Virginia Disc 2, indexing was of special importance because the two information storage and retrieval software packages under comparison started from the same data. The PL implementation offered the opportunity for utilizing the fields not touched by Micro-VTLS,

where these fields contained useful subject words. As many fields as possible, containing useful subject words, were included for indexing.

Table 4. List of Useful MARC Fields

FIELD CODE	TITLE	USES
50, 90, 99	CALL NUMBER	DISPLAY
100, 110, 111	AUTHOR	SEARCH, DISPLAY
245	TITLE	SEARCH, DISPLAY
260	IMPRINT	SEARCH, DISPLAY
265	ADDRESS	DISPLAY
300	PAGINATION	DISPLAY
310, 321	FREQUENCY	DISPLAY
350	PRICE	DISPLAY [removed upon suggestion by M.L. Norstedt]
362	PUBLISHED	For future use
410, 411, 440, 490	SERIES	SEARCH, DISPLAY
500, 520	SUMMARY	SEARCH, DISPLAY
525	SUPPLEMENTS	SEARCH, DISPLAY
550, 570, 580	NOTE	SEARCH, DISPLAY
600, 610, 611, 630, 651, 652 653, 690, 691	SUBJECT	SEARCH, DISPLAY
700, 710, 711	ADDED AUTHOR	SEARCH, DISPLAY
730	ADDED TITLE	SEARCH, DISPLAY
780, 785	CONTINUES	SEARCH, DISPLAY

The decision to maximize the number of fields to index was guided by the following facts :

- PL employed a term weighting scheme and was apt to handle too many terms better than too few terms. The term weighting allowed the document words that were more discriminating to be weighted above the other less descriptive terms. Too many fields was deemed better than too few fields.

- During the retrieval phase, PL allows the user to set which fields ought to be searched. The choice of fields to index was thus a retractable decision since the user can alter the search fields during the search session.

Of course fields which offered meager information had to be removed. The MARC fields were classified into :

- Fields which have ample subject terms -- at least one descriptive term for every three occurrences of a field code. These fields were included for searching.

- Fields which have less than one descriptive term in three occurrences but do contain useful terms. These fields were included in the PL source file and definition but were not included as search fields. A PL database definition accompanies each PL implementation and is contained in files with the .DEF extension.

- Fields which almost never contain useful terms. If these fields had no use as display fields, they were not included in the PL database definition.

Empirical observation of the MARC fields provided the best guide whether to include or exclude fields. The AUTHOR, TITLE and SUBJECT field classifications are naturally rich in terms which describe the contents of the document. The SUMMARY fields (e.g. 500 and 520) were suggested by Markley and Calhoun [1987] although the quality of the SUMMARY fields can vary from one manual cataloger to another. The justifications for including the other fields were acquired through observation by this author, applying the criteria mentioned previously.

The decision as to which words qualify as ample subject terms was admittedly subjective. The author used the criteria of judging those words which a user might possibly use (to issue a natural language query) as subject terms.

CHAPTER V

MICRO-VTLS

5.1 MICRO-VTLS PACKAGE

Micro-VTLS is a library automation system introduced by VTLS, Inc. in 1986. It runs on an IBM PC compatible or a network of IBM PC compatibles. The whole application package integrates Cataloging, Public Access Catalog Searching, and Circulation and Statistical Reporting functions of a small library. The Virginia Disc 2 includes only the Public Access Catalog module, i.e., the retrieval system visible to library patrons [VTLS, 1987]. The same type of software and database which would drive a public access terminal at a library is included on Virginia Disc 2.

Micro-VTLS is a database application developed in the Clipper database language developed by Nantucket [Milne, 1989]. Clipper figures prominently among micro languages optimizing the popular dBase III system of Ashton Tate, because of its open architecture and flexibility.

Micro-VTLS has the following features:

- **Searches by author, subject, title, call number, series, item number, ISSN, ISBN, LCCN and Bib-id fields**
  Micro-VTLS indexes on the first four characters of a field. For instance the words APPLIANCE and APPLE stem to the same four letters "APPL" and thus the two words have the same key. A query for APPLE will also yield matches for APPLAUSE. This indexing scheme is done to speed up retrieval in a hard disc system though not nescessarily for a CDROM implementation. The shorter the key, the faster and less overhead there is. The disadvantage is that too many documents are retrieved and the user may wonder why the search is too broad despite the specificity of the key used. With small collections, the user can tolerate the lower precision caused by an abbreviated key.

- **Searches by keywords inside author, subject, and title fields**
  Another part of the search software can look for word occurrences within

a field. This is slower than the abbreviated key version above and also has a different syntax.

- **Accepts Boolean queries**
  All the Boolean operators (AND, OR and NOT) are supported by Micro-VTLS. Complex Boolean commands can be issued in a single command line. There is no need to break down a query into smaller queries and to combine the results later with the Boolean operators.

- **Keeps result of old queries**
  Micro-VTLS stores a reference to the results of previous queries. These old results can be summoned by the user and can even be used to formulate more complicated Boolean queries.

5.2 DATA PREPARATION

VTLS, Inc. prepared two Micro-VTLS databases : one for the Newman library data and the second for the Virginia State library data. VTLS, Inc. regularly offers the service of converting MARC records to Micro-VTLS databases for its customers who are converting to Micro-VTLS for the first time. Funds from the State Council on Higher Education were paid to VTLS,Inc. for their services and to allow use of Micro-VTLS.

Their mainframe HP-3000 handled the MARC format utilizing the utilities of the HP-3000 based VTLS system. There was very little information provided on how the HP-3000 processed the MARC records, but the result was a downloadable file which a Clipper routine called ADDREF converted into dBase III representations of display screens. The reformatted MARC records were downloaded from the HP-3000 to a Novell network of IBM PC/AT compatibles. VTLS did not release the database schema nor any of the source code used by the retrieval program CALLOPAC. As for indexing, Micro-VTLS creates 13 inverted files, one for each access key and an extra index to link the bibliographic screen with the holdings screen. The twelve access keys are AUTHOR, SUBJECT, TITLE, CALL_NO, SERIES, ITEM_NO, KEYWORD, ISSN, ISBN, NETWORK_NO, LCCN and BIB-ID.

A month's delay was experienced towards the end of the project. Micro-VTLS encountered errors regarding the integrity or consistency of the files downloaded from the mainframe, which interfered with the loading of the Micro-VTLS database.

The new release of Micro-VTLS version 2.0 also had problems with indexing words adjacent to punctuation marks. VTLS,Inc. acknowledges the problem and plans to make corrections.

Data preparation for Micro-VTLS entailed the following (see Figure 12) :

1. MARC tapes were read into the VTLS system.
2. Files were downloaded to a microcomputer through a Novell network connected to the HP-3000.
3. The data files were translated to dBase III formats.
4. Inverted index files were built for each of the twelve available access keys and a final index to link the bibliographic screen to the holdings screen. A program named INDEX.EXE, compiled in Clipper, was used to create the index files (.NTX) from the dBase files (.DBF).
5 The database files, programs and support files were placed into one directory, ready for integration with the rest of the CDROM databases and programs.

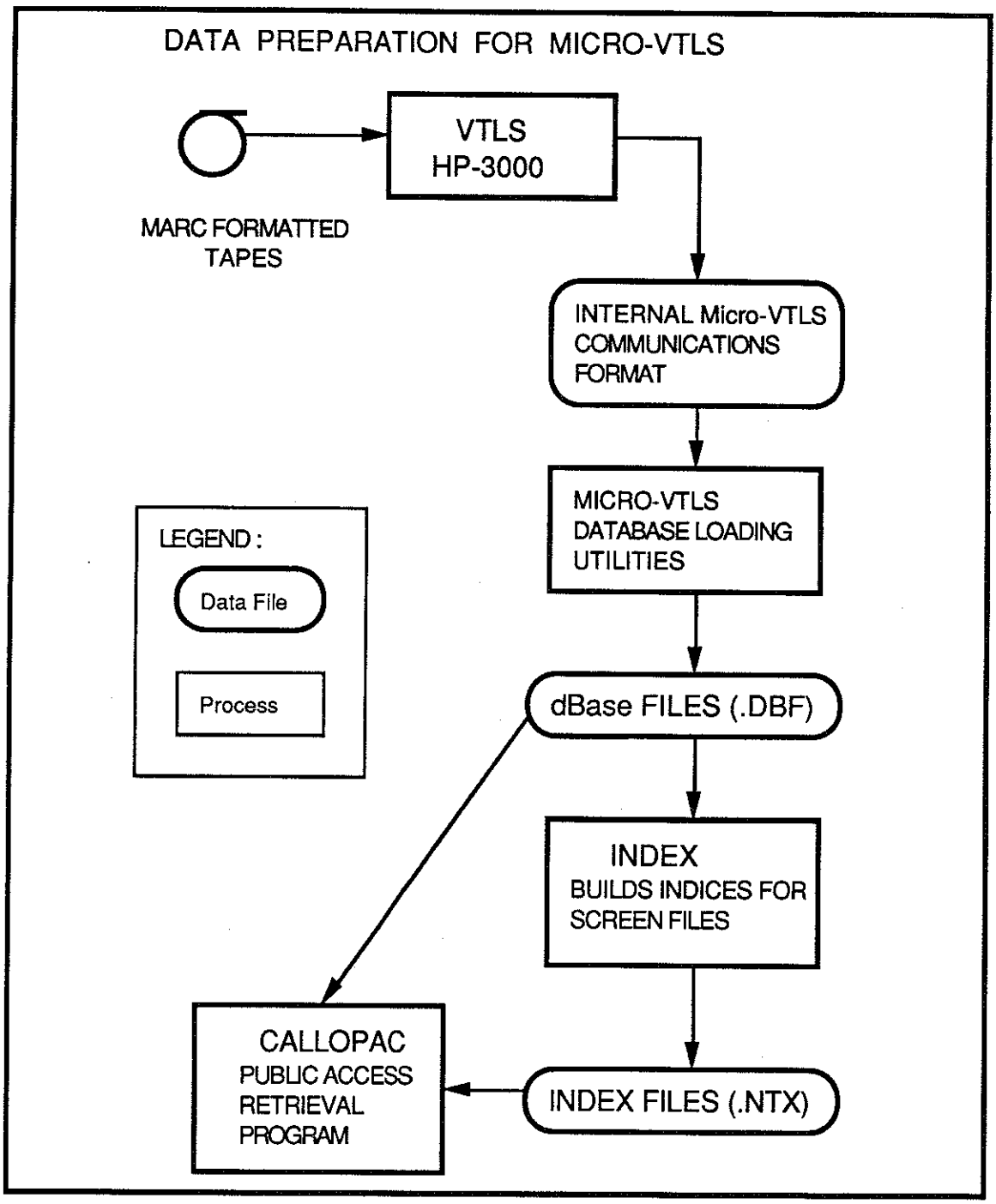


Figure 12. The Micro-VTLS Data Preparation Process

Table 5 is a list of MARC code fields that Micro-VTLS used for indexing and display. The list has some similarity to the list prepared for Personal Librarian. A notable difference is the Call Number field code used. The Personal Librarian implementation uses the Library of Congress call number (field code 050) first if it exists and then tries the locally assigned (field code 090) call number. The Micro-VTLS tries to assign the local call number (field code 090) first. The disparity occurs only in the Virginia State Library collection where both the local call number and the Library of Congress call number occasionally exist for the same record. There was no coordination between the VPI&SU team and VTLS,Inc. on which field code to use. The problem was discovered too late -- after both PL and Micro-VTLS versions had been completed. The Micro-VTLS and PL versions can still be compared with each other even with this problem of the call number used. The problem is mentioned here to reduce the consternation of users trying to compare the two systems for the first time.

