1987

Indexing for the online catalog

Arthur B. Chitty
CUNY Queens College

Indexing for the Online Catalog

A. B. Chitty

The proliferation of online public access catalogs (OPACs) requires some systematic rationale for the comparative evaluation of their designs. Considered as an indexing application, the OPAC can be analyzed by three features: the varieties of bibliographic data processed, the kinds of indexes constructed, and the ways in which the indexes are searched. No one configuration applies to every library research project with equal efficacy or likelihood of satisfying queries. However, the rationale proposed can compare and evaluate alternative library computer catalogs in terms of the library's understanding of the relationship between the library's collections and their use.

The online public access catalog (OPAC) is fundamentally an indexing application, yet the sheer novelty of online access to machine-readable records of library holdings often masks important issues in developing and implementing this application. The various OPACs developed over the last few years exhibit different approaches to providing ready and effective online access to library holdings, and the range itself suggests some of the confusion and miscomprehension involved in this aspect of library systems development.

Rooted in the unexamined contradictions of historical development, this confusion hinders the construction of any rationale more substantial than a checklist of features for the comparative evaluation of online public access catalogs. While much effort has been invested in the theory and practice both of indexing and of the manipulation of machine-readable bibliographic records, the online public access catalog itself has been a haphazard, even opportunistic, though not irrational, development. Commercially driven development, answering as it must to the vendor's perception of market acceptability (and the vendor's cash flow), is mostly a recursive process, proposing and disposing as ideas are implemented, sent out into the marketplace, and returned for review and revision. Considering the central role of collecting, cataloging, and circulating books in the library's organization and work and confronted with the proliferation of paperless media in the distribution of information, the failure to address this lack of rationale is at least risky and potentially subversive of the effectiveness of the library's most precious (at least most costly) possession, its catalog. This paper proposes a rationale for the comparative evaluation of the effectiveness of an online catalog's design for the retrieval of citations relevant to the patron's query from a MARC database.

HISTORICAL ISSUES

The historical dimension of the confusion is an irony of the chronology of library automation. The first great automation project was the machine-readable cataloging (MARC) effort initiated by the Library of Congress more than two decades ago. Its crowning achievement was the standard LC/MARC format for bibliographic data, the basis of the great bibliographic utilities, OCLC, RLIN, and WLN. Even as this project came into final flowering with the

A. B. Chitty is Library Systems Planner, Queens College of the City of New York.

elaboration of LC/MARC formats for nonbook materials (most recently the completion in 1986 of standards for serial holdings), its functional context shifted. The LC/MARC format is a communications format; it provides standards for the communication, capture, and recognition of bibliographic data. The shift follows from a change in the "recipient" of the communication, defined on both machine and human levels.

The original machine recipient of this communication was a catalog card printer; the target is now the online bibliographic database of local library holdings to which the OPAC provides access. In the context of OPAC development, many LC/MARC features are irrelevant or even misleading: in retrospect the LC/MARC format is littered with printing and pseudo-printing instructions. The paradigmatic example is the "main entry," coded in the LC/MARC 1XX tag series, a cataloging concept that has evolved from its origins as the basis for ordering entries in catalogues raisonnees, through its use as a means of controlling reproduction costs in book and card catalog, to its present position as a source of considerable complexity (and confusion and heartbreak) in the rules promulgated by the second edition of the *Anglo-American Cataloguing Rules (AACR2)*, with the attending controversies over imposition, superimposition, and finally desuperimposition.

The communication's human recipient, the "user," has changed more profoundly, from using machine-readable bibliographic records to support copy cataloging to using them in library reference work as sources of topical citations. In copy cataloging, the user of the bibliographic database wants to know if a title actually inhand at the terminal has already been cataloged. The online record of libraries' participation created in OCLC, RLIN, and the large, shared circulation systems permitted an extension of this use of the database, the online union catalog (OLUC) of holdings. This utilitarian extension has been a vast, profound enhancement of the librarian's reach, but the improvement is quantitative rather than qualitative: the searcher must still have previously identified the citation sought. In front of the reference desk or the catalog, the would-be user of the database usually does not yet know the relevant titles and certainly does not yet have them at hand. The answer demanded from the database has changed from "yes/no" to "try this, and this, and this."

