ARCHIVAL INFORMATICS NEWSLETTER

Part 1 of Archival Informatics Newsletter & Technical Reports

ISSN 0892-2179 Volume 1, # 1

FROM THE EDITOR:

The first issue of any new publication calls for explanation. The simplest one is almost certainly the best - I felt a real need among archivists to discuss automation and its impact on their profession and I recognized a lack of vehicles for this dialog. After leaving the Smithsonian last fall, I found myself building a database of archival automation news which I knew had not been available to me when I was Deputy Director for Information Resource Management at a very large, public organization. It was obvious to me that it was not available to others either. Then I found how many terrific applications have never been reported, how many studies, such as the report on archiving of Database Management Systems by Tom Brown and Bill Reader in this issue, cannot find their way to print, and how many thoughtful discussions, like Barbara Cain's treatment of indexing issues in the North Carolina State Archives system manual are buried beyond the reach of colleagial review. I hope the Archival Informatics Newsletter attracts your good ideas and stimulates a vigorous dialog. As always, I will provide lots of my own ideas to shoot at! Please send in snippets or diatribes, reviews or news of your own.

But a newsletter cannot support some of the needs of the community; some extremely important technologies need to be examined in greater deatail, and reported from an archivist's perspective for the use of archives administrators. This issue of the Archival Informatics Newsletter is part 1 of Archival Informatics Newsletter & Technical Reports. This quarter, part 2 is an extensive technical report on optical technologies and their significance for archives today. It is intended to be extremely practical and directly usable - I leave to your judgments whether it works, but I'm sure such technology forecasting and pragmatic advice are needed. Future topics for the Technical Reports include archiving of software, artificial intelligence, requirements for automated archival control systems, and the use of information technologies in collections management and tracking. There are many other topics I would like to research myself, but I would also like to invite others with special areas of technical knowledge to propose future studies and author them as part of the Technical Report series.

David Bearman, Editor

TABLE OF CONTENTS:

AUTOMATED SYSTEMS:

by David Bearman

INHERITANCE OF INDEXING TERMS IN FAIDS

The North Carolina State Archives has recently implemented a system for holdings location and retrieval, information retrieval from finding aids and limited administrative processing control functions on the Prime 9655 mini-computer it shares with the rest of the Department of Cultural Resources. The system runs on PRIMOS, using Prime Information and Blacksmith Intercept software and an in-house data entry program. The system uses 34 essentially standard data elements and plans to produce MARC formatted output. Currently it is implemented for staff use as the database (18,158 records with an average length of 205 characters in early March) is being built. Eventually it is intended to serve as an on-line public access catalog and will implement authority control over subjects consistent with LC headings as far as is possible.

At some time in the future, I hope to review the software, but here I will discuss a number of issues relating to indexing of archival material in hierarchically structured files which North Carolina has attempted to resolve. Like many other archival systems, the database architecture implemented by Dr. Arlon Kemple, Computer Specialist for the Department, in response to requirements defined by the Archives Committee on Automated Finding Aids, " reflects the organization of records by maintaining an explicit hierarchical relationship between record group, series, box, folder and item, linking information entered for the higher level with that entered for the lower levels".

This method of hierarchical linkage, in which all records have the same structure regardless of level but each is assigned an identification number which indicates its structural location within the overall holdings of the institution, is now widely adopted. The structural notation is an automated record relationship manipulation device only, and bears no relationship to 'call' numbers or stack locations.

The North Carolina staff identified a variety of advantages of using such a system in its staff manual on the system:

" information displayed about a document or map or photograph at the lowest level will always include the provenance of that document....

... it also has great flexibility, allowing description to whatever level is best for the specific material -- item level for photographs and maps, box or folder level for most state agency and county records...

... one can browse the structural arrangement of the Archives...request names of record groups or collections that have been entered...a listing of the series in a particular record group, or the boxes in a particular series, or the folders in a box, or the items in a folder..."

In a section of its documentation entitled 'Indexing Rules', the Archives discusses its approach to exploiting hierarchical inheritance:

"Because the archival levels are linked in FAIDS, index entries for material described at one level are always applied also to the lower levels contained within that unit. If a personal name, corporate name, geographic name, or subject is basically involved in all the records or papers in a record group or collection, or in a series or a box, even if it is not actually named in every folder, the index entry should be made at the appropriate higher level. Orange County would be entered at the record group level for Orange County records, and Ashe, Samuel A'Court would be entered at the record group level for the Samuel A'Court Ashe papers. If the name or subject is searched for alone, the reference given would be the entire record group or collection. If it is entered as part of a Boolean combination search, it will match with index entries made for every computer record at every level below it. Orange County + Wills would match at the series level. Orange County + Wills + John Doe would match at the folder or item level.

a. In such a system, the computer is bound to come up with unwanted references from time to time. If Education and Schools are entered as subjects for the State Board of Education Record group, a search for School + Pender County is going to produce references to swamp land in Holly Shelter Swamp, since Swamp Land Records is a series within the Board of Education record group. The researcher will usually know from the title of the reference that the record is not of interest to him, but if in doubt, he can call up the full description in FAIDS.

b. To reduce the number of inappropriate references produced, care should be taken to index each entry at the best level. For example,

it may be better to enter 'Schools' at the series level instead of the record group level."

Because the experience of North Carolina is of general interest and the instructions they provide for their indexers are so clear, it is useful to use their discussion as a point of departure for a critique of index term inheritance issues in archives.

Although other software environments, especially those used in artificial intelligence systems employing semantic networks with frame representation (which inherit by exception), can permit this issue to be finessed, most archives would consider themselves fortunate to have a system which inherited through the relatively mechanical hierarchy described by North Carolina and which supported a query which used both assigned and inherited terms. What issues does this raise?

