January 1987

Artificial Intelligence, Expert Systems, and Archival Automation

Peter B. Hirtle
National Library of Medicine

Follow this and additional works at: http://digitalcommons.kennesaw.edu/provenance
Part of the Archival Science Commons

Recommended Citation
Available at: http://digitalcommons.kennesaw.edu/provenance/vol5/iss1/8

This Article is brought to you for free and open access by DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Provenance, Journal of the Society of Georgia Archivists by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.
Artificial Intelligence, Expert Systems, and Archival Automation

Peter B. Hirtle

The success of archival automation during the past two decades cannot be questioned. From the development of the SPINDEX (Selective Permutation Indexing) program at the National Archives, through the work of the National Information Systems Task Force (NISTF) committee, to the development of the MARC AMC (MACHINE Readable Cataloging, Archives and Manuscripts Control) format, automation's progress, while perhaps slow in comparison to the library profession, has been sure.1

The relative success of current archival automation programs is especially surprising when one considers that automation in an archives promises few of the cost savings available in a library context. By sharing cataloging records, librarians can hope to increase the quality of their catalog and decrease the per item cost of cataloging. In an archival environment, however, most materials are unique; there can be little sharing of cataloging records, and no decrease in overall cataloging costs. In most instances it is likely that conforming to national standards will actually be more expensive than cataloging according to local practice. Improved control and access for archivists and users, and not cost savings, have been the driving forces behind archival automation.2

Of course, the use of computers in archives has lowered some of the clerical costs associated with processing. On the local level, finding aids, guides, and inventories prepared on word processors have either been less expensive or of a higher quality than their typed cousins. In addition, some archives have experimented with using database management systems
for inventories and other finding guides descriptions as well. On a national level, the MARC AMC format holds out the promise of becoming a machine-readable supplement to, if not a replacement for, the manually generated National Union Catalog of Manuscript Collections. Widespread acceptance of the MARC AMC format should result in improved archival reference services.

Yet, the professional activities in an archives, including the arrangement and cataloging of collections, still remain the task of the trained archivist. While quicker access to collection or series descriptions can be achieved by placing the descriptions into database management systems or national bibliographic utilities, a professional archivist still must describe the collections before the descriptions can be added to a database. In addition, a professional archivist is needed to translate researchers' questions into terms which can be used to interrogate the database. Archival automation has made the clerical tasks involved in describing collections in archives easier; it has not altered or replaced, however, the intellectual and professional skills archivists bring to their jobs.

Recent developments in the field of artificial intelligence (AI) may change this picture. In particular, with the development of expert systems, it has now become possible to foresee a time when archival automation may actively assist in the processing of collections and even meet some of the reference needs of the users. No expert system currently exists which is ready for use in an archives. Yet, prototype systems are currently being designed, and from these prototype systems the operational expert systems of the future will develop. For example, expert systems designed to perform automatically many of the cataloger's duties are already under development in libraries. Librarians are also experimenting with expert systems which can respond to reference inquiries, and the National Archives has conducted preliminary investigations into the possible reference use of expert systems in an archives.
The development of expert systems engenders with them, because of their nature, standardization of routines and other activities. Expert systems promise a savings in professional expense, so it is likely that they will be adopted. However, unless the archival community as a whole becomes involved with the pilot projects, the danger exists that the standards developed for other purposes, such as library cataloging (as happened initially with the MARC format), or created because they represent a system analyst's view of an archives and its activities, will be forced upon the profession. This article, therefore, is intended to introduce archivists to the basic concepts and vocabulary of expert systems, and to bring to their attention the pioneering applications work already under way.

For most of its history the computer has been primarily a number-cruncher, capable of doing millions of calculations in the scantest period of time, but incapable of dealing with symbolic representation or abstract thought. Only with the development in the 1970s of the new fields of cognitive psychology and its computer-based cousin, artificial intelligence, could computers begin to live up to the dreams of their first creators in terms of analytical and symbolic reasoning ability. Artificial intelligence has been defined as research efforts aimed at studying and mechanizing information processing tasks that normally require human intelligence. Researchers have discovered that people do not explore equally all possible approaches to a problem when they wish to solve it. Rather, they use their problem-specific knowledge and their knowledge surrounding the problem (their "domain knowledge") to help them understand issues and to limit the possible approaches or solutions to the manageable few that are most likely to succeed.

A chess champion, for example, does not automatically analyze all the possible implications of every move available at any moment. In chess the average number of moves that can be made from a given position is thirty-five; an exhaustive search of the
possible outcome after three moves by each player would require the examination of more than 1.8 billion moves. Instead, chess champions have developed an "expertise" which helps them limit their analysis to those moves which are most likely to lead to positive results. Expert systems, an application of the findings of artificial intelligence studies, are computer programs which try to embody the heuristic, or rule-of-thumb, reasoning of experts.

An expert system is built by first creating a knowledge base containing the knowledge of the expert. In most systems the knowledge is in the form of rules, expressed in a series of if/then statements, though some systems can induce rules from examples provided by the programmer. Any essentially rule-based expert knowledge system is a prime candidate for development into a knowledge database.

The knowledge database is processed by the heart of the system, the inference engine. Inference is most often achieved through backward chaining; when the answer (the "then" part of the statement) is known, the computer then works backwards through a series of "if" statements. The system expresses which of several alternatives is more likely to lead to useful results, usually by giving the answer as a probability, and can often show the user how the solution was reached. A recent development in the field has been the proliferation of relatively inexpensive expert system "shell" programs, many of which are designed to run on microcomputers, and which only require the addition of a profession's rules in order to function.

Most professional tasks, such as those of archivists, are guided by a set of heuristic rules. In some cases, these have been articulated into a set of clearly stated rules. In other cases, they are implicit in the knowledge of the professional, but have never been expressly articulated. Expert systems identify and codify the implicit or explicit heuristic rules present in most professions, and then apply them. Several by now quite famous expert systems, embodying the knowledge of professional
practitioners, have been developed. Included among them are MYCIN, a program designed to assist in the diagnosis of bacterial infections; DENDRAL, which identifies chemical spectrographs; and PROSPECTOR, which aids in the search for minerals and other natural resources.10

While MYCIN and the other programs were developed by first identifying and then transferring to the computer the heuristic principles which the practitioners in each discipline follow, the task of developing an expert system is made easier when the rules are already codified. An example of the latter would be library cataloging. In the Anglo-American Cataloguing Rules, 2nd edition, (AACR2), librarians (and archivists) have an established set of rules guiding how material is to be cataloged. An expert cataloging system would embody the rules in AACR2 and then apply them to a book or archival collection in order to catalog it automatically. Rather than have a cataloger who knew the rules apply them, the computer would.11

The development of a microcomputer based cataloging program would be of tremendous use to archivists. It could be the perfect cataloging assistant for those archivists who wished to catalog their collections in the MARC AMC format and according to national standards and, yet, who are not in and of themselves expert catalogers. Although there have been as yet no publicized attempts to develop an expert archival cataloging system, preliminary efforts at developing an expert system embodying the general cataloging rules in AACR2 have been made at the University of Exeter and at Linköping University in Sweden. Their general findings on the possibility of incorporating AACR2 into an expert system should be of interest to archivists faced with the new task of cataloging.

Both groups of researchers concluded that while AACR2 may be understandable to an expert cataloger, it is not intelligible to computers.12 The rules, while on first glance exact in their formulation, are on closer examination quite inexact; cataloging requires
a high degree of judgment and interpretation on the part of the cataloger. The development of a fully automated expert cataloging system, therefore, would require a new set of cataloging rules--ones solely in the form of production rules, which are unambiguous in their application and, hence, comprehensible to a computer. 13 An example of a cataloging production rule would be, "If a work is a monograph and the work is by one personal author then the main entry equals the heading for personal author." 14 Preliminary talks are under way at the Library of Congress and the other national libraries to discuss the development of a new set of computer compatible cataloging rules, but it would appear that it will be a long while before cataloging is performed by a computer.

While the Exeter and Linköping Universities researchers held out little hope for the development of an expert cataloging system without first effecting major changes in the nature of the cataloging rules, they did uncover certain elements of the cataloging process which could be effectively performed by the computer. In particular, they independently came to the conclusion that the second part of the AACR2 code which deals with the selection of access points could be expressed in the production rule format required by expert systems. Robert Burger has independently suggested that artificial intelligence systems could be used in authority control. He envisions a system which would link automatically all forms of a given name under which a user might search and thus, in effect, remove the idea of a primary access point altogether.15

The use of an expert system in the formulation of authority headings may be one of the easiest, and hence first, practical applications of expert systems in libraries. It is also a clear example of how the standardization which increased automation engenders may be inimical to existing archival practice. As Steven Hensen has noted, the selection of access points has been seen "as a potential minefield for archivists and thought to be best avoided."16 In particular, the emphasis on common usage rather than
administrative hierarchy in the formulation of corporate headings in AACR2 has been viewed as a stumbling block for archivists. Hence, the impact of an automated expert authority system on archivists could be considerable.

If archivists are going to intermingle their records with more standard bibliographic descriptions, as is happening in the national bibliographic databases, they will have to be involved actively in decisions regarding authorities for the databases. The Society of American Archivists's Committee on Archival Information Exchange and the different AMC working groups in the bibliographic utilities must stay alert to developments in computer-assisted authority control and its possible impact on archivists, to insure that the interests of archivists are not overlooked.

Much library research into possible applications of artificial intelligence and expert systems has been in the area of technical services and, in particular, on the application and interpretation of AACR2. Many more potential applications for expert systems exist in the library and archives, however. Expert systems, for example, could be developed for the training of new archivists. If it were possible to condense an expert's knowledge into a database—in effect, to capture the institutional memory of an experienced archivist—it would then be possible to design a system which would lead an untrained or poorly trained archivist or a researcher through the same retrieval steps which an archival master would follow.

Further, expert systems could be part of a records management program and assist in the selection, scheduling, and arrangement of records. Perhaps, through the application of fuzzy reasoning and other artificial intelligence concepts, an expert system could be developed which could assist with records description. It may even be possible to borrow from the research into natural language systems (also being conducted as part of artificial intelligence research) in order to construct an expert system which would respond effectively to questions posed directly
by the user, thus removing the intermediary role of the archivist altogether.

These are all visionary predictions of what expert systems may be able to do in the future. The most thoroughly explored and currently available expert system application in archives focuses on the retrieval of records. The National Archives has built a prototype expert system for reference inquiries using the commercially available Ml expert system shell. The system was designed to emulate the thought processes of an expert archivist in order to answer patrons' inquiries. Using as its test database a portion of the archives's holdings of Department of Interior records, the system first captured an expert archivist's knowledge of the nature of the originating offices, the content of the papers, the filing arrangements of government agencies, and the heuristic approaches the archivist would take in answering inquiries about the collections. It then used this knowledge to translate subject inquiries into references to specific record groups and series which were likely to contain material of interest.

The results were surprisingly good. In a test which, in response to a set of test questions, compared the number of relevant series retrieved by the expert archivist to the number retrieved by the system, the system performed credibly. While incomplete indexing limited the total system recall to seventy-four percent of the correct entries, the system did identify eight appropriate series which the archivist failed to pick up the first time through. 21 This suggests that the system may at least have a role to play as a memory supplement for expert archivists.

At least two elements of the expert system test at the National Archives are troubling, however. The first concerns the expert system's dependence upon the quality and completeness of machine-readable information available to it. An expert system can be no better than the data available to it. If, as happened at the National Archives, series are poorly described or improperly indexed, or if a rigorous thesaurus is not used in the database design and
construction, the system's rates of recall and precision will fall. The few library-based expert reference systems are similarly limited by the nature and quality of the databases which they can search.

Does this mean that to use an expert system to retrieve records archivists must begin to index thoroughly and to supply subject terms to their collections, a practice which many have avoided in the past? If one were to emulate the National Archives system, the answer would appear to be yes. A second approach, however, may be more feasible.

Since inventories and other finding guides are, to a degree, an index to and subject analysis of a collection, they could, if they were in machine-readable form, serve as the database which an expert system would interrogate. A simple and practical expert reference system for archives would consist, therefore, of an expert system shell coupled to a database of finding aids, most likely read into machine-readable format through the use of an optical character recognition (OCR) scanner. It is likely, however, that to employ expert systems in their institutions, archivists will need to rethink their descriptive practices radically.

Secondly, and also troubling, is the implicit standardization of archival practices which the development of expert systems engenders. The prototype system at the National Archives was designed to search a select group of records, and hence could be designed with that group of records in mind. One of the underlying concepts in the project, however, is that archivists have many practices in common. Furthermore, the designers of the National Archives system suggest that there are "general principles of archivey (sic) which apply across all record groups." Again the NISTF experience is helpful. While their studies suggested that there may indeed be general principles of archival sciences, they could not agree on them among themselves. Archivists are now faced with the possibility that the "general principles of archivey" will not be drawn up by their colleagues, but by systems analysts and programmers.
Discussions about expert systems for cataloging and reference use are proceeding apace in the library world, and the National Archives has designed a prototype expert system which could assist the naive user in navigating his way through the specialized arrangement of an archives. While there are no practical applications for archives currently available, expert systems do hold the promise of making automation cost-effective for an archives because of their ability to complement or replace professional, rather than clerical, duties. Hence, archivists would be well advised to experiment, if possible, with one of the relatively inexpensive microcomputer-based expert system shell programs to see if they can devise applications useful for their particular archives and collections.

If the promise of expert systems in archives is to be fulfilled, it is essential that archivists pay close attention to and participate in the development of artificially intelligent expert systems in related fields, such as library science. Failure to do so may mean that one day archivists will be forced to make do with an expert system knowledgeable about everything except archives.

Peter B. Hirtle is Curator of Modern Manuscripts at the National Library of Medicine. This article is based on a paper prepared for a seminar on archival automation led by Frank G. Burke at the University of Maryland. It was written by Hirtle in his private capacity. No official support or endorsement by the National Library of Medicine is intended or should be inferred.

NOTES

1 The history of archival automation efforts has yet to be written, though information on early developments can be found in Richard M. Kesner, ed., Automation, Machine-Readable Records, and Archival Administration: An Annotated Bibliography (Chicago:

2 Lytle, "NISTF," 361.


4 Smith, "Artificial Intelligence," 69.


6 Waltz, "Artificial Intelligence," 119.


8 The reverse of this process, called (naturally enough) "forward chaining," is also used.

9 One of the conclusions reached by the NISTF committee was that in spite of the seeming anarchy in the archival profession, there was a surprising degree of similarity in processing, and in particular in the


14 Davies and James, "Cataloguing," 286.


22 Ibid., 17.