Despite the shift, the development of online catalogs of local library holdings rests on the LC/MARC format for bibliographic data. The economies of copy cataloging are too deeply embedded in library funding and budgets, and the utility of online access to large collections of holdings of known items is too important to library service for the MARC format to become obsolete. The question for OPAC development is not how the format might be improved but how best to use it. Indexing is the key to its use.

Indexing cannot be considered in isolation. The utility of any indexing technique depends both on the way the index manipulates the data and how the index is searched.

BIBLIOGRAPHIC DATA

The LC/MARC record contains various data encoded in various ways: the record identification and control data, the identification of the bibliographic entity, printing instructions, and a variety of data elements designed to allow the bibliographic record or the item to which it refers to be juxtaposed to records or items similar in some significant feature. These "relational" data elements are "headings" or "access points," so-called from their use in filing and finding cards or entries in a catalog or list. Headings are identified in the LC/MARC record format by type: the LC/MARC three-digit tags specify who was involved in the production of the text cited by the record, the various "names" (titles) given to the text, and what the text is about.

The headings data in the LC/MARC record, aside from control codes of various kinds, can be analyzed according to their content and arrangement. Some data are essentially arbitrary numeric codes; others are textual. Some are composed of "free" elements taken from the material described; others are assigned from controlled vocabularies. Some are arranged in more-or-less logical hierarchies; others follow natural language grammar. Any heading will fall into one category or the other along each of

these three dimensions, though not all headings are equally pure examples of the category. Decimal classifications are purely numeric, but alphanumeric classifications like Library of Congress use letters mainly as extensions of the cardinal decimal numbers. The alphabetic logic of subject (and institutional author) heading arrangement is conventional, though the subdivisions can be either grammatical or hierarchical, depending on whether or not the subfield code is processed as part of the text. Since these subfield codes specify certain logical relationships between the subdivisions and the head term—chronological, geographical, and the like—they permit the deployment of some filing or display order other than strictly alphabetic. The "grammar" of personal authors is highly conventional (witness the variations in filing rules according to the "nationality" of the name). Uniform titles are an especially complex hybrid.

The facility with which LC/MARC bibliographic data is manipulated can be summed up in whether the indexing process recognizes the various demarcations and codes within the data—fields, subfields, indicators—and how the codes are processed. Not all logically potential categories of access points have members (hierarchical free text is hard to imagine), nor are all LC/MARC access points relevant to the OPAC (ISNs and CODENs, for example). LC/MARC data elements other than traditional library catalog headings can be topically significant, serving as access points to bibliographic citations, as parameters for refining a retrieved set of citations, or as alternative index structures accessing the database. In contrast to the "precoordinated" controlled vocabulary headings, contents and abstracts can provide "postcoordinated" vocabularies. For topical indexing purposes (as opposed to bibliographic identification), these elements resembles titles. Format, imprint, date, edition or issue, and language data likewise can be topically significant, though rarely at the initial headings level. In the case of two types of bibliographic data elements, a new possibility is created by the use of the LC/MARC record as input to an OPAC. In both classifications (at the 0XX tag level) and in subjects (at the 6XX second indicator level), the MARC record preserves an option not readily available to the library that relies on a card or printed catalog: the creation of multiple alternative catalogs. The OPAC can provide a Dewey-classed catalog of a collection shelved in LC class order by building an 082 index. It can provide a MeSH (650–2) index to an LSCH collection. At a rudimentary level, this is accomplished merely by annotating headings from different sources with a print constant derived from the indicator. More sophisticated approaches involve segregating the indexes, with perhaps the ability to transfer among them at cross-referenced points, and in the case of alphanumeric indexes such as classifications, providing an indexed textual description of the subject.