First, context is always relative. In North Carolina, the terms North Carolina is not entered, but Tennessee is when the records have that subject. If North Carolina was entered, the United States would not be. Inheritance permits the assignment of broader terms at higher levels and narrower terms at lower levels, but it does not serve as a thesaurus. This is a problem, both because of what we tend not to index (N.C., USA) , and because of what it requires us to index. A true thesaurus would be a valuable aid here.

Second, archivists have a bad habit of deriving guide entries in which some terms from different levels of the description are featured in an abbreviated description, not because they are important to the collection so much as for their importance to the audience. Thus, we have many guide entries which index by Presidents of the United States or by important events, even though these are relatively minor features of the collections. If this practice is deemed uniformly misleading, using only terms from higher in the hierarchy would be an improvement, but how do we identify a level from which such terms will be drawn so as to distinguish between holdings?

Third, the primary advantage of searching both assigned terms and inherited terms is to reverse the effect of indexing by the most specific available entry and assigning such terms at the lowest level to which the analysis procedes; this effect is term scattering, which can only be allowed in authority controlled indexing with term hierarchies. At the same time, we must be careful that by permitting assignment of broader terms at higher levels, many general terms do not rapidly become useless for retrieval because of the numbers of items with which they will be associated.

At the core of the quandry is an extremely valuable indexing precept - that the narrowest possible term which adequately reflects the content of an item should be assigned at all times. As North Carolina has discovered, inheritance permits the archivist to follow this useful guideline by assigning narrow terms at lower levels except in the case of non-homogenous groups (swamps in school records) when departure from the guideline again risks overwhelming the system with too many occurences of certain terms.

What the North Carolina system introduces by indexing hierarchical record levels is equivalent to maintaining a hierarchically controlled vocabulary so that Holly Shelter Swamp will reference from Orange County without an added explicit reference at each level in the record. Nevertheless, they have found a middle ground which seems viable. I would like to invite readers with other solutions to the overloaded indexing problem, and other methods of supporting narrower and broader searches to write.

WHY ADOPT MARC?

For several years now I have met archivists who ask me why the archival community should adopt MARC. I almost invariably give three answers which have nothing to do with sharing information. I would like to repeat them here, if only to stimulate discussion.

First, we should adopt MARC because it is an insurance policy for automated data. Other formats in which we might store the data are non-standard (witness the ART Task Force report on DBMS's in this issue). If we want to be certain that we will be able to move the data we laboriously compile about our holdings from one automated environment to another, we need to be able to read it into and out of a format which allows for "migration" between hardware vendors and software systems.

Second, we should adopt MARC because the AMC format embodies standard archival practice, and as a very small profession our best chance of being able to leverage sufficient investments to get a reasonable variety of software products from which to select something which meets our needs, is to be able to provide developers of such products with a standard set of data they must address.

Third, we have an opportunity to disseminate

information about our holdings through a vast array of other agencies simply by being able to output it in MARC format. If we had developed a new food product and wanted to reach the marketplace, we would adopt barcoding on our packages because the largest grocery chains would not sell our commodity unless it was barcoded. As archivists, if we want to reach the information dissemination marketplace, we need to use the facilities of the national networks and database vendors. To refuse to place MARC on our packages is akin to not recording barcodes. We will find ourselves standing on a corner with a small vending stand, wondering why we haven't reached a larger audience.

What is noteworthy about this list of arguments is that none of them hinges on any assumptions about archivists sharing, or not sharing, data with each other. None of them depends at all on the realization of national archival information systems. And by themselves, none of them suggests anything about the promise of MARC for cooperative documentation or cooperative collections management . It is not irrelevant that some of the data we want in our archives - like state publications cataloging or newspaper cataloging - is already in MARC formats which we can't integrate unless we can also read MARC into our systems, but the three arguments aduced before remain the strongest reasons to adopt MARC, and don't alude to this reality of the information environment either!

ARCHIVAL PROJECTS USING OPTICAL DISK:

In the first quarter of 1987, several major archival projects employing optical disk moved off the drawing boards. The attention of the traditional archives community was attracted by a nearly $1 million contract awarded by the National Archives to SDC for an optical system to store over one million military service records. Although the award culminates a number of years of study, the Archival Research & Evaluation Staff still considers it a pilot project. In February 1984, that staff first recommended digitizing pension, bounty land and military service records (over 8% of NARA holdings) and has been assessing the technology since.

Meanwhile, two larger projects have boldly committed to operational optical disk systems. The University of Syracuse has been awarded a grant of $3.7 million from the W.K. Kellogg Foundation to convert 650 linear feet of adult education materials to an optical resource. The archive includes records of most of the associations which have been active in the adult education field for the past 50 years, plus audio and video tapes, photographs and the papers of prominent academics in the field. Project Director Roger Hiemstra reports in the project newsletter that QUALOG, a logic programming interface developed at Syracuse and now in use to assist in qualitative analysis of text materials, will permit users to "ask a question and get a new response, ie. one that was not specifically programmed into the computer."

On an even larger scale, the University of Pittsburgh is proceeding with the digitization and optical storage of full text data bases, image bases and clinical data on all patients in its hospitals. The system will store all the data collected during the course of patient stays on 500,000 patients, or approximately 10 million records and images. Using partitioned data/image sets driven by microprocessors attached to optical drives, the Vice President for Biomedical Informatics, John K. Vries, anticipates a system equivalent in power to a supercomputer for a cost of several hundred thousand dollars. The Medical Archives System (MARS) employs full-text automatic indexing with artificial intelligence driven, thesaurus assisted indexing to capture medical dictation, directly word processed records and scanned digitized text. A retrieval system employing Boolean operators and artificial intelligence parsers to express user queries in the same canonical forms stored during indexing, assures that users can use natural language and get responses which are highly specific. Most of the system is now implemented, and having seen it, I am certain that we are looking at the next generation of archives.

