STUART WEIBEL

Research Scientist
OCLC Office of Research
Dublin, Ohio

Automated Cataloging: Implications for Libraries and Patrons

ABSTRACT

Many changes in cataloging practice have been and will increasingly be technology driven. Bound lists and drawers of cards defined the form and function of catalogs for as long as they existed until the advent of digital computers. Even today, however, MARC records are as much a derivative of catalog cards as the reverse. The additional functionality of computer catalogs affords opportunities to increase the effectiveness of the cataloging process and improve the value of the catalog itself. Three main research areas are examined with regard to their anticipated influence on this evolution. Automated cataloging research, focusing on the application of rule-based systems to cataloging, represents a novel way to address the cataloging process per se, but has as yet made only modest progress. The incremental implementation of a variety of computer-assisted methods for addressing aspects of cataloging represents a second, more conventional approach to advancing the state of the art in cataloging automation. This approach shares the goal of the first—to build intelligent capabilities into cataloging systems—but the focus remains on human cataloging systems and the methods of implementation are more conventional. The third area is not part of traditional concepts of cataloging at all, but will have a major impact upon what is available in catalogs in the broadest sense of that term. This "nontraditional" cataloging involves automated processing of documents to extract bibliographic information as well as full text. It will expand the range of cataloged objects to include items not generally

cataloged due to resource constraints. Automated processing of such materials will be characterized by lower quality and less complete cataloging, but will nonetheless promote improved access to materials that are currently lost to bibliographic control.

INTRODUCTION

Mark Twain is said to have remarked that when a writer is tempted to use the word "very," he should write "damn" instead, prompting the editor to excise the offending word, thereby improving the quality of the writing. We might apply Twain's advice to the phrase *artificial intelligence*. We are in fact concerned about making library processing more efficient and providing greater value to the patron. Whether this is under the guise of artificial intelligence (AI) or simply intelligent system design is of little consequence to a librarian or a patron. Some of the projects described here fall into the classical (if that is the proper word) AI category, but others are simply the well-considered application of human intelligence captured in the idiom of what once was described as intelligent programming. In aggregate, these projects promise more intelligent systems that ease the burden of catalogers, patrons, and perhaps even library budget officers.

There are many ways to characterize the procedures and results of cataloging processes, and the particular perspective one adopts will necessarily influence the characterization of progress and prospects. It is therefore useful to provide a perspective on the salient issues that will serve as a common foundation for further discussion.

The Cataloging Process: Two Perspectives

The library's perspective on cataloging is as a technical process that can be measured in terms of books processed, items added, shelf lists lengthened, and catalogers employed. The concern of the library is to efficiently and expeditiously provide access to the holdings of the library. Thus, cataloging activity is a bottleneck in making available newly acquired material to the patron and represents a major commitment of staff resources. The existence of OCLC, RLG, Utlas, and other technical service companies is ample testimony to the incentive to reduce the cost of cataloging through resource sharing. Automation of cataloging processes represents a further opportunity to reduce these costs and will therefore be a major concern of the library.

The patron's perspective on cataloging has to do with what she or he can or cannot find in the catalog. Electronic catalogs have made the searching and identification of materials more flexible and effective,

and future improvements in online public access catalogs promise further benefits. To the extent that cataloging practice influences what is available and how it is located, the patron is influenced.

The Cataloged Object

The changing technology of today's library automation environment is having a major impact on the catalog. There is by no means complete agreement as to the desirability of this change, but there can be no argument that the change is ongoing and is having dramatic consequences. The traditional catalog was a collated list of monographs, periodicals, and a variety of special objects; this notion retains a vigorous existence in most of our minds. Increasingly, however, a catalog entry is a surrogate for an item that may exist at a location distant from the physical location of the catalog and is available either by loan or online. It is more often a *work* that is the object of our desire rather than the physical object itself.

Patrick Wilson (1989) captures this notion in his essay entitled "The Second Objective." Wilson proposes a rethinking of bibliographic organization that emphasizes the organization of *works* rather than a particular manifestation of a work and provides access to these works independent of their geographic location or current state of revision or reprinting.

The Catalog

The idea of the catalog itself is enlarged by advancing technological capability. John Duke (1989) proposes the notion of a virtual catalog with "tripartite record structure," a three-tiered catalog that encompasses everything but the physical artifact.