INDEX STRUCTURES

There are two basic types of indexes to bibliographic records, and a third that

Table 1. Taxonomy of MARC Data Elements

Classifications (ABC)	Numeric, Controlled, Hierarchical
Institutional authors (aBC)	Textual, Controlled, Hierarchical
Subject headings (aBC)	Textual, Controlled, Hierarchical
Uniform titles (aBc)	Textual, Controlled, Grammatical
Personal authors (abc)	Textual, Free, Grammatical
Titles (abc)	Textual, Free, Grammatical
Series (abc)	Textual, Free, Grammatical
Contents (abc)	Textual, Free, Grammatical
Abstracts (abc)	Textual, Free, Grammatical

Key: A = Numeric B = Controlled C = Hierarchical
a = Textual b = Free c = Grammatical

NOTE: *Controlled* refers not to the use of authority to control variations in appearance, as in authors and series, but to the control of the vocabulary itself, as in subjects and classifications.

Table 2. Taxonomy of Index Structures

Structure	Access points	Context
Heading	Few	Present
Permuted	Many	Present
Keyword	Many	Absent

combines elements of both: heading, keyword, and permuted. The three vary in the ways in which they combine the number of entries (access points) generated with the extent to which the entry preserves the context from which it was extracted. The heading index closely mimics the card or book catalog: the bibliographic data element is treated as a single character string and filed once, in alphanumeric ASCII or ALA filing order left to right. The keyword index treats each identifiable component of the bibliographic data element as a single unit, filing an index term for each component identified, usually word-by-word as identified by blank spaces or field delimiters. The permuted index, like the keyword index, files an entry for each identifiable "word" but, like the heading index, preserves and displays the context in which the word has been found. Keyword indexes are also called KWOCs (KeyWord Out of Context); permuted indexes are also called KWICs (Key Word In Context) or "rotated headings." The permuted index is in principle the richest of the three structures.

SEARCH TECHNIQUES

Online searching can use two techniques: browsing and selection, and any particular search may combine both in some sequence. Browsing is quite familiar; selection is less well understood, and even the terminology is unsettled.

Browsing permits the searcher to review an alphabetically or logically ordered portion of an index. Selection permits the searcher to review a set of citations that have common features. The techniques differ chiefly in the demands they make of the user. Browsing is "error tolerant"; that is, rigid keystroke and lexical accuracy are less crucial to the search's success, while the success or failure of a selection search depends on both mechanical and conceptual accuracy. Browsing takes more time but requires less forethought, an equation reversed in selection. (Browsing, of course, may be used to aid forethought, and selection may be used to speed up browsing.) In general, browsing produces fuller retrievals, while selection retrievals are more narrowly relevant. (Of course, relevance and comprehension are also directly determined by the index vocabulary and structure.) On the other hand, selection can use some sophisticated methods to refine the retrieval, such as Boolean combinations of features.

Table 3. Taxonomy of Search Techniques

Feature	Technique	
	Browsing	Selection
Error tolerance	High	Low
Sophistication	Low	High
Retrieval relevance	Low	High
Comprehensiveness	High	Low

ASSESSMENT

While searching techniques are the most visible elements of the entire indexing configuration, their effectiveness directly follows from the way in which the technique takes advantage of the index's structuring of the bibliographic data. Not every combination of search technique, index structure, and bibliographic data responds equally to every kind of query or every kind of need for library information, nor does each combination manipulate with equal effect the bibliographic data available in the database derived from LC/MARC format records.

Some general principles can, however, be proposed. To the extent that they permit different approaches to online catalog design to be laid out side by side, they can be used to construct a comparative evaluation of various OPAC structures. Of course, no library is exactly like any other library, either in actual operation or in ambition, and there remain many other features of an online catalog that enhance or detract from its overall utility. Still, the flexibility of any particular OPAC in manipulating various kinds of bibliographic data into a responsive structure to be used by a technique appropriate to the query must be a crucial consideration.

Consider the various kinds of biblio-

graphic data. Numeric data are more arbitrarily coded than textual data and can be juxtaposed more comprehensively in logical arrangements than an alphanumeric order of controlled or free natural language terms can achieve. Likewise hierarchically structured terms permit more orderly arrangements than grammatical terms. Such arrangements are designed to be browsed and support more comprehensive retrievals. On the other hand, the effective assignment of controlled vocabulary terms of whatever content and structure to a text depends ultimately on the sophistication and finesse of the cataloger, while free or natural language text indexing permits the searcher more direct access to the terminology applied to the text by its author and undoubtedly follows trends in specialized terminology and technical jargon more closely than any controlled vocabulary. It does, of course, risk eliminating from the searcher's consideration those other citations accessible only through the intervention of a trained cataloger assigning standardized terminology to the text's bibliographic description.