While numerous integration issues and virtually all the standards questions remain to be resolved, the rate at which the technology is being adopted for official records keeping virtually assures that the legal issues relating to optical records will be resolved (by the best available evidence rule?) in favor of the new technology. The IRS is, after all, keeping tax returns in this format. In Florida, the Department of Corporations is implementing an optical disk for storing filings of Florida businesses. New South Wales, has just given the go ahead to a $5.8 million registry of births, deaths, and marriages which will transfer 30 million records to an optical disk system.

MACHINE-READABLE: VIEWS

by Thomas Elton Brown, PhD.

[Editors Note: Tom Brown, Archivist, Administrative Staff, Office of the National Archives has agreed to be a regular contributor to the Archival Informatics Newsletter. His column will feature issues which he feels are important to archivists concerned with machine-readable records. I am delighted to have Tom reporting on a field he knows so well, and taught most of us.]

KENTUCKY MACHINE-READABLE RECORDS PROJECT:

A draft of a report on the Kentucky Machine-Readable Records Project (KMRRP) has recently been completed. The project was an effort by Kentucky's Department of Libraries and Archives to gain archival control over the current automated information systems in the state government. The excellent report, written by Margaret O. Adams (who has since joined the Machine-Readable Records staff at NARA) is titled Managing State Government Computerized Public Records.

As archivists have attempted to cope with the growth and increasing technical complexity of current automated records systems, they have moved away from inventorying computer files as first proposed by Everett O. Alldredge in 1971. In its place, three approaches have been developed. These are: 1) microcontrol which manages automated systems at the data element level; 2) system component level description to control the different record components within the automated systems; and 3) a capsule approach which develops inventories of abbreviated descriptions of automated systems.

When the Kentucky Department of Libraries and Archives proposed to develop a data dictionary to manage the state's computerized records, it was pursuing a strategy of microcontrol. The initial concept of the data dictionary was an automated system which would track the data elements in Kentucky's automated systems. As this report documents, the project has moved away from the original concept of the data dictionary. While nothing in the report invalidates the microcontrol approach, the report candidly acknowledges the practical problems which Kentucky faced during the project and which led them away from the original concept of the data dictionary. Without a common understanding of what was the core of the project, it is not surprising that communication was hampered. Another recurring theme concerned who would provide information for input into the data dictionary.

The report concludes with two recommendations. First, the author argues that records officers should become "information managers" who would manage the information in an agency and a resource of that agency. Part of their responsibility would be to promote the use of the agency's information by other state agencies and, if appropriate, release it to the public. Secondly, the report urges that the Kentucky Department of Libraries and Archives establish a Kentucky State Government Repository and Service Center for Machine-Readable Records. This would be a records center for computer files in which providing reference service to other agencies and to the public would be stressed. It would provide an opportunity for the archivists to gain experience in processing, describing, preserving and referencing automated materials.

Clearly, these recommendations raise two critical points for the archival profession in dealing with automated records. First, archivists and their colleagues in records management need to become more concerned about controlling the information in the records rather than controlling the physical medium of the record. Secondly, most archival institutions have not addressed the problem of establishing records centers storage and procedures for computerized information. But, as the problem of non-current records which are not eligible for disposition grows, various solutions will be proposed. The report on the Kentucky project offers one of the first concrete proposals.

While raising two important issues, these two recommendations reflect an underlying theme to this report. Namely the driving force behind an archival program for automated information systems should be reference service. This stress on reference service can at times give rise to troubling implications. For example, the release of some information may be prohibited because of privacy considerations or other legal restrictions. In this situation, emphasizing reference may result in archival control missing such restricted information. Or a high degree of current research interest may be transitory, and the information which is the object of such interest may have little or no long term value.

Conversely, information with significant long term value may have little or no current interest. The emphasis on reference causes the resources to be devoted to the information with high current interest regardless of long term value. Such an allocation of resources may be inappropriate for an archival institution.

FEDERAL GOVERNMENT ISSUES REPORT:

Archivists responsible for the administration of computerized records frequently find valuable information in obscure and unusual places. A recent Federal report is a case in point. The Office of Management and Budget, the General Services Administration, and the Department of Commerce jointly produced A Five-Year Plan for Meeting the Automatic Data Processing and Telecommunications Needs of the Federal Government: Volume 1 (September 1986). Despite the unlikely title, the report contains two discussions very relevant to any archivist concerned with automated information.

The first item of interest is the report's discussion of the current state of technology and its forecast of future developments. This technology overview appears in detailed discussions of several information systems projects and in a brief review of agency information technology initiatives which lay the basis for a section entitled, "Forecast of Technological Trends". The report later presents an extremely lucid account of fourth generation languages and data dictionaries. Clearly, the applicability of such information is far wider than just the Federal establishment.

The second discussion of importance to archivists is a section entitled, "Information Management: The Impact of New Technologies." The report argues that the creation, maintenance, and preservation of written documentation is critical for operational and social needs. Claiming that information management must ensure the existence and availability of this information, it outlines the management principles through the information life-cycle: creation/collection, maintenance, retrieval, and disposition or dissemination. These principles are then discussed in terms of the issues raised by three new technologies: word processing, electronic mail, and electronic collection and dissemination of information. Thus the report succinctly analyzes how automated processing of full text information impacts the life cycle of records.

DATA FILE DESCRIPTIONS ENTERING NATIONAL DATA BASE:

The Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan and the Research Libraries Group Inc. (RLG) have agreed to include the description of approximately 1,300 archival collections of machine-readable data files in the Research Libraries Information Network (RLIN). ICPSR manages what is probably the largest collection of machine-readable social data in archival custody. Placing these descriptions in a national bibliographic network will make scholarly researchers aware of the existence of ICPSR's numeric data files and of the importance of computerized information in general.

The project will begin with a machine-readable version of ICPSR's repository Guide, currently maintained in a local SPIRES data management system. ICPSR will output a SPIRES file with descriptive information on the data collections, and RLG will convert the SPIRES record to a MARC form. The MARC Format will be the one specified for machine readable data files - MARC -MRDF. The University of Michigan Library will then add appropriate MARC-MRDF field tags not provided in the conversion, add Library of Congress Subject Headings, and assign other entries in accordance with the Anglo-American Cataloging Rules. This retrospective conversion is projected to be complete in August. As ICPSR accessions new files, descriptions of them will be added to RLIN.

STORAGE TECHNOLOGY & REFERENCE SERVICE:

Archivists at traditional machine-readable data repositories have fairly well established appraisal criteria, processing guidelines, descriptive standards and preservation procedures. However, most archives are still using reference procedures established in the late 1960's by data archives and data libraries. The current standard seems to be to provide a copy of the data file in a fixed length, flat format on magnetic tape encoded to the researcher's specifications. As technology has progressed, data libraries and data distributors have developed other options which archivists have not adopted. For example, the Census

Bureau and the National Technical Information Service provide selected information of floppy diskettes. Others provide interactive, on-line access to selected numeric information. Reference strategies for computer files, especially full-text information, will hopefully become a major research area at traditional repositories as they examine the possibility of expanding their reference service options. Presumably, an initial step would be to examine options which data libraries and data archives are presently offering.

In this connection, the Data Archives on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP) has announced a new service using the Bernoulli Box. This is a small, portable device that connects to a microcomputer. It reads and writes information on magnetic cartridges about twice the size of a floppy diskette. Physical size, however, is deceiving. Where an IBM-PC floppy diskette can hold up to 360,000 bytes of information and an IBM-AT floppy can hold up to 1.2 million bytes, a single Bernoulli Box cartridge can hold up to 20 MB. The discount price of a Bernoulli box with two 20MB drives is approximately $2,600. Once the initial investment is made, removable 20 MB cartridges can be obtained for only $65 each. DAAPPP is now offering all of the data appropriate for microprocessors on three 20MB Bernoulli cartridges for $750, or 40% less than the floppy diskette versions. Included are the machine-readable raw data and SPSS/PC program statements with printed documentation.

Clearly the development of low-cost, large-volume storage media is changing the options archivists have in providing computerized information to researchers.

Thomas E. Brown

ART TASK FORCE - LEARNING OBJECTIVES:

The SAA Automated Records and Techniques Task Force has updated its "Learning Objectives and Course Content" Guidelines for courses on Basic Computer Concepts, Automated Techniques and Machine Readable Records. As of October 1986, "archivists have used this curriculum for the basis of 32 workshops during annual SAA meetings." In addition, the SAA Task Force on Automated Records and Techniques has used the curriculum to present a two day workshop on machine-readable records on four occasions and a two day workshop on automated techniques, twice. The Guidelines are reprinted here without part A, the computer literacy course.

An Educational Program to Train Archivists in Automation:

B. Automated Techniques

1. The archivist will understand how an automated system can organize and manipulate information.
2. The archivist will learn how to determine if the implementation of an automated retrieval system using information in manual systems will produce desired outputs.
3. The archivist will understand the principles of the decision making process regarding the acquisition of an automated information system.
4. The archivist will create an exercise database on a microcomputer using commercial database management software.
5. The archivist will outline procedures to evaluate, maintain, and expand an operational database.
6. The archivist will understand how to integrate information in an automated information system and the subsequent impact on providing reference service to archival materials.

C. Machine Readable Records

1. The archivist will learn to inventory the components of and then describe automated information systems.
2. The archivist will determine the informational and evidential value of a machine-readable file through an analysis of a) units of analysis, b) the level of aggregation, c) the difference between administrative and survey data, d) linkage potential and, e) the relationship of similar information in hard copy format.
3. The archivist will learn to develop records control schedules for automated record systems based upon information gathered during the inventory and the decisions made during the appraisal.
4. The archivist will determine if sufficient documentation exists to accession a file into archival custody.
5. The archivist will determine how to process machine-readable records in order to make them available for research.
6. The archivist will understand the types of information needed to describe machine-readable data files and how to use standardized formats in describing machine-readable data files.
7. The archivist will discuss the dissemination of files with restricted information.
8. The archivist will discuss the research communities for machine-readable data.
9. The archivist will discuss preservation techniques to ensure the integrity of machine-readable files in archival custody.
10. The archivist will discuss trends in computer and information systems technology.

WHAT ARE/IS INFORMATICS?

And especially, what/who is Archives & Museum Informatics?

In November 1986, I formed Archives & Museum Informatics as a sole proprietorship, which, as the State of Pennsylvania so aptly puts it, means that, at the moment, Archives & Museum Informatics is a "fictitious name" for David Bearman. Perhaps, someday, others will be involved as well.

Archives & Museum Informatics was formed to conduct research and publish. I adopted the term "Informatics" from an emerging usage in biomedicine where the importance of information technologies (like computers), information techniques (like full-text retrieval or digitizing radiographic images) and information theories, especially those of linguistic analysis, artificial intelligence, indexing and retrieval, are coming together in new ways of practicing medicine. The term, while relatively new, is used enough so that in this issue of the Newsletter, without my making any effort to do so, it appears in the title of one (other) journal and the title of a person's job. To me the importance of the concept is that is replaces automation, or records, or computerization with a system oriented view of the synergism of information based activities.

Archives & Museum Informatics is also a consulting firm, and the term "Informatics" in the name expresses a philosophy of looking at the cultural information missions of archives and museums broadly in order to address a range of new approaches, techniques, and technologies which can enhance an organizations profile and achieve its mission.

In addition to publishing and consulting, I have developed courses to teach a number of skills which I believe are overlooked by college programs which are available to professionals in their own communities and by the specialized training programs thus far offered by SAA, AASLH, ARMA, NAGARA or the museum associations. These skills range from Planning for New Technologies (a course for managers which focusses on risk assessment, needs analysis, standards, introducing change in organizations etc.), to Understanding Vocabulary Control (a course for archivists involved in record description and/or reference, which focusses on the kinds of vocabularies which can be controlled and effective tactics for controlling them.) A course on Planning for Automation which presents each step in an automation project and gives hands on experience in forms analysis and writing of functional requirements is designed to help form a team in an archive in which a decision "to automate" has been made but the implications of the decision have yet to be worked out.

Originally I had intended to give all these courses as public seminar/ workshops, but the costs to participants of travelling to Pittsburgh has convinced me to offer them instead under state archives, regional organizations, or universities sponsorship where I have to travel, but many of the participants are near home. I welcome inquiries about these and other seminars.

Over the past three months, Archives & Museum Informatics has been involved in a number of projects for non-profit organizations which I will report in a future Newsletter, including:

- Mapping a proposed data dictionary for description of Architectural Drawings into a MARC format for possible use by the J.Paul Getty Art History Information Program.
- Developing Policies & Procedures for the Establishment of a Software Archives by the Computer Museum (Boston) in conjunction with the Smithsonian Institution. This is phase two of a project, the first phase of which was a feasibility study submitted to the sponsors in February.
- Designing a method, for the National Security Archive, of indexing at an item level very large numbers (30,000+ p.mo.) of recently declassified documents relating to American Foreign Policy.
-Assisting the Florida State Archives staff to develop an automation plan, by conducting a workshop with them on automation planning.
- Critiquing and analyzing terminology proposed for inclusion in the Art and Architecture Thesaurus.

Although Archives & Museum Informatics is clearly a commercial proposition, I hope that it can also make a contribution to archives and museum automation through conducting kinds of research activities which cannot be supported by cultural institutions acting on their own. One aspect of this effort is to encourage contributions by others to this newsletter in order to increase the speed of flow of technical information within our community. Another aspect is to ask for your ideas for future research and publications, and to provide means to pool resources in order to conduct such studies. I look forward to hearing from you.

THE ARCHIVAL MANAGEMENT OF MACHINE-READABLE RECORDS FROM DATABASE MANAGEMENT SYSTEMS:
A TECHNICAL LEAFLET

By Thomas Elton Brown & William A. Reader
Society of American Archivists, Task Force on Automated Records & Techniques

INTRODUCTION:

Data Base Management Systems (DBMS's) pose problems for records managers and archivists. This technical information leaflet addresses these problems. It begins with a discussion of archival issues associated with DBMS's. Since most of these concern the transfer of DBMS- managed information to an archives, another section recommends transfer procedures for specific DBMS packages based on a survey of vendors of DBMS packages. Vendors were asked to recommend a means by which information in one of their systems could be transferred to an archives. These procedures are presented here so that an archivist or records manager can convey these instructions to the database administrator of a system which contains historically valuable information. Using the recommended procedures, the database administrator will then be able to overcome the technical difficulties associated with transfer of archivally valuable DBMS information to a repository. Since the universe of DBMS software packages is constantly changing, the leaflet also announces a clearing house function. The purpose of the clearing house is to maintain the currency of the leaflet information by providing a point of contact to correct, refine and supplement the recommended procedures for transferring information from specific DBMS packages to an archives.

I. Archival Problems Created by Data Base Management Systems:

In recent years, organizations generating archival data have created databases and installed database management systems at a phenomenal rate. Unlike conventional archival tape files which are software independent and can be processed on any computer system, DBMS-managed databases organize their data in ways that confer benefits for users but create serious potential archival problems.

There are several advantages to placing data in a database. One does not have to access the entire database in order to use just some of the data. Thus a bank posting check transactions needs only to access checking accounts and does not have to read the other files in the database. This cuts down on processing time, permits simultaneous processing of multiple applications and can limit access to sensitive data areas of the database to only certain users. Without a DBMS, applications programmers must know the location and structure of the data when writing a program to access or process it. With a DBMS, programmers need only know the names of the data elements and thus can focus on turning out applications programs. All this, in turn, cuts costs and improves operating efficiency.

In a DBMS-managed database, data elements and records are physically delimited by control characters embedded in the data. Records consist of either physically separated data fields or data elements in different tables which are linked into a logical record by pointers or database keys. Records, in turn, are logically organized into sets which, according to the type of DBMS, consist either of identical record types defined by database keys, pointers which locate the record's constituent elements, or hierarchically linked record defined by pointers tying one record to another.

These DBMS characteristics, however, have practical implications which cause potential archival problems. First, in a DBMS-managed system, there is no correspondence between the physical and logical order of the data. Such relationships must be created either by the DBMS software, which maintains definitions of data elements, records, sets and linkages, or by the routine that dumps the data into an archival file. Second, different generic types of DBMS's give very different definitions of data elements and records, provide different means of linking data and records together to provide data to the user, and supply different delimiters and pointers to indicate the specific physical location of data fields, records and table elements. In addition, even within one class, software-defined data fields, record delimiters, and keys may differ from specific system to specific system. Consequently, data managed by one DBMS cannot be processed by a computer using a different DBMS.

Recently, the problem of DBMS data transportability

has been aggravated by computer and DBMS manufacturers embedding certain recurring software operations on special purpose computer chips to free up computer memory for other instructions and data. As more software becomes embedded in computer hardware, the data stored in DBMS systems becomes hardware specific.

Thus, for an archival institution concerned with the permanent preservation of historical data, the archiving of database information presents a substantial problem. Specifically, how does one transfer historically valuable information resident on a specified DBMS to an archives? Traditionally, archives have sought to avoid becoming repositories of obsolete hardware and software systems because they have neither the funds, expertise nor space to acquire and maintain antiquated computers and programs. Consequently, archives have acquired machine-readable data in flat fixed-record length formats free of any extraneous control characters or delimiters.

To follow this practice for DBMS data, archivists must determine whether a DBMS has an "unload" routine that will output all or selected fields into linear sequential files without any embedded control characters. Unfortunately, many DBMS's have "unload" routines which correlate physical and logical order, but leave extraneous control characters which can be read only by the DBMS that unloaded the database. Thus, their output is only transportable to another computer using the same DBMS. To make the unloaded data transportable, one has to write a program to reinterpret the control characters, and, consequently output the information in a software independent format. This programming would be tedious, time consuming and expensive.

Even if the creation of an archival, linear, sequential, and fixed-length record is possible, there is often a related question of practicality. Since DBMS's are structured to minimize duplication of data elements, the creation of linear sequential files from a DBMS will result in numerous separate files or a massive linear sequential file in which data fields are constantly repeated. For example, a DBMS containing a sample survey of cohort data would have separate files of geographic area data, household data, person data, and separate time period interview data for each person. Creating linear sequential files could necessitate separate sets of geographic, household, person and interview files. Or it could mean the creating of an interview file incorporating the data elements from all the files. In this case, geographic, household, and person data common to each separate interview would be repeated for each interview record. This would result in a large quantity of redundant data.

The extent of the archival problem with a DBMS is dependent on the paricular software package that manages the information. To examine the issue in terms of specific DBMS's, the Society of American Archivists Task Force on Automated Records and Techniques surveyed vendors of Data Base Management Systems. The goal of the survey was to determine what procedures the vendors would recommend for transferring DBMS-managed information to an archival repository.

II. Survey of Vendors:

To determine which DBMS vendors to survey, we reviewed Datapro Research Corporation's In-Depth Analysis of Leading Database, Data Management, and Data Dictionary Systems (1985) to identify vendors with a large number of installed systems. Letters were sent to those vendors in June 1986, asking how information being managed by their DBMS product could be transferred to an archives. The letter invited vendors to propose any or all of three possible strategies: output of a flat, rectangular format free of extraneous control characters; use of an international standard designed for exchange of information between systems (ISO-8211); or any other means the vendor would recommend.

Initial responses varied greatly in quality. As a result, follow up letters or telephone calls were used to clarify the initial responses. Tentative results were presented at the Society of American Archivists annual meeting in September 1986 and follow-up contacts continued throughout the fall.

It is noteworthy that no vendor reported implementing ISO-8211, the International standard for the exchange of data files. Among those vendors who recommended any specific procedure for transferring information from their system to archives, almost all focussed on techniques for creating a linear, fixed length data file on magnetic tape.

Nothing in this report should be construed as an endorsement of any particular systems by the authors, the SAA or the publisher. Furthermore, the exclusion of any particular system does not mean the vendor does not have a large number of systems in use; vendors with numerous installations may not have responded to the initial letter and a lack of time may have prevented follow-up. The instruction for transferring information from a specific DBMS to an archives may need correction. The recommended procedures are based on discussions with the vendors' representatives, review of user manuals or interpretations of letters, any of which can be misinterpreted. In addition, no tests of the vendors recommended procedures have been made to independently validate their claims. Because of the potential for misinterpretation, because no tests have been run, and because vendors are continually adding to and altering their products, these procedures will

need updating. Also, this technical information leaflet does not address all of the DBMS's which archivists and records managers will confront and will need to be updated as more DBMS's are developed and encountered and as archivists and records managers validate these procedures in their experience with different systems. To correct or expand on this report, archivists and records managers are invited to contribute to a "Clearinghouse" by contacting the authors, c/o ARCHIVES & MUSEUM INFORMATICS, 5600 Northumberland St., Pittsburgh, PA 15217.

III. Vendor Recommendations:

APPLIED DATA RESEARCH, Inc.
ADR/DATACOM/DB

This DBMS package provides a single utility, called, DBUTLTY, for creating and maintaining the database. It includes two routines which have applicability for creating output capable of being transferred to archival custody.

The first routine is BACKUP. While the primary purpose of this routine is to create security back copies of the database and its index, careful selection of options relating to DATACOM's "keywords" can produce a variable length, sequential file which contains standard IBM four-character-length identifiers. Specifically, the DBUTLTY program should include the following command formats: CMPRS=NO and SEQ=NATIVE. Since this routine can INCLUDE or EXCLUDE specific tables or files from the database, the execution of one routine could output all of the information of archival value from the database.

The second routine is EXTRACT. This routine writes each table as a fixed-length record in fixed-block format. Extract the table or file in a native key sequence (SEQ=NATIVE). Since only one table or file can be extracted by this routine, a separate execution is required for each table with information of archival value.

CULLINET SOFTWARE Inc.
IDMS/R & Culprit

Culprit is the report generator associated with IDMS/R, the Integrated Database Management System/Relational. Approximately 95% of installed IDMS/R packages have Culprit. Through the use of Culprit, a portion of the database is specified, then a specific database, then the data, and then the data to be written to magnetic tape. This is known as SELECTIVE COPY. The data will be written to tape in a decompressed format and will be accompanied by a "listing" of the copied data elements. Since the data is decompressed, this listing serves as a record layout. This process requires the selection of each data element for transfer.

Infodata Systems Inc.
INQUIRE

This DBMS is primarily for text rather than numeric information. It can unload data files into a sequential file format through the SAVE command. It selects particular data fields from the database for output, beginning in a specific field location and in a linearized format. If outputting text information, the return characters could be interpreted to output the information in paragraph format or could be suppressed to output the information in a long sequential string. This procedure requires the selection of individual data elements. Another possibility is to dump the information into Infodata Standard Input (ISI) format which is a field delimited file in which field names appear in columns one through eight of an eighty column card image and the value of each field wraps between columns ten and seventy-two. To employ this utility, execute SAVE ISI. The system prompts for the database name, then asks for record description to select certain records from the database, and then unloads that database to a sequential file in ISI format. Since the file is first put out to disk, one would have to use the operating system utilities to move the file to tape for archival storage.

Information Dimensions Inc.
(Formerly Battelle Memorial Institute)
BASIS

This DBMS is primarily for text rather than numeric information. The package includes at least ten optional modules. One of the more commonly installed modules is REPORT. Within the module, REPORT WRITER specifies each field and the field length. In the execution, the data definition is automatically provided for a complete codebook. If one is dealing with only numeric data, the REPORT module can define an output file in the format called for by a statistical analysis package such as SPSS or SAS. In this situation, the BASIS data could be moved as an SPSS or SAS file. Then the utilities within those statistical packages could produce a sequential, fixed-length, file.

I.B.M. Corporation
IMS (Information Management System)

IMS is one of the oldest DBMS's; the first installation was made in March 1971. Each IMS system has a LOAD or UNLOAD utility for use in maintaining and compressing the data in the DBMS. However, using UNLOAD will produce a sequential file embedded with IMS control characters and thus still dependent on the IMS software. Therefore, IMS does not normally have a

utility which can create a rectangular output file free of control characters. Since 1984, IBM has been marketting a relational database system called DB2. In order to permit IMS users to migrate to DB2, a DATA EXTRACT utility was created for IMS to produce a VSAM format of IMS data. The use of the DATA EXTRACT utility allows a choice of transferring a complete table or just portions of a table or file. The DATA EXTRACT utility requires the user to specify the path through the data hierarchy. Since a VSAM format can be moved by IBM system utilities to a rectangular tape format, the use of this utility provides the intermediate step for the creation of a rectangular data file. However, very few IMS installations will have acquired this utility, since its purpose is migration to DB2.

I.B.M.Corporation
DB2

Since the DATA EXTRACT utility is the DB2 approach to maintaining and compressing information, almost all DB2 installations will have it. This utility will output from the DB2 software a VSAM format of DB2 information. The use of DATA EXTRACT allows a choice of transferring a complete table or just portions of a table or file. The DATA EXTRACT utility allows the user to relate different files in different combinations. Since a VSAM format can be moved through IBM system utilities to a rectangular tape format, DATA EXTRACT provides the intermediate step for the creation of a rectangular file.

Oracle Corporation
Oracle

EXPORT is an Oracle database file backup facility which produces an encrypted, compressed file. Users may send flat sequential datafiles from SQL*PLUS onto tapes. Oracle also contains a standard utility called ORACLE DATA LOADER (ODL) to read and store flat sequential ASCII or EBCDIC files, but apparently does not have a routine for the reverse process.

SAS Institute Inc.
SAS (Statistical Analysis System)

Originally designed for statistical research for academics, this package is increasingly used for general data management in other organizational settings. The package consists of collections of instructions, called steps, and of individual instructions within steps called statements. To create a rectangular file (or "raw data file" in SAS terminology) from a datafile in the SAS system (or a "SAS dataset" in SAS terminology), one would use the DATA step. With the DATA step, two statements are necessary. The first, is FILE, a statement which will specify the format of the output file. A FILE statement can have different options. Whether and how these options are used will depend on how the information in the SAS dataset is structured, e.g. a hierarchical file. The second statement is PUT. The options with the PUT statement can specify whether the entire file or just selected variables within a file will be moved to the output file. The vendor promotes SAS as self-documenting. As a result, a record layout and codebook can be created using the CONTENTS statement in the PROC step.

SAS Institute Inc.
System 2000 DBMS

This DBMS does not have a utility program through which a sequential rectangular file can be created. After SAS Institute Inc. purchased System 2000, about 2 years ago, a procedure was developed by which a System 2000 hierarchy can be incorporated into a SAS dataset. The routine is PROC S2K. When using this procedure, it is necessary to specify the path or paths through the hierarchy. With the creation of a SAS datset, one can use SAS procedures to produce a linearized, sequential file.

Software AG of North America, Inc.
ADABAS

This DBMS does not have a routine to use in the production of a fixed-length output file. However, over 90% of ADABAS installations have a programming language called NATURAL. A Three statement program in NATURAL would produce an output file. A READ statement would specify the file or tables; a KEY statement would identify the variables being selected; and a WRITE statement would stipulate that the output is to be a fixed-length record. The KEY command must be followed by the database names of the variables which will be transferred to output, thus the system permits only the selection of individual variables rather than complete tables or files. In the 10% of ADABAS facilities which do not have NATURAL, a COBOL or PL1 program can be written to effect the same output.

UNISYS
DMS-II

This DBMS, a product of the former Burroughs Corporation, includes ERGO (Extended Retrierval with Graphic Output) which is a report generator. Within ERGO, the EXTRACT commands could produce a flat, sequential, file on tape. But to exercise this program, it is necessary to identify each data element to be transferred to the output file.

CONFERENCES:

The Quarterly Update will not publish a comprehensive calendar of upcoming conferences; instead it will list conferences which we intend to attend and may report on in future issues or conferences reported to us as having sessions of special interest to readers.

May 7-9, 1987 Mid-West Regional Archives Conference in Chicago. Linda Evans reports that the meeting will include exhibits of LS2000, MicroMarc/AMC, ZyIndex, and Cactus Software.

May 8-9, 1987 Mid-Atlantic Regional Archives Conference in Baltimore. Sessions on Automation for Reference at NARA (Ted Weir) and NY University (Tom Frusciano); on expert systems in archives (William Holmes, Pat Molholt, David Bearman); and on PC's and MARC/AMC (Donna Wells & David Anderson).

May 17-20, 1987 American Society for Information Science Mid-Year Meeting, Kings Island, Ohio. Focus on laser optical disks and video information systems

May 19-22, 1987 International Association of Social Science Information Service and Technology (I-ASSIST); 13th annual conference. Vancouver, BC; Sessions include workshop on accessioning computerized information into traditional archives. Tom Brown plans a report in the next issue.

June 4-6, 1987 Association of Canadian Archivists, Hamilton Ontario.

June 7-11, 1987 American Museum Assoc. Meeting, San Francisco.

July 11-13, 1987 International Conference on Databases in the Humanities and Social Sciences, Montgomery Alabama, includes keynote by Frank Burke on ISO standards 8211 and electronic archives and 6 other papers on electronic archives projects.

July 22-25. National Association of Gov. Archivists & Records Administrators. Atlanta Georgia. Pre-conference workshop on "Starting an Electronic Records Program" with Margaret Hedstrom, post conference workshop on "Document Base Management Systems" with Edward Johnson, and numerous sessions on automation in between, including one devoted to discussing the RLG/Seven State Archives Project.

ART LIBRARIES AND ARCHIVAL AUTOMATION DEVELOPMENTS

The meeting of the Art Library Association of North America (ARLIS/NA) in DC in February, included demonstrations of a large number of computer projects which indicated the impact which emerging standards are having in visual arts cataloging. The revision of the MARC-VM format, the publication of Betsy Betz Parker's: Graphic Materials: Rules for Describing Original items in Historical Collections, and the growing acceptance of the Art & Architecture Thesaurus were revealed, along with the increasing sophistication of systems employing these standards.

Cathy Whitehead of the Art & Architecture Thesaurus and Amy Lucker of the Brooklyn Museum demonstrated a generic non-book format, mapped to MARC tags, using *MARCON* software by AIRS Inc. The system provided active authority control from seven AAT hierarchies to the appropriate MARC field, supporting both alphabetical and syndetic displays of AAT terms. Jeanne Keefe Watkinson reported on the Rensselaer Polytechnic Institute *SPIRES* based system which employs the AAT Styles and Periods hierarchy to create call numbers. Merrill Smith demonstrated a database, using *Informix*, to the M.I.T. Rotch Visual collection videodisc which is also designed to employ AAT descriptors. A linkage of AAT, MARC and videodisc was demonstrated by Angela Giral of the Avery Fine Arts Library (Columbia University). MARC based records of outdoor sculpture were demonstrated on a modified GEAC library system by Christine Hennessey of the National Museum of American Art. The system is designed to employ AAT terminology in on-line authority control when that database can be mounted.

Numerous other projects of note were displayed, including a database developed by Pat Barnett of the Metropolitan Museum of Art which is a "Clearinghouse on Automation in the Visual Arts" maintained on *MARCON*. Several slide cartaloging programs, using *Smart Software* at SUNY Cortland, *SPIRES* at Rensselear Polytechnic Institute, and *Data Ease* at the University of Oregon, were featured. The Computer Project Demonstrations testify to a very active Special Interest Group within ARLIS with which print archivists have much to share.

FOLLOW UP ON THE NEWS:

VALENTINES DAY ERASURE:

The New York Times of February 15, 1987 reported that all the computerized records of the town of Prescott Valley, Arizona had been erased "leaving officials with no idea how much money has been spent this year or how much cash the town has left." I wondered - do any State Archives offer a security magnetic tape storage program to local governments analogous to their security microfilm programs?

DISCLOSED ERASURES:

The lesson is still unlearned, even if Ann Gorsuch was burned by electronic tracks earlier in this Administration! Now we discover that Admiral Poindexter and Colonel North forgot about PROFS, the office automation system through which their internal correspondence flowed. Much of what the Tower Commission learned, came from the PROFS audit trail; Government archivists may want to consider what this implies for the historical record, since such transactional audit trail records are nearly certain to have extremely short retention periods, if they are even scheduled.

ERASE THIS ADVICE:

The most recent edition of the Fellowes Manufacturing Company's Records Management Handbook, distributed as a free service, has been updated to include information about the preservation of magnetic media. Unfortunately they neglect to note that magnetic tapes should be rewound under constant tension at least annually and checked for errors, which, if found, call for copying onto a new tape. When I pointed out the problem, a Fellowes spokesman promised to take this into account "when it comes to publishing our tenth edition". Until then, beware or find your tapes erased before then.

ARCHIVES NEED NOT APPLY:

What I read about the Library of Congress Correspondence Reference Program, which now refers reference questions received by LC to 41 State Libraries so that answers to patron queries directed to LC may be provided by the patron's local library service, sounded like a terrific idea. So I wrote to ask if they had considered cooperating in a similar program with state archives.

R. David Myers, Head of the Section of the General Reading Rooms in charge of the program, answered that they had not, but that archives would be welcome if the task could be coordinated through Trudy Petersen at NARA. If nothing else, it would be an excellent education for LC staff; his answer ("nor do we receive many questions related to archives") suggested that the idea that archives could be an information source, rather than the subject of a query, was novel!

SOUNDS UNREAL? - MIGHT BE!

Optiram Ltd. is offering a text conversion service to publishers and archivists which they claim can scan any typefont and even handwritten texts and convert them automatically into ASCII characters with limited up front "teaching" of the font or handwriting style. Considering the dimensions of the breakthrough this implies, one would expect Optiram to be aggressively licensing its technology. Instead it is acting only as a service bureau, and won't demonstrate its system or answer my mail about its technology. Personally, I think we are dealing with another company which uses Phillipine, