
Union Records and Dossiers: Extended Bibliographic Information Objects

Michael K. Buckland
<buckland@otlet.berkeley.edu>
Mark H. Butler
Barbara A. Norgard
Christian Plaunt

School of Library and Information Studies
University of California, Berkeley
Berkeley, CA 94720

ABSTRACT

The growing number and sophistication of online bibliographic and network based information systems is starting to blur the once clear boundaries that separated print documents. Two concepts emerge as a consequence of these developments, first the union record, an entity which combines multiple catalog records for a single bibliographic item into an extended information object; and second, an information dossier, a hypertext-like information object built by linking several distinct but related bibliographic entities.

KEYWORDS: union records, extended bibliographic objects, information dossier, bibliographic representation, hypertext

INTRODUCTION

The process of automation is a slow one that often leads to unforeseen consequences, some positive and some negative. Technology ordinarily first computerizes "current practice," sometimes awkwardly, and only later, after much thought and experiment, are the "advances" made. Online catalogs are a good example of this tendency. First generation online catalogs merely put the linear catalog cards in a box and made them nearly impossible to browse.
Buckland et al.

Second generation catalogs added boolean access, and emerging third generation systems are now breaking out of the linear presentation model (Buckland, 1993b). We are now beginning to see that online catalogs can be much more than automated card catalogs (Norgard, 1993).

The storage of text online has followed a similar path, slowly leading from 80-column punched cards to hypertext. Even hypertext, once confined to links between a group of files on a single computer, is now starting to proliferate across the network in the form of HTTP clients and servers (Nickerson, 1992). Just as the linear nature of the card catalog is giving way to new forms of presentation, the linear model of online text is being replaced by a new linked model (Ardo, 1992; Bertha, 1992). However, this linking has brought with it something else. It is no longer a simple matter to identify the extent of a document. Where does one document end and the next begin? What is the appropriate unit of measurement for online “documents”? What is the right way to measure the emerging new forms of information object space?

In this paper we explore these ideas in the context of online bibliographic information systems. First we look at how large catalogs might start to exploit overlapping bibliographic records for individual items in order to form union records, merging the various bibliographic records for individual documents to form a larger, more informative representation. We then look at how multiple, related items might be merged using similar techniques to form a different kind of extended information unit, an information dossier. And finally, we draw attention to some of the problems associated with such work.

THE ONLINE ENVIRONMENT

The online public access catalog, originally a prisoner of the walls of the library in which it did its work, is now the gateway to the Internet for many library users. It is no longer "books as far as the eye can see." Now it is "information as far as the network can reach" (Schwartz, 1992; Émtage, 1992).

Searching Multiple Databases

Libraries are continuing their trend toward offering their users access to greater numbers of databases under the umbrella of their online catalog. An example is the MELVYL system provided by the University of California system-wide administration which currently offers access to close to a dozen databases, including both its union catalog of monographs and a variety of abstracting and indexing databases. In addition, it offers pass-through access to a number of other resources over the Internet such as FirstSearch from OCLC, EUREKA from RLIN, DIALOG, US Forest Service information, and online catalogs worldwide.
The Internet allows users to access multiple databases in different online catalogs. It makes available a wide variety of bibliographic records, technical reports, hypertext, full-text, and government documents, information services, and numeric data, and the list grows daily. Commercial services such as DIALOG and OCLC are increasingly becoming more accessible over the Internet through which they, too, can be easily searched.

**Heterogeneity of Data**

Given the variety of data resources readily available on the Internet, it is now possible, using a workstation with Internet access, to search several of these resources at once, either in parallel or sequentially (Ardo, 1992; Berners, 1992). One of the obvious results of such searching is the heterogeneity in the bibliographic records one is likely to find. Unfortunately, there is no accepted standard for the format of the data and even the ubiquitous MARC records of library catalogs are unreliable. On systems where the user can, usually with some effort or by using "hidden" options, obtain the MARC record, it is presented by the system in such a way as to no longer be a MARC record, e.g. the fixed fields are tagged or converted to "human readable" forms, lines are wrapped, leaders are left off, and the field/subfield/record delimiters are converted to printable characters.

However, much of the bibliographic data in online catalogs are becoming available in tagged formats. Here again, however, the tags are inconsistent, often even between datasets in the same system.

Another form of heterogeneity arises from the fact that each catalog has, for each item held, its own version of a catalog record for that item, and that the various cataloging agencies are unlikely to have perfectly identical records for any item. That is, many of the records are duplicates -- either based on copy cataloging a MARC record or different records for the same object -- but the data in these records may be heterogeneous. The same object may have different values for the same attribute or differ in the attributes represented. A further confounding factor is the ever changing nature of catalog codes and the relative infrequency with which they are retroactively brought into conformance with the current standard.

In online catalogs such as MELVYL that provide access to multiple databases, records for the same object in different databases have different information. Some of these records provide very simple bibliographic data, others provide an extensive set of subject headings, still others may have an abstract, and some may even provide the full text.

This chaotic mix can be seen as an insurmountable obstacle if one wishes to enforce a one record per object relationship. But there is an alternative: this heterogeneous mass of data can instead be viewed as an asset to be exploited and taken advantage of to build a union record -- a record that
Buckland et al.

consolidates all these different bits and pieces of information into a single extended information entity.

THE UNION BIBLIOGRAPHIC RECORD

The problems of coping with heterogeneous data are being investigated within the database management community and among those interested in resource discovery on the Internet. One example within the database management community (Bright, 1992) is the concept of a multidatabase system. It gives users one common interface to multiple heterogeneous databases without adversely affecting the autonomy of any of the individual databases. The integration of these heterogeneous databases can be performed with either a global schema to integrate the underlying databases, or a multidatabase language system which uses query language tools to perform the integration.

The Netfind application on the Internet (Schwartz, 1992; Schwartz, 1993) exploits heterogeneous data about individuals, stored in publicly accessible locations, to dynamically create a union record. The user asks Netfind about an individual's e-mail address by providing the individual's name and their potential location. Netfind looks at information from a variety of Internet sources (WHOIS, finger, SMTP, DNS) to help users discover that individual's e-mail address. The concept of data mapping is central to Netfind's success (Bowman, 1993). Data mapping uses mapping protocols for collecting information and agreement protocols for correlating information. Mapping protocols collect information and reformat it. Agreement protocols allow resolution of conflicts between different information sources either by eliminating one set of data, or allowing multiple sets of conflicting data to co-exist within the emerging record. The result of a Netfind search is a record that is greater than the sum of its parts.

The union bibliographic record creates a more complete record by combining various pieces of bibliographic information about a particular item from multiple sources. Some online catalogs already combine a limited amount of data to form a more complete record. MELVYL consolidates data from multiple cataloging departments for their monograph records -- although this union record is not displayed unless a user specifically requests it. MELVYL also attaches library location data from its periodicals holdings database to the display of records from its abstracting and indexing databases.

In its simplest form, a union record can be built by simply combining all the information that can be obtained about a bibliographic object. That is, one can build up a union record that includes different subject headings, one or more abstracts, citations, table of contents, and, perhaps, even the full text.

As a simple example, consider a case where there are multiple records for the same object. In this case, the search begins in MELVYL's Current
Contents (CC) database\(^1\) since it has very broad coverage and provides a bare-bones record from which a union record can be built. Most of the examples that follow have been edited to some extent either for readability or to save space.\(^2\)

\begin{verbatim}
AN    2902408
DT    ARTICLE
PA    FOLTZ PW
PA    DUMAIS ST
[...]  
AT    PERSONALIZED INFORMATION DELIVERY -  
      AN ANALYSIS OF INFORMATION FILTERING METHODS
JT    COMMUNICATIONS OF THE ACM
DP    1992
IS    DEC
[...]  
ZZ
\end{verbatim}

The INSPEC database\(^3\) also has a record for this article which in addition to duplicating the information contained in the CC record, includes an abstract and an extensive list of subject headings.

\(\text{---}
\)

\(^1\)The MELVYL Current Contents Article database (CC) [contains] records for journal articles in the physical, social, and biological sciences, arts, humanities, and other fields. Article citations in the CC database are indexed by the Institute for Scientific Information, Inc. The database includes citations indexed from July, 1989 to the present, representing publication dates since early 1989. As of 2/14/94, the database contains 5,669,462 citations from over 6500 journals. (Excerpted from the MELVYL online catalog Explain page.)

\(^2\)All were taken directly from MELVYL in late January and early February of 1994.

\(^3\)The MELVYL INSPEC databases [contain] citations to articles in over 4,000 scholarly journals, conference proceedings, books, reports, and dissertations in physics, electrical engineering and electronics, computers and control, and information technology. Most citations include abstracts. As of 2/13/94, the current database contains 1,169,828 citations for materials indexed by the Institution of Electrical Engineers from Jan. 1, 1989 to the present. (Excerpted from the MELVYL online catalog Explain page.)
Finally, the COMP database record sometimes provides, in addition to its own subject headings and a different abstract, the full text of the article, including its citations.⁴

⁴The MELVYL Computer Articles database (COMP) contains records for articles in computer-related magazines and journals. The articles are indexed by Information Access Company and include subjects and abstracts. As of 2/13/94, the database contains 366,950 citations for articles published from January 1, 1988 to the present. (Excerpted from the MELVYL online catalog Explain page.)
Research organizations generate a considerable amount of information, which makes it difficult to keep employees up to date on the work proceeding, and only a small amount of the information [...]