Indexing structures are easier to assess. The heading index, for all its long service in card and book catalogs, is extremely limited in utility except as it juxtaposes terms that have been hierarchially structured. Its best use is probably in classification and shelflist arrangements. The keyword index is likewise limited except as it exposes terms embedded in larger terms whether structured logically in a hierarchy or naturally according to grammatical conventions. Its best use is probably in free text indexes. The permuted index combines the strengths of heading and keyword indexing but is not uniformly useful, or even feasible, for all kinds of bibliographic data.

Searching techniques differ in their ability to utilize various combinations of index structure and bibliographic data. Browsing is close to pointless in a free-text keyword index except as a memory aid or a way to eliminate irrelevant homographs. Selection can be applied to controlled term heading indexes, but the results are quite haphazard unless a fully automatic and well-developed cross-reference structure or thesaurus underlies the index. Both search techniques work well enough in a permuted index, but browsing is more effective if the index permutes hierarchical terms, and selection does not use the permutation's preserved context to advantage unless Boolean combinations are used. The various possible configurations begin to make sense in the context of a particular library's collections, services, and patrons. Selection from a free-text keyword index provides a quick search for any relevant item. Browsing a controlled term permuted index provides a systematic review of library resources on a topic.

Subjects and titles pose particular problems. The extent of a keyword selection can be easily expanded by including controlled-term keywords from subject headings in the index, though the dissolution of controlled-term hierarchies may be a high price to pay for the convenience. Likewise title keywords are useful topical descriptors, but the value of title browsing cannot easily be dismissed: permuted titles are undoubtedly a convenience for the searcher who cannot quite remember the name of the text desired (a case of "almost known item" searching). Collections for which LC/MARC access data is quite inadequate might also be taken into account. Materials using term descriptors, such as ERIC, need keyword indexes made of controlled terms and searched by Boolean selection, while other materials using concatenated headings, such as *Art and Architecture* terms, require permuted indexes searched by browsing. The effectiveness of any particular configuration of bibliographic data manipulation, index structure, and search technique depends on the patron's information and patience and the relative importance of relevance versus comprehensiveness in the retrieval's satisfaction of the query. These vary from library to library and indeed from patron to patron within a library and from time to time for any patron. However, the fundamental issue for the design of the OPAC is whether the search exploits the index to expose the appropriate bibliographic data to retrieval and, we hope, to intellectual use.

VARIETIES OF OPACS

Current OPACs show considerable vari-

Table 4. Summary Comparison of Selected OPACs

	ATLAS	BLIS	DOBIS	GLIS	INLEX	LIBS	LS2K	NOTIS	PALS	TOMUS
Bibliographic data manipulation										
Fields	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Plus Subfields	(1)	YES	YES	(1)	NO	(1)	(1)	(2)	(1)	NO
Plus Indicators	NO	(3)	YES	(3)	(3)	NO	(3)	(3)	NO	NO
Indexes										
Heading	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO
Keyword	YES	YES	YES	YES	NO	YES	YES	NO	YES	YES
Permuted	NO	YES	YES	NO	NO	NO	NO	NO	NO	NO
Searches										
Browsing	YES	YES	YES	YES	YES	YES	YES	(4)	NO	NO
Selection	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES

NOTES: (1) Use of the keyword index gives, in effect, access to subfields, though browsing is difficult. (2) Since NOTIS customers can specify and program how bibliographic records are to be loaded, within the heading logically ordered subdivisions can in theory be created. (3) Source indicators in 6XX tagged fields can be used as print constants to distinguish different kinds of terms. (4) Selections are presented to be browsed.

ety in their indexing configurations: the summary chart reflects the state of the art circa December 1986. The selection is somewhat fortuitous: these systems were proposed to Queens College of the City University of New York during 1985 and 1986. At least one (BLIS) is no longer readily available, and several others (IBM's DOBIS, Geac's GLIS, Sperry's PALS) have not made much market headway since 1985.

Even in this arbitrary selection, there are substantial variations. INLEX and TOMUS are mirror images. Only INLEX browses headings only, in effect mimicking the card catalog. Only TOMUS selects keywords only, the most radical departure from the card catalog. NOTIS is much closer to INLEX than to TOMUS: although local or bespoke programming can extend the range within which NOTIS manipulates MARC data, and although the initial entry into a NOTIS index is in fact selection of a key term, without either permuted or keyword indexes and fully realized selection searching, the interior of the heading remains accessible only indirectly and serendipitously. PALS is much closer to TOMUS than to INLEX, though unlike TOMUS it preserves the headings from which the keywords can be selected.

A keyword selection search, provided as an alternative to a headings browsing search, does give direct access to the interior terms of the headings in several systems (DRA's ATLAS, Geac's GLIS, CLSI's LIBS, OCLC's LS-2000, Sperry's PALS). At least one (LIBS) permits browsing near-homographs selected by a truncated keyword. However, the logical and hierarchical relationships are lost in the keyword selection search.

BLIS and DOBIS cover the range of features, though each has its problems. BLIS had developed only the very rudimentary "print constant" technique of marking variant topical terminological systems from the 6XX second indicator contents. DOBIS has not yet developed a uniform solution to filing order for permuted terms, though several solutions are implicit in the various DOBIS/MARC processing options.

OPAC development tends toward more rather than fewer options. Within certain limits, the marketplace demands it, and the history of CLSI's OPACs illustrates the direction. CLSI's first OPAC, the SEARCH process, imposed a selection technique on a headings index. Their second OPAC, then called PAC, browsed headings. Their third and current OPAC, once called PAC II, now CL-CAT, selects keywords or browses

headings. With the addition of a print constant distinguishing among various kinds of subject authority systems, this configuration is likely to become standard. Even NOTIS and INLEX—the two least fully developed OPACs—will surely soon catch up, if development capital continues to be available to them.

However, the major variations coincide with machine differences. The radical departures are based not on minicomputers, but on cascaded microcomputers (TOMUS) or mainframes (BLIS and DOBIS). Since machine capacity (in both memory and storage) is only a momentary constraint on computer system development, this standard is likely also to be momentary. So far as the use of the LC/MARC bibliographic record for the local OPAC goes, the permuted index structure and segregated topical indexes glimmer fitfully on the near horizon.

APPENDIX A. GLOSSARY

1XX tag: In the LC/MARC record, the main entry tagged fields.

AACR2: *Anglo American Cataloguing Rules*, 2d edition.

Access point: In a catalog record, text that serves as a filing element.

ASCII: American Standard Code for Information Interchange, a seven-bit-plus-parity code established by the American National Standards Institute (ANSI) to achieve compatibility between data services.

ATLAS: A Total Library Automation System, developed by Data Research Associates (DRA).

Biblio-Techniques: A software-only vendor with connections to WLN. The Biblio-Techniques system runs on IBM and lookalike machines, using ADABAS and COMPLETE system software from Software AG.

BLIS: Bibliotechniques Library Information System.

Boolean: George Boole (1815–64) developed the use of algebraic notation to express logical relationships, where the variables stand for statements and the connectives refer to the logical operations used in truth tables. Since all Boolean statements can be reduced to a binary number, Boolean logic is peculiarly adapted to computer processing. In library automation, a tiny part of the theory of Boolean algebra is used to define and modify by combination/exclusion sets of citations having common features.

Browsing: In OPAC searching, reviewing entries in an index forward and backward from a starting point.

Carlyle: Developer of TOMUS, an online catalog mounted on cascaded, dedicated microcomputers. The system and vendor have roots in the University of California CLASS project.

CL Systems: The largest of the turnkey local library systems vendors; located in Newtonville, Massachusetts. *CL* originally stood for Computer Library. The library automation system is called the LIBS-100 and runs under CLSI's proprietary operating system, FLIRT, on DEC PDP machines.

CL-CAT: The current name for the OPAC application in CLSI's LIBS system.

CLSI: C-L Systems, Incorporated.

DOBIS: DOrtmunder BIbliotheks-System, the library automation system developed by IBM Corporation for the University of Dortmund and enhanced at the University of Leuven. Also known as DOBIS/LIBIS (Leuven Integraal BIbliotheek System) or as DOBIS/Leuven.

DRA: Data Research Associates, a library automation vendor in Saint Louis, Missouri, long known primarily for library automation systems designed for the blind and physically handicapped. DRA's system is called ATLAS and runs on DEC VAX machines under VAX/VMS.

Free text: In information retrieval, unformatted or "natural language" text.

Geac Computers: A Canadian computer services vendor specializing in banking and library automation. *Geac* has no particular meaning. The library system is called GLIS (Geac Library Information System).

Heading: In library usage, a string of letters and numbers used as a filing element for a bibliographic entry.

Inlex: A software house in Monterey, California; developers of the INLEX/3000 library automation system. Inlex used to be called Electric Memory (EMI), has developed other applications besides library automation, and now uses mainly Hewlett-Packard HP/3000 machines and operating system software.

Keyword index: In library automation, an index with entries derived from individual words identified by machine in free or formatted text.

KWIC: Key-Word In Context, a keyword index that preserves, uses, and displays the words surrounding the individual word that has become an index entry.

KWOC: Key Word Out of Context, a keyword index that preserves, uses, and displays only the individual word that has become an index entry.

LC/MARC: Library of Congress Machine-Readable Cataloging, a standard format for communicating cataloging copy.

LIBS: "LIBrary System," CLSI's library automation system.

LS-2000: The Local System for the year 2000, developed by OCLC Local Systems. Runs on Data General equipment provided and supported by OCLC in a turnkey package under the MIIS/MUMPS operating system.

Main entry: In cataloging, the heading under which the fullest bibliographic description of the title is filed. The concept originally developed as the proper name of a text (usually literary), under which versions, editions, adaptations, and commentaries might be collated for convenient reference. Scholarly bibliographical description sometimes still uses the idea in its original sense.

MARC: Machine-Readable Cataloging. The original MARC project was undertaken by the Library of Congress, and MARC usually refers to cataloging copy recorded in the communications format developed by the Library of Congress and used by the major bibliographic utilities such as OCLC, RLIN, WLN, and Utlas. The main defects of LC/MARC follow from its history as a record for communicating instructions to a card set printer; its main virtues are its ubiquity and use as a standard copy cataloging format.

Natural language: Language spoken or written by real people, as opposed to computers and their minions.

NOTIS: NOrthwestern Total Information System, developed by Northwestern University in Evanston, Illinois, and now marketed as a software package for IBM and lookalike computer systems.

OCLC: Originally named the "Ohio Conference on Library Computing," OCLC is the largest of the online bibliographic utilities. OCLC Local Systems is OCLC's local library automation systems division, and OCLC's local system is called the LS-2000.

OPAC: Online Public Access Catalog.

PALS: The integrated system developed at Mankato State University and marketed by the Sperry Corporation.

Permuted index: A KWIC index.

Postcoordinated and precoordinated: In searching terminology, an index that embeds in its terms the logical relationships among descriptors is precoordinated. A retrieval system in which the searcher specifies any logical relationships is postcoordinated. Library catalog headings are precoordinated; a Boolean combination of free-text keywords is a postcoordinated search.

RLIN: Research Libraries Information Network, sponsored by the twenty-six-member Research Libraries Group (RLG), an OCLC clone which serves large academic research-oriented libraries.

Selection: In OPAC searching, retrieving a list of citations that are indexed by a specified term or combination of terms.

Sperry Corporation: A computer company known mainly for defense contracts, but also in the library automation market with the PALS system running on Univac machines.

TOMUS: The Online Multiple User System, developed by Carlyle; the most radical of the OPAC designs.

UTLAS: A Canadian bibliographic utility sometimes used by U.S. libraries.

WLN: Once the Washington Library Network, now the Western Library Network; the youngest, smallest, and most sophisticated of the big three bibliographic utilities. ■■