1. Document Surrogates: the traditional notion of an abbreviated, formalized citation structure.
2. Document Guides: synopsis of content—tables of contents, indexes, abstracts, summaries; an "enriched" record.
3. Document Texts: the work itself in electronic or digital format available for distribution.

It is this enlarged idea of cataloging (and the catalog itself) *a la* Wilson and Duke that will have a direct, substantive impact on the library and library patrons. In order to describe what this impact might be, it is useful to distinguish three distinct areas of research and the prospects for each.

AREAS OF CATALOGING AUTOMATION RESEARCH

These areas are complementary and overlapping. There are activities that easily fall into more than one of them, but they cover, in aggregate, the range of automation activities that have a strong influence on cataloging practice.

- fully automated cataloging; cataloging untouched by human hands;
- computer-assisted cataloging; tools or utilities, either active or passive, that could enhance the human cataloger's productivity; and
- nontraditional cataloging; automated processing of materials not typically included in conventional cataloging workflow.

Fully Automated Cataloging

The concept of a cataloging robot lies at the center of this area. The goal is to embed cataloging expertise in a system that has access to machine-readable versions of items to be cataloged and generate appropriate bibliographic surrogates and guides in an automated way. No one working toward this goal can long harbor illusions about the near-term prospects in this area. Nonetheless, the results of such work can have important side effects for real cataloging systems and can as well, perhaps, teach us something about what the successor to AACR2 should look like. Indeed, this last outcome is suggested by some to be the most important potential result.

The research environment supporting this area is a difficult one. Conceptual analysis is helpful, but at some point the proof is in the cataloging, and without convincing demonstrations of actual cataloging by machine, the effort is sterile. Building prototype systems is a difficult and costly activity with a number of seemingly intractable problems.

The first study of the feasibility of automating the cataloging process was in a dissertation written by Martha Fox at the University of Illinois at Urbana-Champaign. This study set out to "determine whether the human intellectual process of cataloging bibliographic materials could be simulated by automatic, namely, objective, non-intuitive, computer techniques" (1972, p. 3). One of Fox's conclusions merits mention in the context of current work in this area:

> Finally, if librarians are to consider a system in which automated cataloging is to play a part, it is essential that the intellectual structure of the existing cataloging process be reexamined in light of the capabilities and operations performed by machine. (pp. 304-05)

Davies and James (1984) published the first attempts at actually encoding some component of cataloging rules, and although their attempt was somewhat bogged down in the implementation aspects of building a system, Davies (1986) subsequently described many issues

which anticipated later efforts in this area. Note that Davies is not an advocate of fully automated cataloging systems, but rather proposes that the rule-based system work interactively with a cataloger.

Helga Schwarz (1986) of the Deutsches Bibliothekinstitut proposed an approach to automating the extraction of bibliographic descriptors from title pages in a three-step process: (1) recognition of types of data, (2) recognition of the function of data, and (3) applying appropriate rules to formulate a cataloging record. Unlike Davies, she counts herself among the advocates of the cataloging robot, acknowledging that it may not be a reality in the near future.

Elaine Svenonius and Mavis Molto have two papers (1990, in press) that address issues in automated cataloging. The goal of these studies is to advance the theoretical underpinnings of automated cataloging as well as to provide pragmatic methods that could be incorporated into actual systems. Among the virtues of these studies is their foundation in real data. The authors randomly selected English language monographs from the UCLA stacks and systematically applied their ideas to the data. The results are useful heuristics that can be employed in practical systems that could be implemented today.

These studies emerged from previous work (Svenonius et al., 1986) in this group addressing conceptual issues in cataloging that bear on the rules supporting the choice of name-access points. These efforts in aggregate illustrate the close interaction of rule structures and the consequences these structures have for automated systems. The obvious question emerges: Should changing technology influence the way rules are structured or should the technology simply implement the rules?

Ling-Hwey Jeng (1986, 1988) has explored the potential for automating cataloging using title page information incorporating AACR2 into a structure suitable for application in an automated environment. Her recent work addresses the structure of AACR2 rules and the implications for implementation in an automated environment (1990).

Dissertation research at UCLA by Zorana Ercegovac, under the direction of Harold Borko (Borko & Ercegovac, 1989) approaches another aspect of cataloging: map cataloging. These investigators explored issues in the application of written and unwritten procedures for assigning map authorship. This study recognizes that necessary expertise in such tasks extends beyond that which is articulated in formal rule sets, and such considerations must inform any successful attempt at automating these processes. The technological impediments of automated map cataloging far outweigh those of monograph cataloging, mitigating against application of such ideas; however, the authors suggest that their work might profitably be applied to training of catalogers.

The Automated Title Page Cataloging Project (Weibel et al., 1989) at OCLC represents an attempt to demonstrate the feasibility of

descriptive cataloging from title page images without the intervention of humans. The prototype was implemented as a rule-based system in Prolog; the objective of the system was to generate a first-level bibliographic description from information on the title page.

Sample title pages were selected randomly from current cataloging on the OCLC Online Union Catalog at the time of the study. Scanning and optical character recognition (OCR) were not (and are not now) sufficient to generate accurate representations of the title pages, so machine-readable versions of these title pages were rendered in a typesetting language and parsed automatically for the tests. In this way it was possible to tackle the conceptual problems associated with format recognition without being unduly handicapped by the realities of the technological limitations of scanning and OCR.

It is this thread of unreality that pervades to some extent all the automated cataloging studies alluded to above. They share an earnest attempt to address the conceptual problems in this area and a willingness to overlook the practical limitations that loom as large obstacles to implementation of production systems. This is not to say that such studies are fruitless; the value of these efforts lies in three areas:

1. providing a better understanding of the problems that must be solved to automate cataloging procedures,
2. pointing to productive ways to restructure cataloging procedures such that future automation attempts will have greater prospects for success, and
3. developing teaching simulators to enhance cataloging education.

They are unlikely to change technical processing in the library in the next five years, however.

Computer-Assisted Cataloging

Virtually all cataloging now performed in libraries is in some sense computer-assisted cataloging. For the purpose of this discussion, included somewhat arbitrarily are those tools not in common use but which will become more widespread in the near future. The implementation of such tools will have a major impact on technical processing departments. Are these artificial intelligence? Robert Burger (1984), in a paper entitled "Artificial Intelligence and Authority Control," made the statement: "artificial intelligence is already used in libraries...one of the major responsibilities of catalogers, machine-based authority control, is a form of artificial intelligence"(p. 344).

Whether such efforts should be considered artificially intelligent is moot; one may simply understand such capabilities as part of the naturally evolving capability of intelligent systems which support the

cataloging effort. Several such projects in progress in the OCLC Office of Research illustrate the point.

Duplicate Detection

The Duplicate Detection Project at OCLC (O'Neill, 1989) is a good example of a practical implementation of a rule-based system that involves no specialized languages or unconventional techniques. It is the embodiment of a high degree of expert knowledge—the knowledge of an experienced cataloger—in combination with matching similarity algorithms to detect duplicate records in the OCLC Online Union Catalog. As such, it is a useful cataloging utility that has a variety of applications in a cataloging workstation as well as in its current batch processing implementation. A program such as this one which monitored cataloging input could contribute to preventing the addition of duplicate cataloging records rather than cleaning them up after the fact.

The current implementation has identified 80 percent of the duplicates in test samples with less than 0.5 percent misidentification of pairs of nonidentical records as duplicates.

Subject Heading Correction

A review article of online database quality control (O'Neill & Vizine-Goetz, 1988) describes a variety of error correction techniques that can be applied to databases. One of the authors, Edward T. O'Neill, currently is leading an effort to correct errors in subject headings in the OCLC Online Union Catalog. Two million records have been corrected in the initial phases of this effort; a million or more are expected to be corrected in a second phase. These efforts improve the quality of cataloging databases, thereby making cataloging more effective and making catalogs more useful to patrons.

Cataloger's Assistant

Diane Vizine-Goetz (1989) is leading a team that is developing a prototype cataloger's workstation for use in actual cataloging production. The system is now being tested at Carnegie-Mellon University Libraries for reclassifying a mathematics and computer science collection and applying subject cataloging to new items in these subject areas. The prototype makes available the Dewey Decimal Classification (DDC), machine-readable Library of Congress Subject Headings (LCSH-mr), and OCLC cataloging data in a HyperCard interface on the Apple Macintosh. The goal of this study is to explore the following issues in a production cataloging environment:

- How can the structure of DDC and LCSH best be conveyed to the user? How should these systems be linked?

- What browsing and navigational tools are appropriate for this application?
- What searching capabilities are necessary to support effective usage of these resources?

Nontraditional Cataloging

The term "nontraditional cataloging" is intended to describe the processing of materials that do not command the attention of a complete cataloging effort but should nonetheless be available for retrieval at some level, typically in an electronic database or catalog. The so-called "grey literature" or fugitive documents have traditionally fallen outside the body of fully cataloged items due to resource constraints or perceived lack of importance. Journal articles, pamphlets, correspondence, and office documents come to mind as examples of materials for which identification and retrieval are often substandard.

In addition, there are new forms of communication that are becoming widespread, such as E-mail and electronic newsgroups. Are such items worthy of cataloging? The answer to this question is a pragmatic one; they will be cataloged (in the broadest sense of the term) to the extent that the benefit is perceived to justify the cost. To the extent that cost is low and the process automated, more materials will be cataloged.

The goal is to capture in an automated way something like a cataloging record that is useful for search and retrieval. It is unlikely that automated systems will provide records of quality equivalent to human cataloging, but the increased access to an otherwise poorly accessible body of information should nonetheless be useful.

Project ADAPT

Project ADAPT is an ongoing project in the OCLC Office of Research to automate the conversion of paper documents to SGML-structured, machine-readable form and to provide searchable indexes for retrieving such documents.

The document representation continuum (see Figure 1) extends from the physical document (or its image) to a structured logical representation that includes the indexed text, associated graphics, and functional role of document objects (title, author, abstract, etc.), all represented in a database structure that will afford multiple views of the document and will support a wide variety of retrieval and presentation options.

The goal of Project ADAPT is to move incrementally from one end of the continuum toward the other. Image-based systems are now being produced for archiving and preservation activities. However, more sophisticated document representations can be expected to improve the utility and flexibility of such systems.

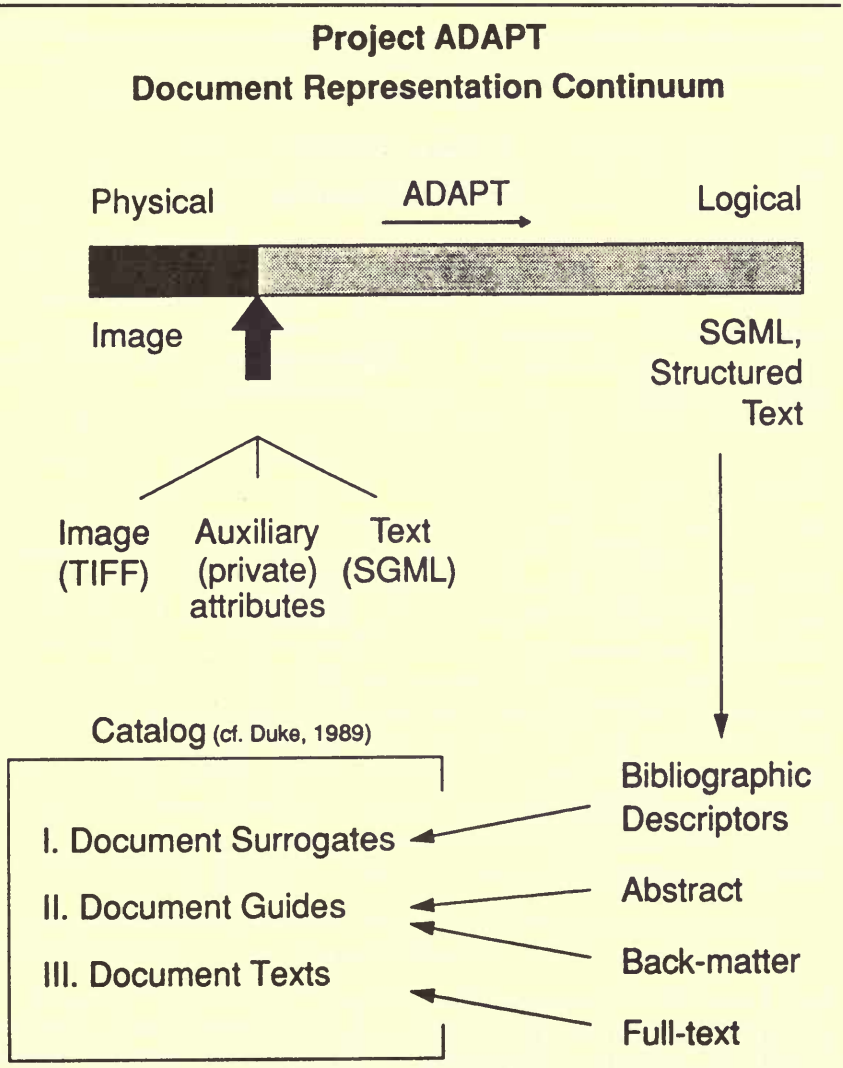

Figure 1. Document representation continuum

System Overview

Figure 2 represents the overall system design for a completed document processing system. The details of user interfaces and formatting of output are important production system considerations, but are of only minor concern to our project activities. The Newton Database server is also largely a production concern; it exists as the result of an intensive development activity in OCLC's development

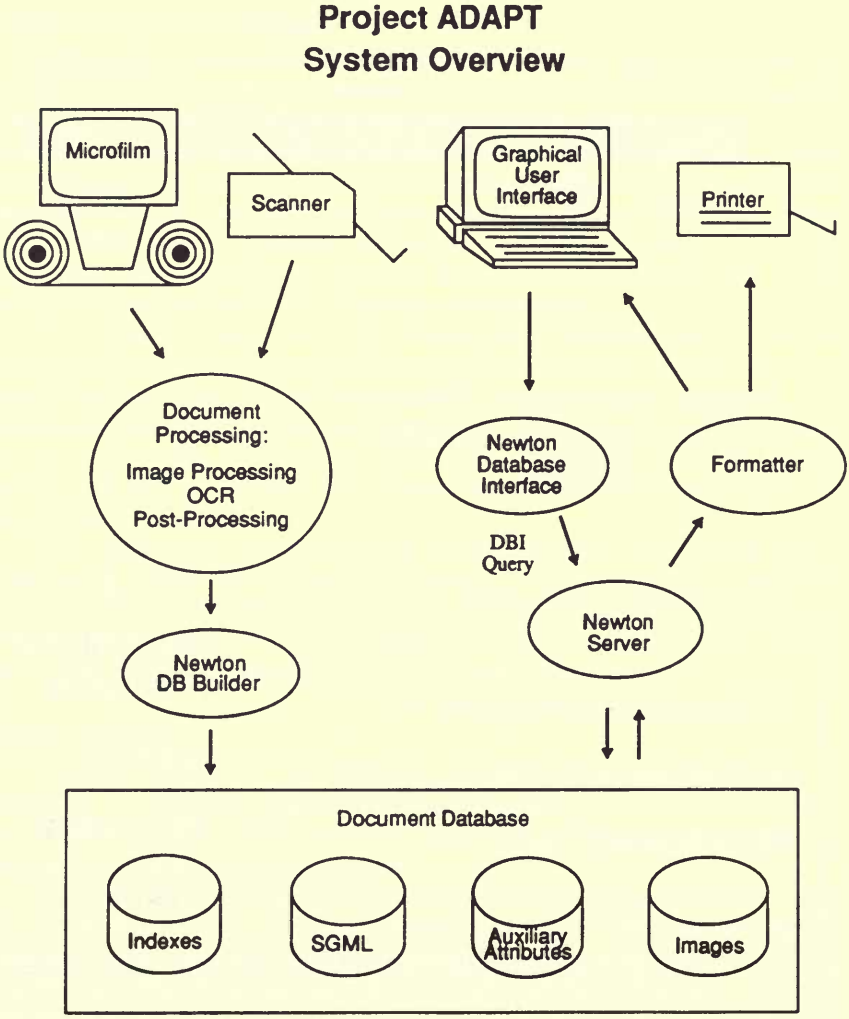

Figure 2. Project ADAPT system overview

department and is central to a number of OCLC products and research projects.

The overall strategy is to add value to commercial OCR capabilities by pre- and post-processing of the documents:

1. *Document Image Pre-processing:* The first processing stage is image pre-processing, which entails segmentation of the document image and classification of the segments into layout objects of several types. The segmentation process identifies rectangular layout objects for which a variety of statistical attributes are subsequently generated. These statistical attributes are then used to classify the objects as text, graphics, or extraneous noise. Text objects can then be passed through commercial optical character recognition devices.
2. *OCR Processing:* All OCR processing is done with commercial OCR systems. We use the Calera CDP 9000 system, but the process is designed to be independent of the OCR device used. Indeed, one of our strategies for reducing error in the processing is to employ multiple OCR processing and merge the results. This approach has resulted in reduction of errors by approximately 40 percent.
3. *Text Post-processing:* Post-processing includes activities from error correction to analysis of the structure of the document and markup of the document (in SGML) to reflect that structure. Document structure analysis involves the coordination of OCR output with associated layout objects such that the bitmap location of a particular line of text and other attributes of that line are accessible. This information can be used to determine the role of a particular text object (for example, titles, authors, and abstracts). This record, in conjunction with machine-generated indexes, then affords access to a work that might otherwise have little or no other means of retrieval.

Related Projects

There are many related projects in this country and elsewhere, each with somewhat different focus, but all with the common goal of making various types of written materials more readily available for search, retrieval, and distribution. Table 1 identifies representative projects. Not everyone will agree that the results of such processing will be uniformly good. Speaking about full-text access to cataloging records, Helen Schmierer (1989) writes: "Librarians and users will soon discover in online files of only moderate size that word access, while powerful, produces some bewildering results" (p. 112).

It is probably true that access to every word in a cataloging record or the document itself will raise many problems and much bewilderment, but the response should be to solve the problems rather than back away

from them. In Duke's broad sense of the catalog, these systems are "cataloging" systems and they will ultimately promote greater access and availability of materials not now well represented in current catalogs. The occasional bewilderment and inherently greater complexity of retrieval in such a full-text world is a price that must be accommodated. Some of these problems will be mitigated by the maturation of the technology. In the long run, the patron will be well served by such capabilities.

TABLE 1
EXAMPLES OF TEXT DIGITIZATION PROJECTS

Organization	*Reference*	*Project Description*
Hochschule Darmstadt, Dept. of Information Science	Endres-Niggemeyer (1987)	AUTOCAT: OCR and automated cataloging of journal articles in the physical sciences
National Library of Medicine	Thoma et al. (1985)	Prototype system for electronic storage and retrieval of medical journal articles
OCLC Office of Research	Weibel et al. (1989)	Project ADAPT: automated document structure analysis and SGML markup
Nuclear Regulatory Commission	Bender (1988)	Optical disk-based system to deliver text, images search, and retrieval capabilities
National Agricultural Library	Andre & Eaton (1988), Zidar (1988)	Text Digitizing Project: scanning and OCR of agricultural documents for electronic retrieval and delivery
German National Research Center for Computer Science	W. Putz (personal communication, January 5, 1990)	Prototype system for conversion of paper documents to SGML structured documents
University of Strathclyde, Glasgow	F. Gibb (personal communication, January 24, 1990)	SIMPR: Software tools for indexing, retrieval, subject analysis, and structured information management

CONCLUSION: PROSPECTS FOR THE FUTURE

The projects described above are part of the foundation for future advances in the technology of cataloging. These projects provide a useful horizon to help gauge the future of cataloging, but the rate of progress

toward such goals (and even the solidity of the goals themselves) is difficult to predict. Incremental progress will be made by implementing useful, practical cataloging tools—duplicate detection, more advanced authority checking, subject authority correction algorithms—in relatively conventional production environments.

The systems which result will be intelligently implemented rather than intelligent, they will be real rather than artificial, but, most importantly, they will make the cataloging process more practical and more efficient.

The cataloger will have increasingly sophisticated tools to augment the traditional process of cataloging, resulting in a better product at a lower cost. The patron will benefit from this by virtue of the indirect benefits of a more efficient operation.

As these parts of cataloging systems mature, research in the conceptual structure of cataloging and the automation of cataloging processes should provide a foundation to support longer term changes in cataloging and the systems to support it. When such changes take place, they will have also resulted from incremental progress toward an understandable goal. Some of the techniques that will have been applied to reach these goals are included in what are commonly understood to be artificial intelligence techniques; others will have been more conventional.

The other realm of potential improvements will come from the low end of the cataloging spectrum: the conversion of paper or microform to more accessible media—the electronic document. The large number of documents now in relatively unaccessible paper or microform that is not indexed or cataloged by humans can be rendered more accessible through a process of conversion to electronic format and machine indexing. The high level of research activity in this area suggests that systems to automate this conversion will have a major impact on cataloging information in the broadest sense of the word.

REFERENCES

Andre, P. Q. J., & Eaton, N. L. (1988). National agricultural text digitizing project. *Library Hi Tech, 6*(3), 61-66.

Bender, A. (1988). An optical disk-based information retrieval system. *Library Hi Tech, 6*(3), 81-85.

Borko, H., & Ercegovac, Z. (1989). Knowledge-based descriptive cataloging of cartographic publications. In *Annual review of OCLC research, July 1988-June 1989* (pp. 49-50). Dublin, OH: OCLC.

Burger, R. H. (1984). Artificial intelligence and authority control. *Library Resources & Technical Services, 28*(4), 337-345.

Davies, R. (1986). *Cataloguing as a domain for an expert system.* Chichester, England: Ellis Horwood.

Davies, R., & James, B. (1984). Towards an expert system for cataloguing: Some experiments based on AACR2. *Program: Automated Library and Information Systems, 18*(4), 283-297.

Duke, J. K. (1989). Access and automation: The catalog record in the age of automation. In E. Svenonius (Ed.), *The conceptual foundations of descriptive cataloging* (Papers presented at the Conference on the Conceptual Foundations of Descriptive Cataloging, 14-15 February 1987) (pp. 117-128). San Diego, CA: Academic Press.

Endres-Niggemeyer, B. (1987). Eine reprasentation der informationsstruktur von fachzeitschriften [A representation of the information structure of scientific journals (AutoCat)]. *Nachrichten fur Dokumentation, 38,* 333-340.

Fox, A. S. (1972). *The amenability of a cataloging process to simulation by automatic techniques.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Illinois.

Ling-Hwey Jeng. (1986). An expert system for determining title proper in descriptive cataloging: A conceptual model. *Cataloging & Classification Quarterly, 7*(2), 55-70.

Ling-Hwey Jeng. (1988). The language of a title page. In C. L. Borgman & E. Y. H. Pai (Eds.), *ASIS '88: Proceedings of the 51st ASIS Annual Meeting* (23-27 October 1988) (pp. 31-35). Medford, NJ: Learned Information.

Ling-Hwey Jeng. (1990). *The general structure of a knowledge base for Anglo-American Cataloguing Rules: Final report* (Technical report of a project funded under the OCLC Library and Information Science Research Grant Program).

O'Neill, E. T. (1989). Duplicate detection. In *Annual review of OCLC research, July 1988-June 1989* (pp. 15-16). Dublin, OH: OCLC.

O'Neill, E. T., & Vizine-Goetz, D. (1988). Quality control in online databases. In M. E. Williams (Ed.), *Annual review of information science and technology* (Vol. 23, pp. 125-156). Amsterdam: Elsevier Science Publishers.

Schmierer, H. F. (1989). The impact of technology on cataloging rules. In E. Svenonius (Ed.), *The conceptual foundations of descriptive cataloging* (Papers presented at the conference on the Conceptual Foundations of Descriptive Cataloging, 14-15 February 1987) (pp. 101-116). San Diego, CA: Academic Press.

Schwarz, H. (1986). Expert systems and the future of cataloguing: A possible approach. *LIBER Bulletin, 26,* 23-50.

Svenonius, E.; Baughman, B.; & Molto, M. (1986). Title page sanctity? The distribution of access points in a sample of English language monographs. *Cataloging & Classification Quarterly, 6*(3), 3-21.

Svenonius, E., & Molto, M. (1990). Automatic derivation of name access points in cataloging. *Journal of the American Society for Information Science, 41*(4), 254-263.

Svenonius, E., & Molto, M. (in press). Studies in automatic cataloging. *Journal of the American Society for Information Science.*

Thoma, G. R.; Suthasinekul, S.; Walker, F. L.; Cookson, J.; & Rashidian, M. (1985). A prototype system for the electronic storage and retrieval of document images. *ACM Transactions on Office Information Systems, 3*(3), 279-291.

Vizine-Goetz, D. (1989). Cataloger's assistant. In *Annual review of OCLC research, July 1988-June 1989* (p. 8). Dublin, OH: OCLC.

Weibel, S.; Oskins, M.; & Vizine-Goetz, D. (1989). Automated title page cataloging: A feasibility study. *Information Processing & Management, 25*(2), 187-203.

Wilson, P. (1989). The second objective. In E. Svenonius (Ed.), *The conceptual foundations of descriptive cataloging* (Papers presented at the Conference on the Conceptual Foundations of Descriptive Cataloging, 14-15 February 1987) (pp. 5-16). San Diego, CA: Academic Press.

Zidar, J. A. (1988). National agricultural text digitizing project: System startup and operation. In M. E. Williams & T. H. Hogan (Eds.), *National Online Meeting Proceedings—1988* (Proceedings of the Ninth National Online Meeting, 10-12 May 1988) (pp. 443-448). Medford, NJ: Learned Information.