Categories and Subject Descriptors:
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.4 [Information Storage and Retrieval]: Systems and Software

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Indexing methods, information filtering, information retrieval, latent semantic indexing, retrieval models, user models, user profiling

The result is a union record very rich in "clues," both about the object and about objects closely related to it, for the user-- both for retrieval purposes and as a representation.

There are several interesting things of note here: the two abstracts are different; the full text includes the citations (albeit in a format difficult for machines to manage); there are keywords provided for online search; and the ACM classification codes are provided as well.

Even when the MELVYL system does not provide information to fill all the gaps in union record, the Internet can be used to access other available information resources. For example, those pieces of the puzzle which MELVYL may lack can often be found in one of the commercial databases such as the ISI citation databases on DIALOG. These can be used to add citation information to an existing union record, or to search for cited references -- other records that cite the current one (or that the current record cites) which enables the user to do both forward and backward chaining with citations.

By exploiting multiple sources of existing human cataloging and indexing, we can create very complete knowledge about individual bibliographic items. But this is not the only direction in which rethinking traditional bibliographic boundaries can offer new possibilities.
THE DOSSIER

There is a long history of manually interconnecting related bibliographic items, such as bibliographies, citation indexing, book reviews, classification systems, subject headings, commentaries and the like. More recently, the hypertext model has encouraged a freely associative type of object linkage which seems to reflect the way we think and organize our work areas -- some neat and tidy, others inconsistent. In addition, we are also now seeing both of these models spreading, albeit in a very anarchic way, over the Internet.

Online catalogs offer us the opportunity to codify these connections in simple and useful ways. These links can be made explicit within the framework of an online system. For a given topic or query, we can create and make available an electronic dossier, that is, an assembly of objects, all related in some way to one another. Such related materials might be groups of union records, numeric and/or statistical data, Uniform Resource Locators (URLs) which point to some similar or perhaps source documents, or any other Internet resources such as bibliographies and USENET newsgroups.

One simple example of a dossier is that of a book and a group of reviews of that book. Here is a MELVYL catalog record for Brendan Kehoe's *Zen and the Art of the Internet*:

| AN | 7883733 |
| DT | BOOK    |
| PA | Kehoe, Brendan P. |
| MT | Zen and the art of the internet: a beginner's guide |
| ED | 2nd ed  |
| PL | Englewood Cliffs, NJ |
| PU | PTR Prentice Hall |
| DP | c1993    |
| PG | xv, 112 p.; 23 cm |
| SE | Prentice Hall series in innovative technology |

This record can then easily be linked to its reviews, two of which are located in IAC's Magazines database on MELVYL and follow here.

| AN | 13462811 |
| DT | ARTICLE - Review |
| PA | Sheehan, Mark |
| AT | Zen and the Art of the Internet: A Beginner's Guide. (book reviews) |
| JT | Online |
Buckland et al.

| AN | 12549166 |
| DT | ARTICLE - Review |
| PA | Polly, Jean Armour |
| AT | Zen and the Art of the Internet, 2d ed. (book reviews) |
| JT | Library Journal |

Note the reviewer's grade field, "NT" and that in some cases the full text of the review will be available in field "XX."

Book reviews are not the only type of related material easily and automatically derived. The records can be related in a wide variety of ways, including citations (both forward or backward); subject headings, classification numbers, or in fact, any field of the bibliographic record may provide a basis for establishing relatedness.

One particularly useful sort of dossier can be composed of the catalog records and full texts of a group of citation-related items. For example, good use can be made of the full texts available in the COMP database, which have full references at the end. By following these references, a dossier of closely related union records can be built.

Our earlier example, *Personalized Information Delivery*, includes a citation for Belkin and Croft's article on information filtering (Belkin, 1992) from the same issue of CACM. We can retrieve a record for the Belkin and Croft article and build a union record from it, piecing together representations of it from MELVYL's MARGS (Magazine), INSPEC, and COMP databases, which includes 3 different abstracts, two dozen subject headings and the full text. In addition, union records can be built for each of its 33 references. The result is a web of related materials that starts to take on a shape and extent of its own -- a dossier of information about a particular topic as embodied in MELVYL's databases. So far we have described the creation of a dossier by assembling and consolidating excerpts from multiple sources. If, instead, the
dossier were composed wholly or partially in the form of links and pointers, then a dossier in hypertext form would have been created, rather as anticipated by the bibliographer Paul Otlet (1868-1944) early in this century (Rayward, 1994). In a static environment these two forms would be logically identical; in a dynamic environment they would diverge.

This example also gives a hint of what some clever programming could easily provide: a hypertext-like interface to a large multidatabase online catalog's holdings. As we read this article, something interesting might jump out at us such as this:

"It has also been shown that Boolean or structured queries can be very effective when used with an appropriate retrieval model [6, p. 21]."

Normally, all we get for our trouble are the short, tantalizing references at the end of the article, which, though useful, are not all that we might hope for.


[...]

5The MELVYL Magazine Journal Articles database (MAGS) [contains] records for articles from popular magazines, general interest journals, and scholarly journals in the humanities, social sciences, and general sciences. Legal, business, and health journals are also included. The records in the database are from the Expanded Academic Index produced by Information Access Company. Citations for articles in medicine, technical fields, and selected other areas include abstracts. There are also many citations for book reviews, especially in the humanities and social sciences. As of 2/13/94, the database contains 1,268,738 citations for articles published from January 1, 1988 to the present. (Excerpted from the MELVYL online catalog Explain page.)
Buckland et al.


On the other hand, if we use these references as entry points to their respective union records, we get much more. In this case, the union record for the second reference contains, among other things, a useful abstract.


[...] AB An extended Boolean information-retrieval system is introduced that is intermediate between the Boolean system of query processing and the vector-processing model. The query structure inherent in the Boolean system is preserved, while at the same time weighted terms may be incorporated into both queries and stored documents; the retrieved output can also be ranked in strict similarity order with the user queries. A conventional retrieval system can be modified to make use of the extended system. Laboratory tests indicate that the extended system produces better retrieval output than either the Boolean or the vector-processing system.

ZZ

DIFFICULTIES WITH THE UNION BIBLIOGRAPHIC RECORD

Building union records is a task fraught with difficulties. Mapping between tag sets (i.e., AU vs. PN for the author's personal name field) is a relatively simple process (albeit time consuming). The real problem lies in the quality of the data and the many different forms in which it appears. One example is the many different forms in which names appear -- Buckland Michael vs. Buckland M vs. Buckland MK vs. Buckland M.K.
The many name forms may not be surprising, but the differences show up in more unexpected places. Here is an example where the title of an article appears in three different forms. Note the title inversion between the first and second examples.

AN 4383502
[...]  
AT Three views of virtual reality: nonimmersive virtual reality.  
JT Computer  
[...]  
ZZ

AN 13756146
[...]  
AT Nonimmersive virtual reality. (Three Views of Virtual Reality)  
JT Computer  
[...]  
ZZ

AN 2719441
[...]  
AT NONIMMERSIVE VIRTUAL REALITY  
JT COMPUTER  
[...]  
ZZ

These problems can be overcome by triangulating the many different data points present in a record -- periodical title, volume, number, date of publication, page numbers, article title, author(s). While there has been major standardization of library catalogs (AACR2, ISBD, MARC), the next logical step is now required. There must be broader standardization of bibliographic records to achieve interoperability between a wider range of bibliographic services (abstracting, citation, and text indexing) and text repositories.

The Union Bibliographic Record also raises a number of challenging information presentation and interface design issues. How should the union record be represented to the user? How can the user be asked for input regarding the attributes on which related materials should be found (i.e., subject, document type, citations, cited references, etc.)? To what extent should the process of searching be automated and to what extent should the user be involved in directing the search? For example, which location(s) should be searched first and in what order? Should the dossier of related documents
be automatically built up or only built on an ad hoc basis as the user specifically requests each related union record. All of these issues point to much work that needs to be done.

CONCLUSION

In this paper, we have suggested several new ways to view information space. The first is a way of combining multiple representation into a single, more comprehensive and informative object: a union record. Such a union representation offers users a much broader perspective on the scope and depth of the topic of interest. The key to this solution is to view the heterogeneous, but overlapping parts of the databases, not as a problem, but as assets to combine and exploit.

In the second case, we argued that multiple, related bibliographic entities be viewed as a single, extended information object or dossier which borders on a sort of online catalog hypertext document. Such a web of bibliographic data offers a broader vantage point for engaging in the information seeking process.

As online catalogs proliferate, as access services become more interconnected, and incorporate an ever-increasing range of resources, they begin to approach a critical mass where what was once bibliographic data can start to become bibliographic information.

REFERENCES


Union Records & Dossiers