Virginia Disc 2 incorporated the latest software release of Micro-VTLS called ver 2.0. Compared to a prior version, 1.4c, Micro-VTLS can do keyword searching, that is, matching the user's query term with any word embedded in the AUTHOR, TITLE or SUBJECT fields. The previous matching capability of Micro-VTLS compared only the first four characters of a key field and failed to match words located inside a field.

Table 5. MARC Fields Indexed or Displayed by Micro-VTLS

FIELD NAME	MARC FIELD CODES USED
Author Name	100, 110, 111, 710
Subject	600, 610, 611, 630, 650, 651, 652, 653, 690, 691
Title	245
Call Number	090, 099, 050
Series	400, 410, 411, 440, 490
Word	100, 110, 111, 710, 600, 610, 611, 630, 650, 651, 652,653, 690, 691, 245,400, 410, 411, 440, 490
ISSN	022
ISBN	020
Language	008
LCCN	010

CHAPTER VI

PERSONAL LIBRARIAN

6.1 PERSONAL LIBRARIAN PACKAGE

Personal Librarian (PL) is an information retrieval system that can run on a variety of machines including IBM PC systems. It can do full-text retrieval as well as structured field retrieval. Personal Library Software, Inc. of Maryland, graciously allowed use of PL in the Virginia Disc 2 project. PL boasts of the following features:

- **Accepts natural language requests**
  PL can accept English phrases but it doesn't have a true natural language interface. The package does not do syntactic or semantic analysis of the query. PL assumes that the input lists all keywords of interest.

- **Presents the relevant documents in ranked order**
  PL orders the relevant list according to how relevant the documents are by PL's own unique calculation. This is useful in order to limit the user's browsing over probably irrelevant material. This facility is discussed further, later on in this report.

- **Does automatic matching on word stems**
  Word stems are a good fit for matching because the plural form of a noun, past tense of a verb ... etc. could cause some systems to lose possibly relevant documents.

- **Has an expand function to statistically determine a set of terms to replace a given word**

- **Can process a query with specifications for : Adjacency of words, proximity of words, numeric range operators and Boolean operators**

- **Can do truncation and wildcard searching**

- **Accepts a stop word list for each collection**
  A stop word list is a list of words which are too common or otherwise inappropriate to index on.

- **Has a facility for customizing displays**
  PL provides for the facility to run one program after a document has been selected for display. The program could be a user written program to

display the data in a more readable manner. Such a program was written for Virginia Disc 2.

Because PL is a general information retrieval system, several changes were necessary to adapt PL to handle MARC records. These are :

- Programs were written to support the display of MARC records, bibliographic screens, and holdings screens in a clear, intuitive manner.
- Words were added to the stop word list of each collection in order to tune the retrieval performance of PL. Space is reduced and performance improves if one omits common words which do not distinguish between documents or are inappropriate for indexing. Terms like "Virginia" and "Va" appear too often in the Virginia State Library collection were added to the stopword file called NEGDIC.WRD.
- The MARC fields had to be evaluated in order to determine whether they were fit for indexing/searching and displaying (see section 4.5). Reformatting was also necessary for the fields used for indexing and displaying. For instance, some subfields had to be grouped together. Other fields which were too long had to be reformatted into several lines of 70 columns for proper display on a CRT.

6.2 DATA PREPARATION

Data preparation for Personal Librarian was difficult because of some problems associated with the tape transfers. Reading in data from the tapes had to be done twice due to a wrong choice in the tape blocking factor. The Newman library had then disposed of the original copy of the tape and were unwilling to do a retaping. The project was near being aborted for loss of the data. Fortunately, VTLS,Inc through the helpfulness of Deveron Milne lent the author a copy of their tape. VTLS, Inc was doing in parallel a preparation for the Micro-VTLS version of the bibliographic database.

The other major problem to the project concerned the tape from VTLS which was recorded using an HP backup facility. Dr Fox tried several load sequences that succeeded in loading a major portion of the files. Although the files looked perfect at the time, unexpected

sequences appeared during subsequent processing, doubtless related to the proprietary tape backup format. Elaborate programs had to be written to recreate the original data -- more elaborate than the actual processing done to the MARC files. The delay to the project (around one year) can be traced to the early error in a choice of a tape blocking factor and attempt to use data provided in nonstandard form. The key lesson to learn is to always thoroughly check data after each stage of processing.

From the VAX, the FTP file transfer utility (see Figure 13) enabled the downloading of files for manipulation on an IBM PS/2. After writing some programs, it was discovered that the downloaded MARC records were faulty. Many filtering programs were written to remove or skip over damaged records. An example of a filtering program was a program to excise segments of damaged records which caused the data not to align on 2048 byte boundaries. MARC records always occupy multiples of 2048 bytes. The difficulty lay in determining whether errors during new program development were caused by faulty data or erroneous programs run earlier.

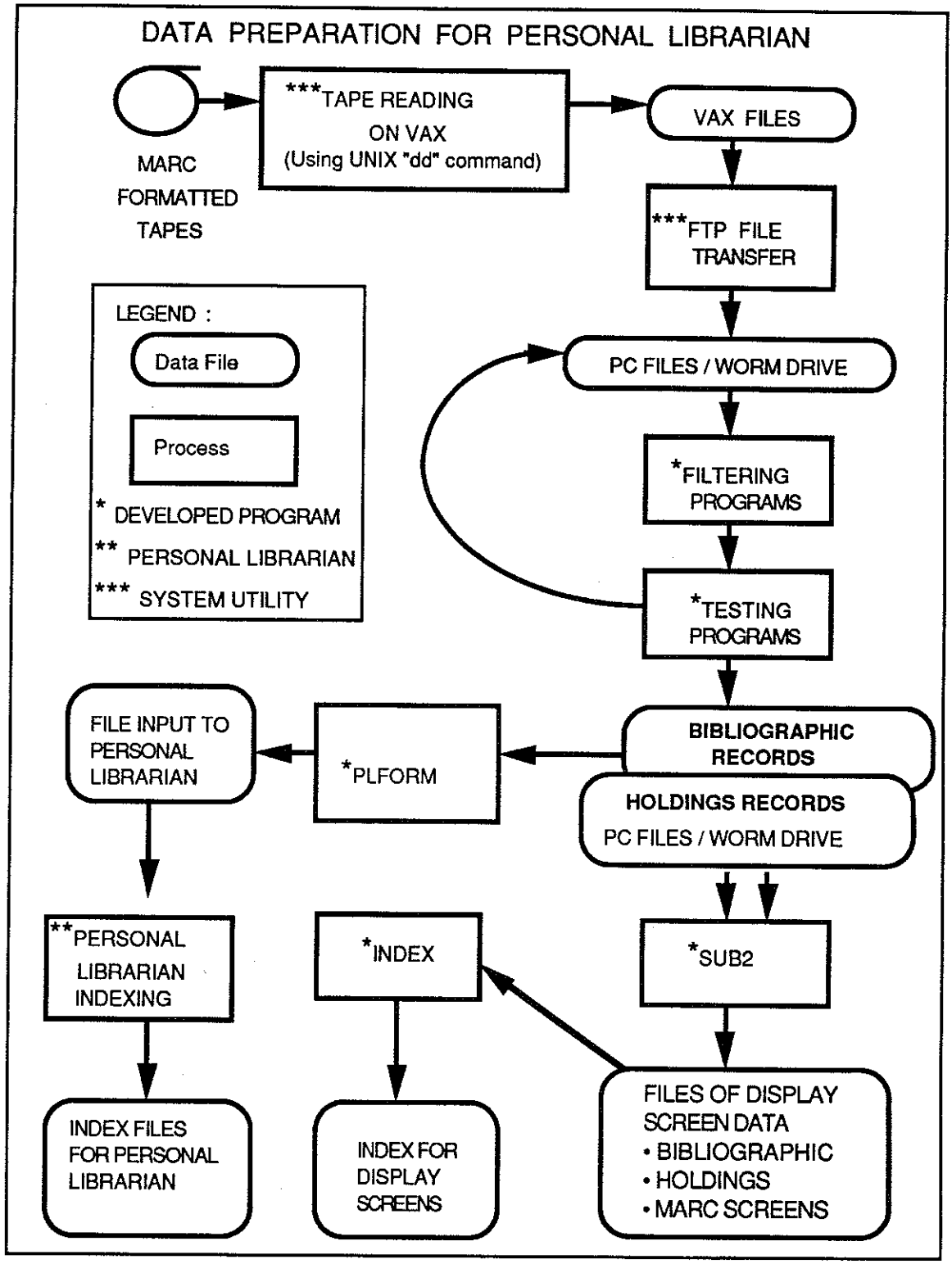


Figure 13. The Personal Librarian Data Preparation Procedure.

The PLFORM program used the MARC bibliographic records to form the input file to Personal Librarian. An example of the input to Personal Librarian is shown in Figure 14. The PL indexing program proved very reliable and numerous alternatives to constructing the PL schema or database structure were tried out.

The SUB2 program combined the MARC bibliographic and MARC holdings records and produced screen display data as illustrated in Figure 15. The screen display data was used by the display enhancement program called Z.EXE. Z.EXE is a facility to display to the user the contents of the document of interest. The INDEX program builds the index file which Z.EXE consults. After the three files at the bottom of Figure 13 were constructed, the indexing phase was complete.

```
-CALL_NO-
 AP7 .M4
-TITLE-
 Meanjin.
-IMPRINT-
 Victoria, University of Melbourne.
-PAGINATION-
 v. ill. 22 cm. quarterly.
-PUBLISHED-
 v. 36- Autumn 1977-
-SUBJECT-
 AUSTRALIAN LITERATURE   PERIODICALS.
-ADDED_AUTHOR-
 University of Melbourne.
-CONTINUES-
  Meanjin quarterly, 0025-6293. (OCoLC)2277744
-END-

-CALL_NO-
 BF173.A2 P69
-TITLE-
 Psychoanalysis and contemporary thought.
-IMPRINT-
 [New York] International Universities Press, 1978-
-ADDRESS-
 315 Fifth Ave., 10016
-PAGINATION-
 v. 23 cm.
-PRICE-
 $40.00 (individuals) $50.00 (institutions)
-PUBLISHED-
 v. 1- 1978-
-SUMMARY-
 "A quarterly of integrative and interdisciplinary studies."
-SUBJECT-
 PSYCHOANALYSIS   PERIODICALS.
-END-

-CALL_NO-
 BJ47 .J6
-TITLE-
 The Journal of religious ethics.
-IMPRINT-
 [Waterloo, Ont., American Academy of Religion]
-ADDRESS-
 CSR Executive Office, Wilfred Laurier University, Waterloo,
 Ont. N2L 3C5 Journal of religious ethics, Scholars Press,
 Missoula, MO, 59801
-PAGINATION-
 v. 23 cm.
-PRICE-
 $6.00 (individuals) $8.00 (institutions)
-PUBLISHED-
 v. 1- fall 1973-
-NOTE-
 Vols. for 1973-77 issued by the American Academy of
 Religion; 1978-spring 1981 cooperatively sponsored by the
 University of Tennessee and Kennedy Institute of Ethics at
 Georgetown University, and: 1980-spring 1981 sponsored by
 the University of Virginia, University of Notre Dame, Emory
 University; fall 1981- edited at Rutgers University at
 University of Notre Dame with the cooperative sponsorship of
 the University of Tennessee, University of Virginia, Emory
 University, and the Kennedy Institute of Ethics at
 Georgetown University.
-SUBJECT-
 ETHICS   PERIODICALS.
-ADDED_AUTHOR-
 American Academy of Religion. cn
 University of Virginia.
 Emory University.
 University of Notre Dame.
 University of Tennessee, Knoxville.
 Kennedy Institute.
 Rutgers University.
-END-
```

Figure 14. A Sample File Input to Personal Librarian

```
-N0000-00660

 CNTL: 003972868  Rec stat: c Entrd: 820401     Used: 861118
 Type: a Bib lvl: s Govt pub:   Lang: eng  Source: d S/L ent: 0
 Repr:   Enc lvl:   Conf pub: 0 Ctry: at   Ser tp: p Alphabt: a
 Indx: u Mod rec:   Phys med:   Cont:      Frequn: q Pub st: c
 Desc:   Cum ind: u Titl pag: u ISDS:      Regulr: r Dates: 1977-9999

  1.    35    0000-00660
  2.    40    INT \c INT \d VPI
  3.    49    VPI$
  4.    90    AP7 \b .M4
  5.   245 00 Meanjin.
  6.   260 01 Victoria, \b University of Melbourne.
  7.   300       v. \b ill. \c 22 cm. quarterly.
  8.   362 0  v. 36-   Autumn 1977-
  9.   690  0 Australian literature \x Periodicals.
 10.   710 10 University of Melbourne.
 11.   780 00   \t Meanjin quarterly, \x 0025-6293. \w (OCoLC)
>

CALL NUMBER : AP7 .M4
Title       : Meanjin.
Imprint     : Victoria, University of Melbourne.
Pagination  : v. ill. 22 cm. quarterly.
Published   : v. 36- Autumn 1977-
Subject     : AUSTRALIAN LITERATURE -- PERIODICALS.
Added Author: University of Melbourne.
Note        : CONTINUES
               Meanjin quarterly, 0025-6293. (OCoLC)
>

   CNTL : 0000-00660   Acq stat: 4   Method :   Int canc:
   Enter: 860204 Ret: 8 Ret cd:     Complet: 0   Copies:
   Lend : a    Repro: a   Lang: eng  Sp ind: 0   Update: 861202
  1.    90    AP7 \b .M4
  2.   245 00 Meanjin.
  3.   780 00 Meanjin quarterly,
  4.   852    0100
  5.   853 11 8 \a v. \b no. \u 4 \v r \i (year) \j (season) \w q \x 23
  6.   863 40 8.1 \a 36-44 \i 1977-1985 \j 23-22
  7.   866 40 1 \a Official
>

CALL NUMBER : AP7 .M4
TITLE       : Meanjin.
LOCATION    : Copy 1 NEWMAN
STATUS      : Currently received

Official
v. 36-44  fall 1977-summer 1985
-N0000-01760

 CNTL: 003946816  Rec stat: c Entrd: 850821     Used: 850821
 Type: a Bib lvl: s Govt pub:   Lang: eng  Source:   S/L ent: 0
 Repr:   Enc lvl:   Conf pub: 0 Ctry: nyu  Ser tp: p Alphabt: a
 Indx: u Mod rec:   Phys med:   Cont:      Frequn: q Pub st: c
 Desc:   Cum ind: u Titl pag: u ISDS: 1    Regulr: r Dates: 1978-9999

  1.    10       78643466
  2.    12    2 \b 3 \j 0 \l 1
  3.    22 0  0161-5289
  4.    30    PCTHDS
  5.    35    0000-01760
```

Figure 15. Contents of a File with Display Screen Data

6.3 CUSTOMIZING THE USER INTERFACE

One of the time consuming tasks for the Virginia Disc 2 project was to prepare a document display facility for PL. The original display screens of PL did not visually appeal to the author and probably not to a library patron. PL essentially echoes the format of the records as they were entered into PL (compare Figure A-9 and Figure A-10). Realizing the inadequacy of the display format, it was necessary to write a program that would display MARC records and aid the user in browsing the retrieved document in a simple manner.

6.3.1 Processing to Prepare the Display Screens

The display screen data files are text files (i.e. created by SUB2 in Figure 13) ready to be manipulated for displaying. There were four display screens for each MARC item in the Newman Library collection. The Virginia State Library collection, which had no holdings information, required only two display screens. These screens were created by extracting the appropriate fields from the MARC bibliographic record and merging this with the corresponding MARC holdings record and formatting the result for display. Table 4 is a summary of the MARC fields which were considered useable for search/index or display purposes.

6.3.2 Runtime Environment of the Display Facility

PL allows a developer written program to be invoked from PL's command line. This facility allows a developer to write a program which is passed the document number of the current document being examined under PL. Documentation for customizing PL by altering a file named STRINGS.DAT is not available in the Personal Librarian manual and the information was learned by the author by examining similar implementations in Virginia

Disc 1. In the STRINGS.DAT file, line number 62 identifies a program name which could be directly invoked from the command line of PL. The name of the program is also the actual command under PL, thus Z.EXE is chosen for the developer defined program. The developer is entirely responsible for what the program does and in most cases, as in this implementation, the custom program takes the document number and with an appropriate file (or files), retrieves several screens about the document under observation. The effort required lies not only in writing the display program but also in preparing the screen images to display (see Figure 16).

Close attention must be paid to ensuring that enough memory is available to the application program space in the PC running MSDOS and Virginia Disc 2. Too many memory resident programs and device drivers might not leave enough space in the first 640 Kbytes for Virginia Disc 2 to run. Moreover, the special display facility for the Personal Librarian runs while two other layers of software, AUTOMENU and Personal Librarian are memory resident. The whole Virginia Disc 2 application successfully ran with 426 Kbytes available for application programs. Lesser memory ceilings in the area available for application programs have not been tried yet by the author.

6.3.3 Iterative Refinement

The software underwent continuous improvement over two weeks.This author repeatedly showed the software product to a variety of persons, solicited their comments, evaluated the suggestions and rewrote some of the display code. The PL command interface could not be changed, so only the display facility was involved in the revisions.

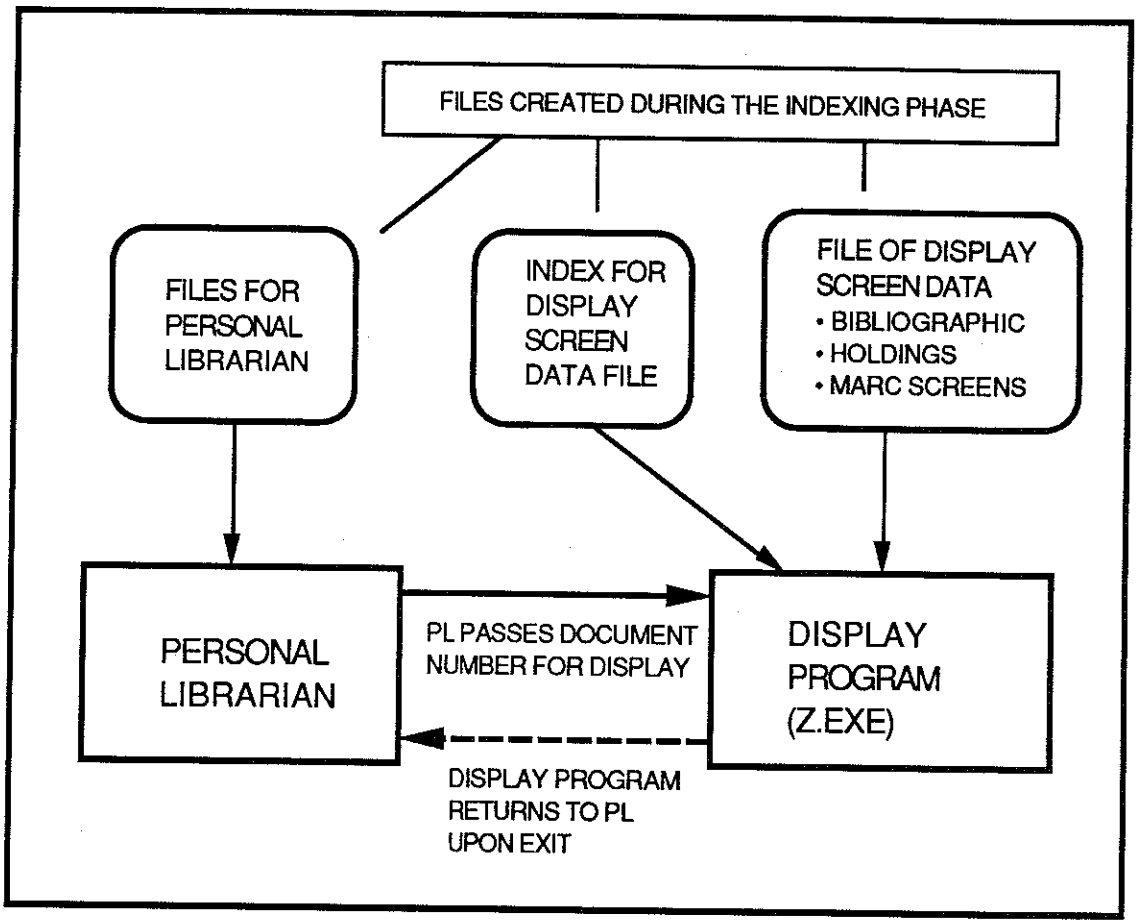


Figure 16. Runtime Environment for Customized Display Program

CHAPTER VII
COMPARISONS BETWEEN PL AND Micro-VTLS

Preliminary tests done on Virginia Disc 2 compared the performance of Micro-VTLS and Personal Librarian. Test 1, which measured the response time of both packages, was conducted in a Meridian premastering system at an AT&T facility in Greensboro, NC. The Meridian system allows a CDROM's realtime performance to be gauged prior to pressing a master disc. Test 2, which deals with recall/precision measurements, was completed with a PS/2 Model 80 with an IBM 3363 WORM (Write Once Read Many) disk system. Both preliminary tests can be extended to a complete performance comparison of the two systems.

7.1 RESPONSE TIME MEASUREMENTS

Response time varies as a function of the hardware, the logical and physical file structures, and the retrieval mechanism (whether Micro-VTLS or Personal Librarian was used). The Meridian Data Publisher was a good venue for a balanced test, barring the availability of the finished Virginia Disc 2 CDROM. The test was conducted with a Meridian Data,Inc. machine and SCSI controller emulating a CDROM drive connected to a PC compatible with an 80286-processor. Table 6 shows the response time in seconds of Micro-VTLS and Personal Librarian when processing simple queries selected randomly by the author. Values were timed on a digital watch with a 1 second resolution. Table 6 shows the response time in seconds of Micro-VTLS and Personal Librarian when processing a single query selected by this author, for each of a number of common forms of query.

Table 6 . Time Response Comparisons of Micro-VTLS and PL

Micro-VTLS on Newman Library Data (Time in seconds)

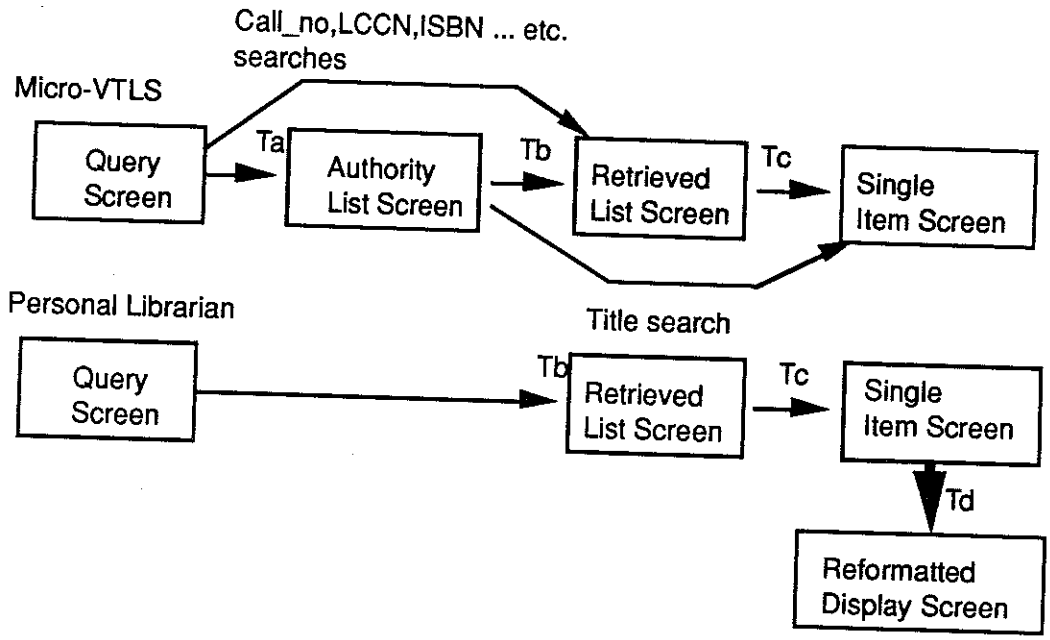
Type of Query	Micro-VTLS Query	Ta	Tb	Tc	Ttotal
Author Search	A\smith	11	12	23	46
Title Search	T\smith	10	-	22	22
Call Number	C\HV1	-	18	21	39
Subject	S\social	18	12	20	50
Keyword search	W\W:hotel	106	18	17	141

Personal Librarian on Newman Library Data (Time in seconds)

Type of Query	Micro-VTLS Query	Tb	Tc	Td	Ttotal
Author Search	smith:AUTHOR	16	0	7	24
Title Search	smith:TITLE	13	0	9	22
Call Number	HV1*:CALL_NO	20	0	8	28
Subject	social:SUBJECT	23	3	7	33
Keyword search	hotel	17	2	7	26

(Notes appear on the next page)

Table 6 . Continued ...
Time Response Comparisons of Micro-VTLS and PL



Note:
The response times of Micro-VTLS and Personal Librarian can be compared when we measure the time it takes to move from the issuance of the query to the display of one document.
Micro-VTLS uses 2 or 3 screen transitions before displaying a screen with one document (see Appendix A). Author, subject or keyword searches need 3 transitions thus incurring a total response time of Ta+Tb+Tc. A title search skips the retrieved list screen and the response time is Ta+Tc. Micro-VTLS searches for CALL_NO, LCCN, ISSN, ISBN, SERIES, BIB-ID, ITEM and NETWORK do not use the authorities screen and have a response time of Tb+Tc.

Personal Librarian does not have an authorities screen and PL's response time is Tb+Tc. The transition time Td accounts for the time it takes to move from the display of one item or document to an enhanced display (see Appendix A). The enhanced display facility was written to display bibliographic information in several easy-to-read set of screens. Td is not a performance penalty that should be added to PL's response time. Td is shown on Table 6to illustrate the speed of the display enhancements.

The transitions of screens is key to understanding the values on the table (see also appendix A). The idea is to measure the time needed for the retrieval mechanism to move from the point the query is issued to the time a single document is displayed on the screen. The user's response time is ignored because the test measures only the retrieval mechanism. Micro-VTLS employs more screen transitions (2 or 3) than Personal Librarian (2 transitions).

7.2 TESTS FOR RECALL AND PRECISION

Precision and recall are two criteria for evaluating an information retrieval system mentioned in [Salton,McGill 1983]. Others include response time, user effort needed, presentation and collection coverage. Precision and recall measures can be quantified but the results depend heavily on the manner in which the test is conducted. Determining which documents are relevant is a value judgement, moreso when the user can only see the bibliographic information of a document. Also, there is the selective view of relevance based on the state of the user's knowledge at the time of the search. The user's skill or vocabulary may significantly affect the performance of searching with the retrieval system. Finally, recall is not measurable in a large collection where human experts are not likely to examine every document.

Despite all these caveats, the following test was done to compare Personal Librarian and Micro-VTLS. Subjective decisions were minimized by fixing the rules prior to the search. The collection used for the test was the Virginia State Library collection of materials related to Virginia's history. The query about hotels was chosen over several other candidate queries like schools, monuments, etc ... (terms which also deal with Virginia's history)

because the documents retrieved by the query about hotels could be individually examined by the author without being too few to invalidate the comparison.

Search Situation : A book about the hotels in Virginia was being written. The book could include hotels of historical value but the thrust is about the major hotels in the state -- the history, ownership and operation.

Further assumptions : The user does not know of any hotel name in or out of the state. The user does not care for peripherally relevant topics like taverns, motels,etc.

Conduct of the Test : All the decisions about document relevance were determined by the author. In case of doubt, the document was considered not relevant.

Tables 7-9 tabulate the results. Table 7 identifies the relevant documents found using Micro-VTLS and those found using Personal Librarian. Table 8 identifies the rejected documents and the reason for rejection. Personal Librarian proved to be superior in terms of recall and precision for this sample query. The database was not scanned individually due to the size of the database. For the recall computation, it was assumed that the union of the relevant PL documents (17) and the relevant Micro-VTLS documents (6) constituted the total population of relevant documents (18). Table 9 and Figure 17 show the superiority of Personal Librarian especially in recall.

Table 7 List of Relevant Documents

Personal Librarian	Micro-VTLS
F232.R4 H67	
F232.R4 H67	
F234.V8 N8	
F234.S7 H82 19--	F234.S7 H2 19--
E481.P4 J5	
F234.P4 J26	
F232.G4 M915 1890	
F234.O4 H82 1917	F234.O4 H82 1917
F234.O4 H8	F234.O4 H8
F232.R67 N262	
F232.R67 N27	
F234.N48 C7	
F234.B4 H8	F234.B4 H8
F234.C47 M7	
F234.B9 B9	
F234.O4 W93	
F234.W7 B59	
	F233.625.H6
COUNT = 17	COUNT = 6

Table 8. List of Non-Relevant Documents

REASON FOR NON-RELEVANCE	Personal Librarian	Micro-VTLS
Talks about an event in the hotel	E186.6 .N39 no.34 F234.W7 G64 F234.O4 A213	F234.O4 A213
Tour organized by hotel or Resort Guide	F233.3 .W5 F233.3 .W5 1938 F233.3 .W5 1940a F233.3 .W5 1950 F231 .R43	F231 .R43 F227 .B25
History (hotel incidental)	F227 .T77 F234.O4 D1 F227 .T77	F232.R7 C77
About sights, not hotel	F232.L92 A5 1886 F232.L92 A5 1890	
Fiction	F232.D7 P8	
About a tavern, not a hotel		F234 W7 C552 F232.F6 F6 no 28-29
TOTAL NON-RELEVANT	14	6

Table 9. Recall/Precision Results for a Query about Hotels.

	Personal Librarian	Micro-VTLS
PRECISION	17 out of 31 0.548	6 out of 12 0.5
RELATIVE RECALL	17 out of 18 0.944	6 out of 18 0.334

Retrieval systems which return a ranked output list of documents can have a recall/precision plot shown for each query. The procedure is illustrated in [Salton and McGill, 1983] and is done here for the ranked output from the query about hotels in Virginia. Since we are comparing two retrieval systems, the relative recall measurement seems adequate and there is no obligation to estimate the total number of relevant documents in the collection. In other words, the measure of recall done on both collections assumes that the union of the relevant documents retrieved by Micro-VTLS and PL comprise the whole set of relevant documents in the collection. Table 10 shows a list of items PL returned, ordered by their (presumed) relevance along with the recall and precision computation.

The values listed in the Table 10 are plotted in Figure 17 to illustrate the better relative recall/precision exhibited by Personal Librarian. The thick lines in Figure 17 form the interpolated recall/precision curve for the sample query about hotels.

Table 10. Recall/precision Table after the Retrieval of n Documents for PL.

Note : X indicates a relevant document.

Rank	Relevant	Call No.	Recall	Precision
1	X	F232.R4 H67	.056	1.0
2	X	F234.57 H82 19--	.111	1.0
3	X	F234. V8 N8	.166	1.0
4	X	E481 P4 J5	.222	1.0
5	X	F234.P4 J26	.278	1.0
6	X	F232.G4 M916	.333	1.0
7		F232.L92	.333	0.857
8		F232.L92 A5 1890	.333	0.75
9	X	F234 .O4 W93	.389	0.778
10	X	F234.O4 H82 1917	.444	0.80
11	X	F232.R67 N262	.5	0.818
12	X	F232.R67.N263	.555	0.833
13	X	F232.R67 N27	.611	0.846
14	X	F234.O4 H8	.667	0.857
15		F233.3 .W5	.667	0.80
16		F233.3 .W5 1988	.667	0.75
17		F233.3 .W5 1940a	.667	0.706
18		F233.3 .W5 1950	.667	0.667
19	X	F234.b4 H8	.722	0.684
20	X	F234.B9 B9	.778	0.7

Table 10. Continued ...
Recall/precision Table after the Retrieval of n Documents for PL

Note : X indicates a relevant document.

Rank	Relevant	Call No.	Recall	Precision
22	X	F234.N48 C7	.889	0.727
23		F234.O4 A213	.889	0.696
24		F234.O4 D1	.889	0.667
25		F233.3 .R52 1911	.889	0.640
26		E186.6 .N39 no34	.889	0.615
27		F232.L92 A5 1890a	.889	0.593
28		F232.D7 P8	.889	0.571
29		F231.R43	.889	0.552
30		F227.T772	.889	0.533
31		F227.T77	.889	0.516
32	X	F234.W7 B59	.944	0.531
33		F234.W7 G64	.944	0.515

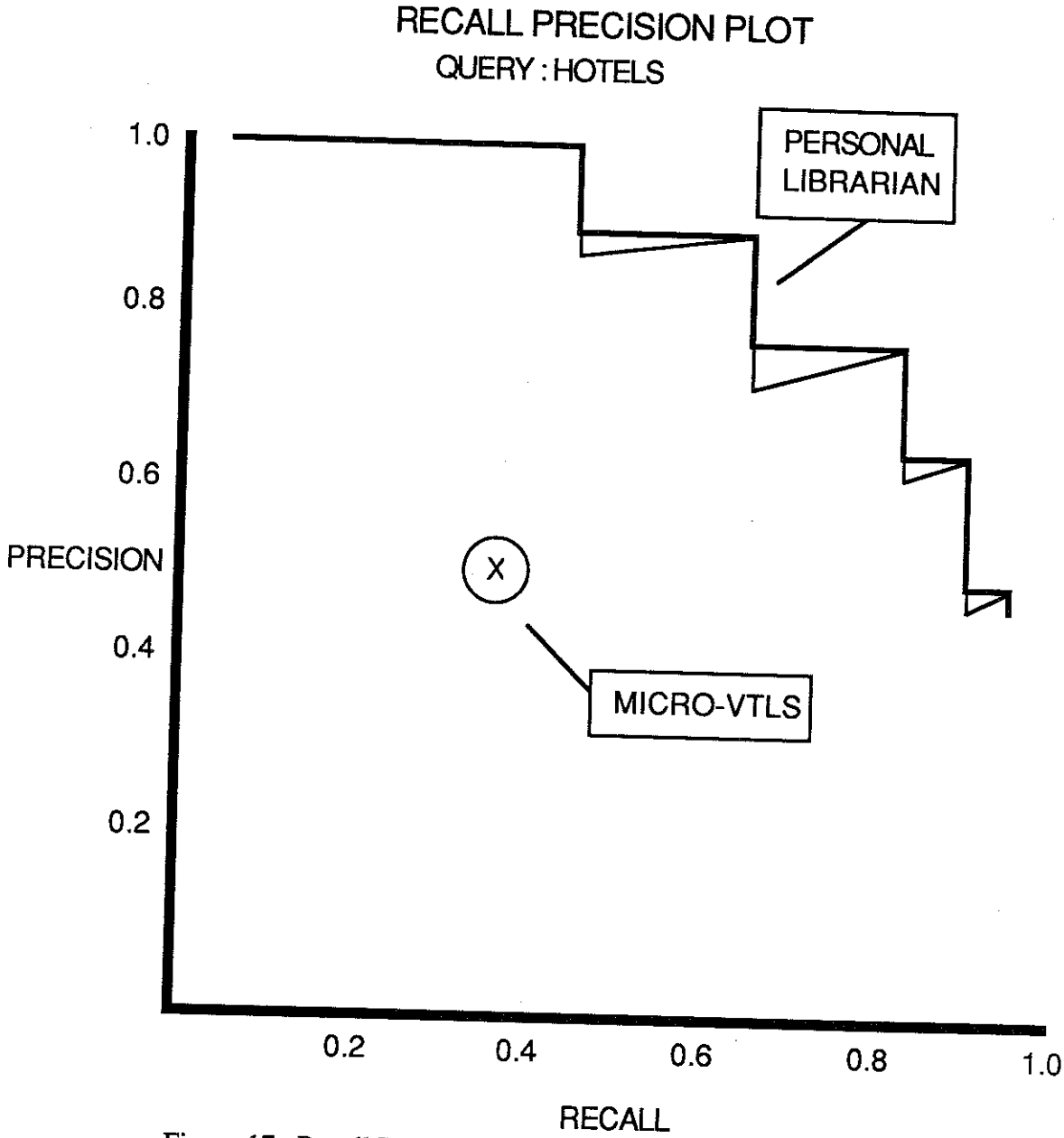


Figure 17. Recall Precision Plot for Micro-VTLS and PL. Micro-VTLS is inferior in the recall/precision measure for the query about hotels on the Virginiana collection.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

8.1 CONCLUSIONS

The findings for the project can be grouped into i) preliminary performance measurements of Virginia Disc 2 and ii) lessons gained about building CDPACs.

Many of the project's goals were met but the delays encountered in the data preparation stage seriously eroded the currency of the information. From an academic viewpoint, Virginia Disc 2 still provides a good vehicle for comparing the performance, under controlled conditions, of a conventional Boolean retrieval system with inverted files and a SIRE-based retrieval system.

Here is a compilation of lessons learned in publishing Virginia Disc 2 :

- Programs that handle MARC records should be written to take advantage of the consistency in the assignment of field codes by the MARC standards committee. This way it is easy to adapt programs for processing one MARC material type, say books, to other material types like serial, maps, films, etc ... .
  The universal presence of MARC records assures that CDPACs can be built from the same type of programs that were used to build Virginia Disc 2. The implementation method can be repeated for different files of MARC records.

- It is imperative to check data after each stage of processing. Sometimes spot checks will not uncover all the problems.

- The usefulness of a common format for transporting information among teams coordinating in a CDROM publication cannot be overstated. Transporting data between VTLS,Inc., VPI&SU and Nimbus was a recurrent problem.

- MARC holdings information may well be inappropriate for CDROMs because of its volatility. Holdings information generally change too frequently to keep on a CDROM.

The comparison of Personal Librarian and Micro-VTLS was done on the following basis :

• Both information storage and retrieval systems started with the same set of MARC records.

• Micro-VTLS utilizes the following fields for indexing : SUBJECT, AUTHOR, TITLE, CALL_NO, and eight other minor fields.

• Personal Librarian has the capability to index on any field. In this implementation, PL utilized other MARC fields like IMPRINT, SUMMARY, SUPPLEMENTS and several other fields not accessed by Micro-VTLS.

• Micro-VTLS's most comprehensive search capability, the keyword search, scans the SUBJECT, TITLE, and AUTHOR fields for exact keyword matches. PL automatically scans all fields on its search list and has the capability to weight its terms so the more descriptive index terms are better emphasized during matching. In addition, PL offers advanced features like word proximity searching and ranking the output documents according to an estimate relating to relevance.

The time measurement tests show that a retrieval system built with an optimizing language for a dBase III system (Micro-VTLS) may not be the best way to construct a retrieval system for CDROMs. The slow latency period of a CDROM might contribute intolerable delay to a system known to be fast with a hard drive.

The merit of developing a special display facility for Personal Librarian which operates on a single MARC record at a time was observed. Keeping a small index on the hard drive instead of the CDROM also helped improve the performance of the customized bibliographic display facility.

8.1.1 Preliminary Comparisons Between PL and Micro-VTLS

Salton and McGill outline six criteria for evaluating retrieval effectiveness. The results that follow (see Table 11) are preliminary and are a good example of how a comparison of PL and Micro-VTLS could be done. The following table shows the result of 3 of the 6 comparisons done. Personal Librarian shows slight superiority in three of the facets listed. A full-blown study will be done later.

Table 11. Preliminary Comparison of Micro-VTLS and Personal Librarian

Criteria	Micro-VTLS	Personal Librarian
1. Response time		superior
2. Precision		superior
3. Recall		superior
4. User effort needed	no data	
5. Presentation	no data	
6. Collection coverage	no data	

8.1.2 On the Statewide Publication of Serials on CDROMs

One of the stated goals of the Virginia Disc 2 project is to make MARC serial records available on a CDROM so that duplication of effort can be avoided among state supported libraries. There may be other impediments (legal or organizational) but from the technical viewpoint a creation of a union catalog of serials is possible on a CDROM. For developers who would be attempting to merge MARC records from diverse sources, they may encounter problems such as:

- detecting duplicate MARC records,
- combining records from different utilities. i.e. RLIN MARC, OCLC MARC, and
- merging holdings information. The MARC holdings format is still a proposal and many facilities use extra fields. For instance, holdings information can be found in the OCLC 045 field and the RLIN 9xx fields [Bills and Helgerson, 1989b].

8.2 FUTURE WORK

A more comprehensive comparison of PL and Micro-VTLS might be done in a few months. Such a study could integrate at least six criteria identified by Salton and McGill [1983] and discussed above.

Other interesting extensions about MARC records exist. Bibliographic publications on CDROMs have been commercially available since 1986. Much work has been done with publishing the MARC bibliographic records. A very effective retrieval system should result if a facility should have access to a large MARC authorities database like the Library of Congress MARC authorities database. Using an authorities database would enhance the match between the user's vocabulary and the cataloger's controlled vocabulary thus

eliminating users' frustrations when words have alternative spellings (e.g. British vs. American English) or when names of persons have variant spellings. The publisher could include cross reference authority records in order to broaden the possible ways to address an authoritative term. Consider the following hierarchical organization :

Household appliances

Domestic appliances (see from Household appliances)

Home appliances (see from Household appliances)

Domestic appliances and Home appliances are subclasses of Household appliances. Household appliances can be taken as a broader term than Domestic appliances. Other linkages like the "see also from" reference can be used to broaden a query. MARC Authority records have all these relations in machine readable form today and a study of how to use the information prior to indexing a MARC bibliographic record would be extremely valuable.

REFERENCES

Armstrong, A. 1986. Premastering and Mastering. *CDROM Optical Publishing*. Microsoft Press. Redmond, Washington. pp.217-224.

Arneson, R.H. 1989. The VANILLA Network: Something is Better than Nothing. *Library Hi Tech*. Issue 25 Vol.6 No.5 pp 20-21.

Bates, M. 1977. System Meets User : Problems in Matching Subject Search Terms. *Information Processing and Management*. Vol.13 pp.367-368.

Blair, D. C. 1984. The Data-Document Distinction in Information Retrieval. *Communications of the ACM*. Vol 27 no.4, pp 369-373.

Blair, D. C., Maron, M.E. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Communications of the ACM*, Vol 28 no.3 pp 289-299.

Belkin, N.J., Oddy, R.N., Brooks, H.M. 1982. ASK for Information Retrieval : Part I Background and Theory. *Journal of Documentation*. Vol. 38 no.2 pp.61-71.

Bills, L,, Helgerson, L. 1989a. CDROM Catalog Production Products. *Library Hi Tech*. Issue 25 Vol.6 No.5 pp.67-72.

Bills, L., Helgerson, L. 1989b. CDROM Public Access Catalogs: Database Creation and Maintenance. *Library Hi Tech*. Issue 21 Vol.6 No.1 pp.67-86.

Bills, L., Helgerson, L. 1989c. User Interfaces for CDROM PACs. *Library Hi Tech*. Issue 22 Vol.6 No.2 pp.73-115.

Bookstein, A. 1985. Probability and Fuzzy-Set Applications to Information Retrieval. *ARIST*. Vol.20. editor: Williams, M.E. Knowledge Industry Publications,Inc. New York. pp.117-151.

Crawford, W. 1984. *MARC for Library Use: Understanding the USMARC Formats*. Knowledge Industry Publications. New York.

Crawford, W., Stovel, L , Bales, K .1986. *Bibliographic Displays in the Online Catalog*. Knowledge Industry Publications. New York.

Crawford, W. 1989. Standards, Innovation and Optical Media. *The Laserdisk Professional*. January, 1989. Vol.2 No.1 pp 31-37.

Davis, W.P., Stephen, P.F., Raithel, F.J. 1989. The Missouri Library Connection : Progress in Statewide Cooperation. *Library Hi Tech*. March, 1989.

Gorman, M., Winkler, P.W. 1978. *Anglo-American Cataloguing Rules*. American Library Association. Chicago, IL.

Hildreth, C.R. 1985. Online Public Access Catalogs. *ARIST*. Vol.20 pp.233-285. Knowledge Industry Publications, Inc. White Plains, NY.

Library of Congress.1976. *Authorities : A MARC Format.* Marc Development Office. Washington DC.

Library of Congress. 1984. *USMARC Format for Holdings and Locations.* Network Development and MARC Standards Office. Washington DC.

Lynch, C.A. 1987. Standard Issues for Optical Publishing. *Bulletin of the American Society for Information Science.* Vol.13 No.6 pp.27-29.Aug/Sep, 1987.

Markley, K., Calhoun, K. 1987. Unique Words Contributed by MARC Records with Summary and/or Contents Notes. *Proceedings of the 50th ASIS Annual Meeting.* Boston, Massachusettes. Vol. 24 pp.153-162.

Maron, M.E. 1977. On Indexing, Retrieval and the Meaning of About. *JASIS.* Vol.28 No.1 pp.38-43. January, 1977.

Martin, B. 1987. The First CDROM Publication. *CDROM Optical Publishing.* Microsoft Press. Seattle. pp.269-282.

McGrath, D.H., Lee, C.R. 1989. The Virginia Tech Library System (VTLS). *Library Hi Tech.* Issue 25. Vol.6 No.5 pp.67-72.

Minker, J. 1977. Information Storage and Retrieval. A Survey of Functional Description. *ACM SIGIR Forum.* Vol.12 no.2 pp.1-108.

Milne, D. 1989. Personal conversation and consultation about the Micro-VTLS system.

Noreault, T., Koll, M. McGill, M. 1977. Automatic Ranked Output from Boolean Searches in SIRE. *JASIS.* November, 1977.

Norstedt, M. 1989. Personal interview about CDROMs and library practices.

Personal Library Software,Inc. 1987. *PERSONAL LIBRARIAN The Information and Knowledge Management Tool.* Personal Library Software. Rockville, MD.

Radecki, T. 1988. Trends in Research on Information Retrieval - The Potential for Improvements in Conventional Boolean Retrieval Systems. Information Processing and Management. Vol.24 no.3 pp 219-227.

Salton, G., McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. New York.

Salton, G. 1986.Another look at automatic text-retrieval systems, *Communications of the ACM.* Vol 29 no.7 pp.648-656.

Shera, J. H. 1980. Philosophy of Librarianship in. *ALA World Encyclopedia of Library and Information Science.* editor: Wedgeworth, R. pp.314-317. American Library Association, Chicago.

Smit, P.M., Kochen, M. 1988. Information Impediments to Innovation of On-line Database Vendors. *Information Processing and Management.* Vol.24 no.3 pp.229-241.

Steinberg, D., Metz, P. 1984. User Response to and Knowledge about an Online Catalog. *College & Research Libraries.* Vol.45 no.1 pp.66-70. January,1984.

Svenonius, E. 1986. Unanswered Questions in the Design of Controlled Vocabularies. *JASIS.* Vol.37 no.5 pp.331-340.

VTLS, Inc. 1987. Micro--VTLS, Exceptional Automation for Smaller Collections. Sales brochure. Blacksburg, Virginia.

Waller, W. G., Kraft D. H. 1979. A Mathematical Model of a Weighted Boolean Retrieval System. *Information Processing and Management.* Vol.15 no.5 pp.235-245.

Wilson, P. 1978. Some Fundamental Concepts of Information Retrieval. *Drexel Library Quarterly.* Vol 14 no.2 pp.10-24. April, 1978.

APPENDIX A
USER INTERFACES

MICRO-VTLS USER INTERFACE

A transition state diagram is a visual representation of an interactive system's interface. The transition diagram shows how the user's commands affect the state of a retrieval system. Each state in Figure A-1 approximates one screen type for Micro-VTLS. The diagram is useful for showing the basis for comparing Micro-VTLS and Personal Librarian's response times. In the timing test, we measured the amount of time it takes to switch from the initial query screen to the screen displaying a single document .

We shall now track the interactions for a sample transaction from the query screen to the display of a MARC screen. Figure A-2 shows the Micro-VTLS Query screen. A query like 's/school', a subject search about schools, yields Figure A-3, the authority list screen. The user chooses an entry by using the up/down cursors coupled with the <return> to yield Figure A-4. A choice under the retrieved list screen displays the record of interest as in Figure A-5. Finally the MARC record format is displayed by typing an 'M' (see Figure A-6).

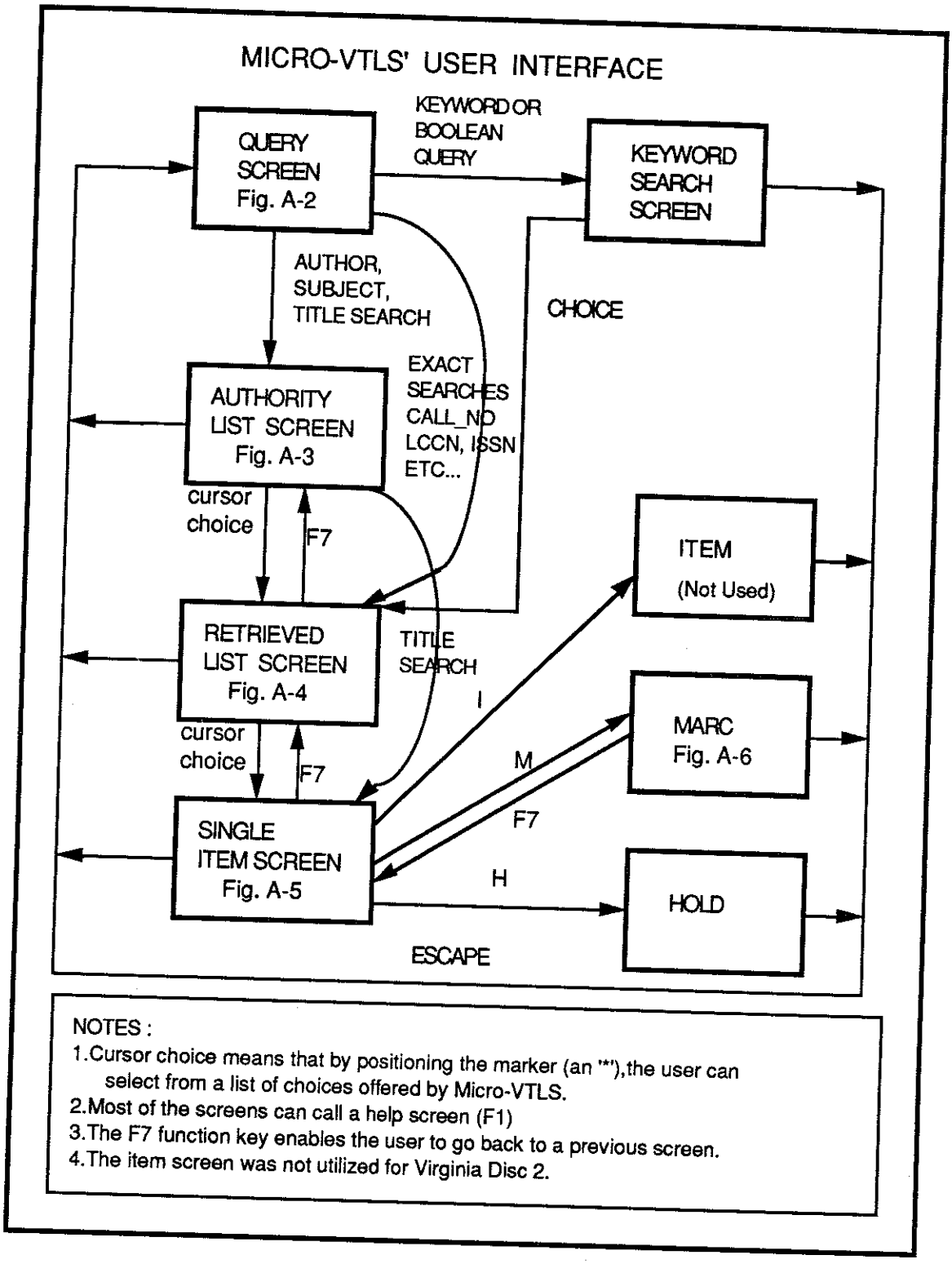


Figure A-1. A State Transition Diagram of the Micro-VTLS Interface

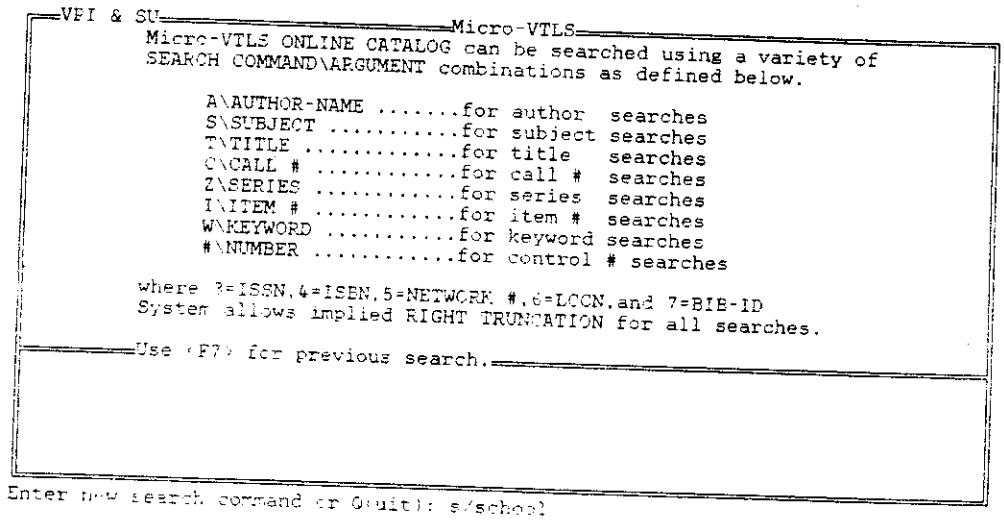


Figure A-2 Micro-VTLS Query Screen

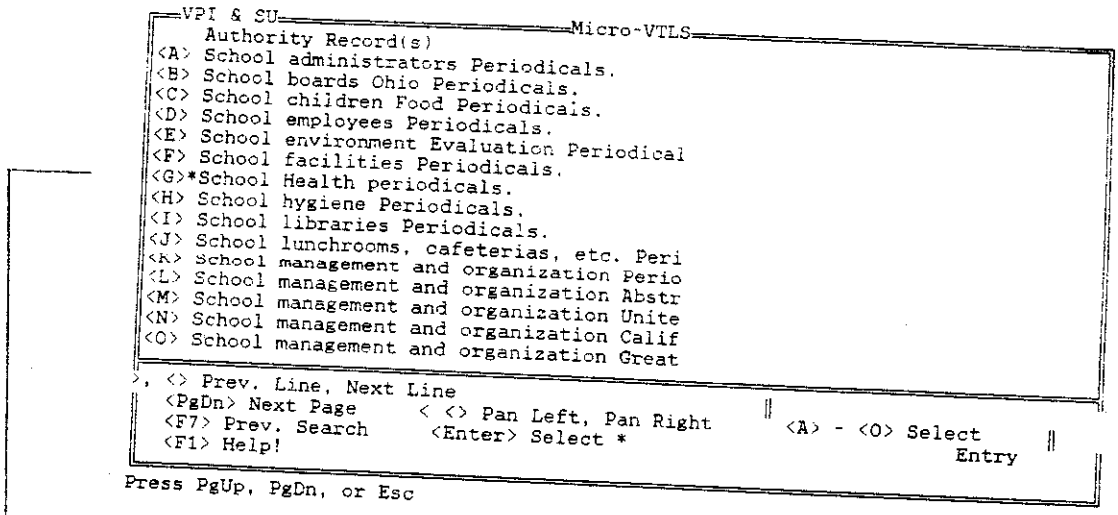


Figure A-3. Micro-VTLS Authority List Screen

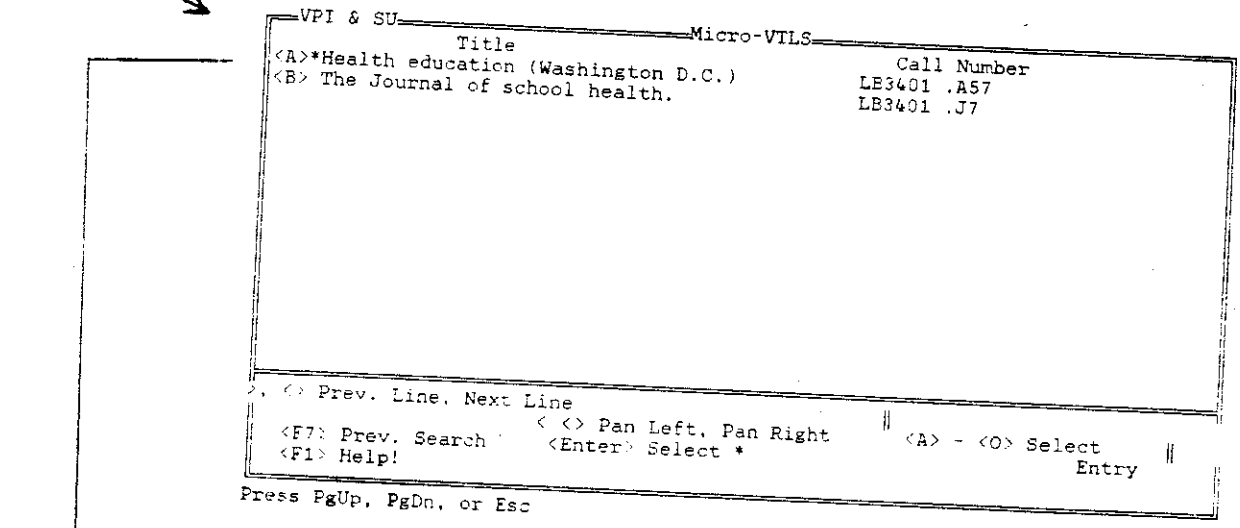


Figure A-4 Micro-VTLS Retrieved List Screen

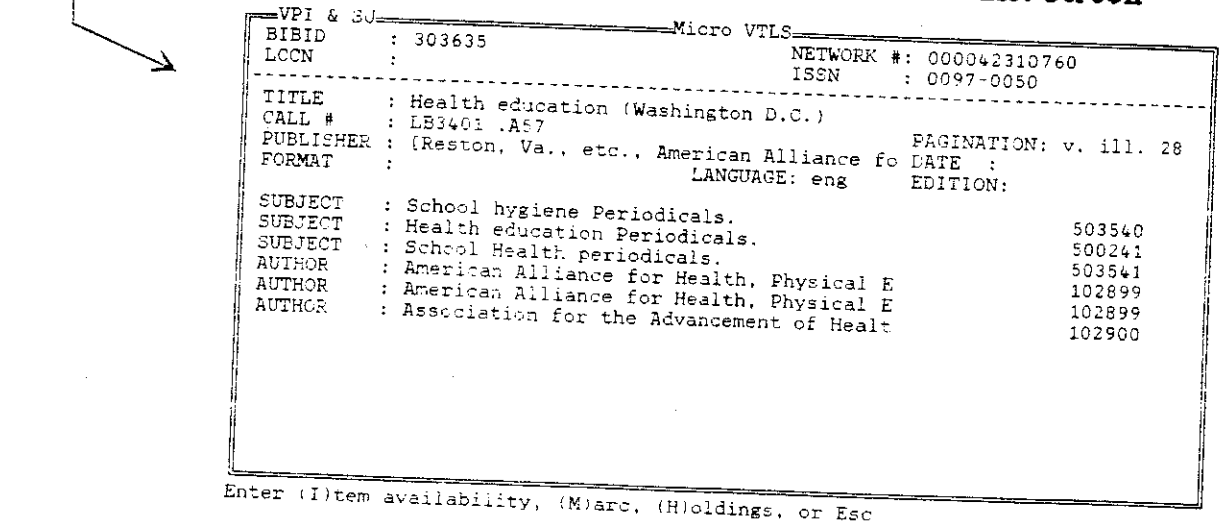


Figure A-5. Micro-VTLS Single Item Screen

MARC BIBLIOGRAPHIC SCREEN

```
16. 246 10 R. & H.D.
17. 246 10 R. and H.D.
18. 246 17 R & HD
19. 260 00 New York, N.Y. : \b Restaurant Business, Inc.,
20. 265    Restaurant Business, Inc., 633 Third Ave., New York, NY
           10017
21. 300    v. : \b ill. ; \c 28 cm.
22. 310    Frequency varies
23. 350    $24.00 (U.S.) \a $30.00 (Canada)
24. 500    "A Bill publication."
25. 500    Description based on: Vol. 5, no. 1 (Jan./Feb. 1983); title
           from cover.
26. 510 1  Trade & industry index \b 1983-
27. 650  0 Restaurants, lunchrooms, etc. \z United States \x
           Periodicals.
28. 650  0 Hotels, taverns, etc. \z United States \x Periodicals.
29. 780 00 \t Restaurant design
           \x 0191-345X \w (OCoLC)4853268
30. 850    CU \a DLC \a IaAS \a InLP \a MiU \a MoU
31. 901     \c Ser
32. 936    Jan./Feb. 1984

Press PgUp, PgDn, <F7>, or Esc
```

Figure A-6. Micro-VTLS MARC Screen

PERSONAL LIBRARIAN USER INTERFACE

We now trace the interaction for a query submitted to Personal Librarian for the Newman Library collection. The initial query screen could be any PL screen because PL can accept new queries from any screen state. PL can accept natural language queries. The word query 'electromagnetism' yields Figure A-8. Pressing the down arrow then chooses the top entry from the relevant list and presents the record as Figure A-9. The special display facility, written by the author, could thereafter be summoned with a 'z' <return>. Figure A-10 shows the resulting bibliographic screen. An 'H' brings the user to the Holdings Screen.(Figure A-11). Two other formatted display screens can be reached with an 'M' and a 'J' for the MARC bibliographic screen and the MARC holdings screen respectively. Figure A-12 shows a MARC bibliographic screen and Figure A-13 displays a MARC holdings screen.

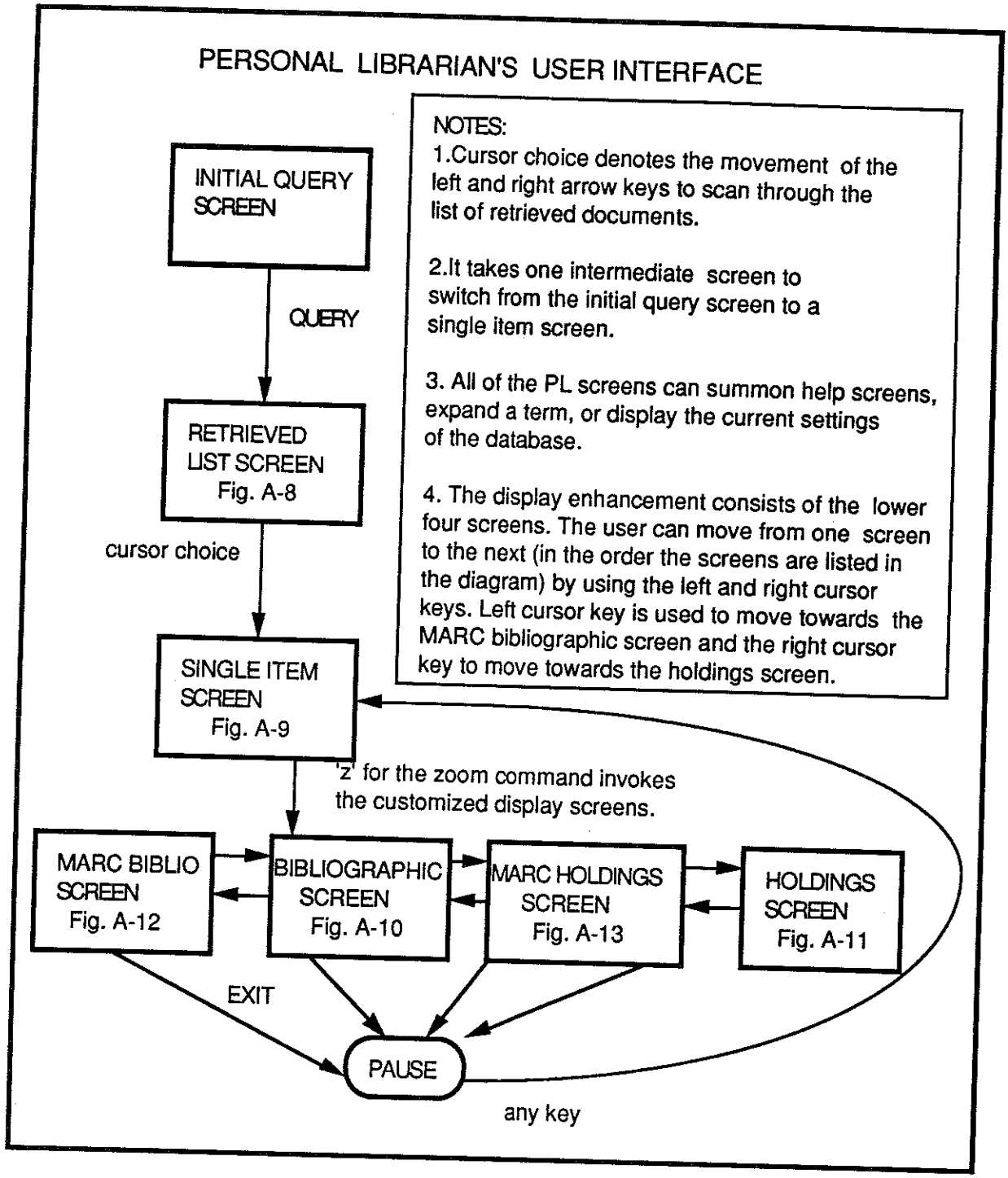


Figure A-7. A Transition State Diagram of the Personal Librarian Interface
The last 4 screens at the bottom comprise the display extensions made for Personal Librarian.

```
get Display Jump > < Quit Help Set List Bar Alpha Expand Past SOrt EDit Write
  Rank
   1) CALL_NO:  TK6553 .I15
      TITLE:  IEEE transactions on electromagnetic compatibility.
   2) CALL_NO:  QC759.6 .E4
      TITLE:  Electromagnetics.
   3) CALL_NO:  TK7800 .J686
      TITLE:  The Journal of microwave power and electromagnetic energy :
       a publication of the International Microwave Power Institute.
   4) CALL_NO:  QP82.2.E43 B53
      TITLE:  Bioelectromagnetics.


Query 1        Retrieved 4
Enter command>
ELECTROMAGNETISM
```

Figure A-8. Personal Librarian Retrieved List Screen

```
get Display Jump > < Quit Help Set List Bar Alpha Expand Past SOrt EDit Write
-CALL_NO-
 QC759.6 .E4
-TITLE-
 Electromagnetics.
-IMPRINT-
 Washington, D.C. : Hemisphere Pub. Corp., c1981-
-ADDRESS-
 Hemisphere Pub. Corp., 1025 Vermont Ave., N.W., Washington,
 D.C., 20005
-PAGINATION-
 v. : ill. ; 26 cm.
-FREQUENCY-
 Quarterly
-SUMMARY-
 Title from cover.
-NOTE-
 Published in cooperation with the Electromagnetics Society.
-SUBJECT-
 ELECTROMAGNETISM   PERIODICALS.
Hit Enter for more
Query 1        Retrieved 4                    QC759.6 .E4
Enter command>                                Doc. # 8808      Rank 2
ELECTROMAGNETISM
```

Figure A-9. Personal Librarian Single Item Screen

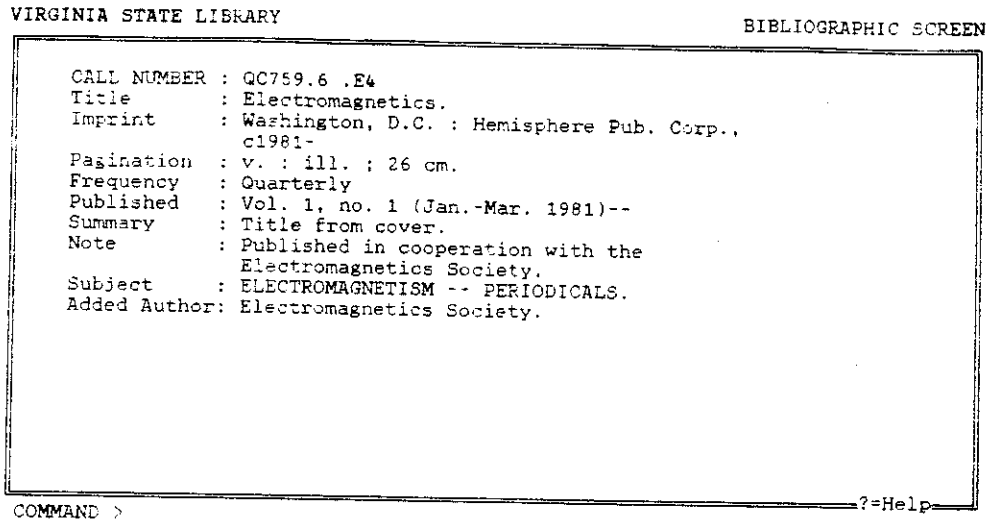


Figure A-10. Display Enhancement Bibliographic Screen

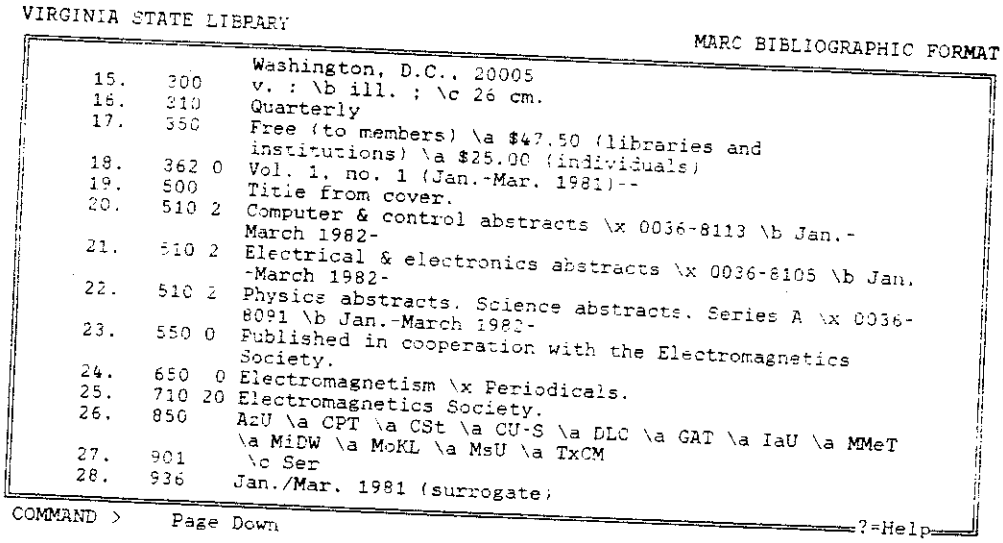


Figure A-11. Display Enhancement MARC Bibliographic Screen

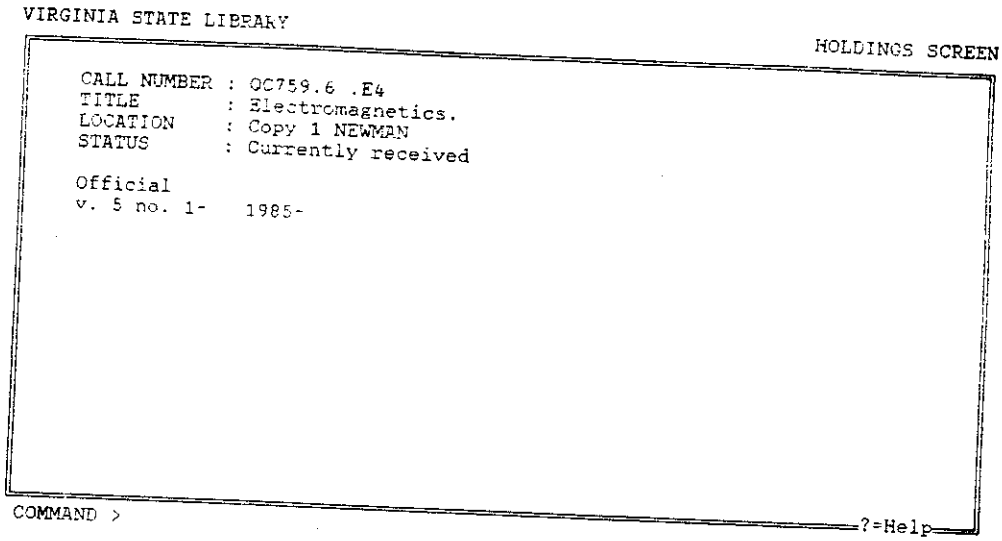


Figure A-12. Display Enhancement Holdings Screen

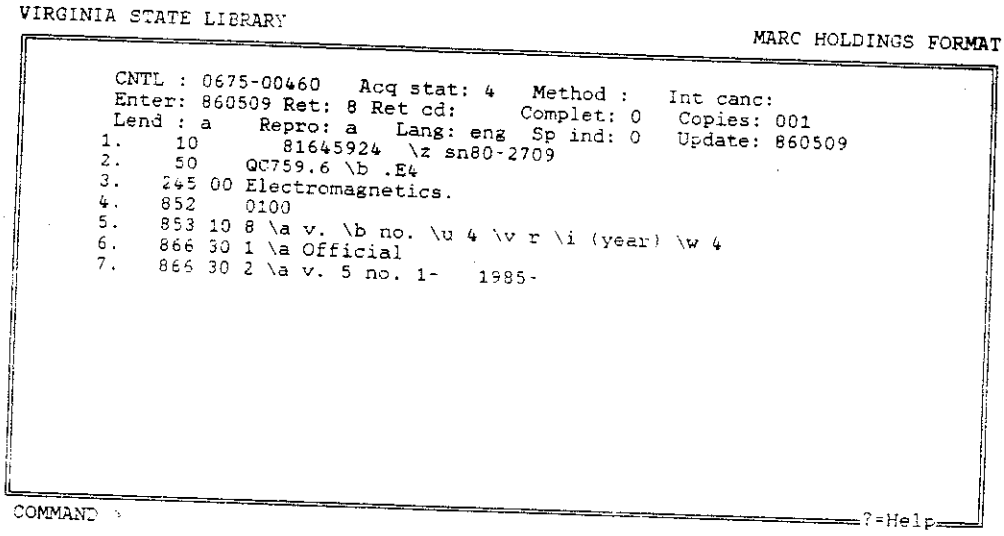


Figure A-13. Display Enhancement MARC Holdings Screen

APPENDIX B

PARTICIPANTS IN THE VIRGINIA DISC 2 PROJECT

Participant	Contribution
1. Newman Library	Contributed 10084 bibliographic records of Newman Library's serial holdings
2. Virginia State Library and Archives	Contributed 5598 bibliographic records about Virginia's history
3. Virginia Tech Library System, Inc.	Prepared Micro-VTLS version of two OPACs
4. VPI&SU computer group	Prepared PL version of OPAC and integrated CDROM
5. AT&T	Premastering services
6. Nimbus Information Systems	Mastering and replication, grant funding
7. Virginia Center for Innovative Technology	Grant funding
8. Personal Librarian Software	Allowed use of PL for database indexing and retrieval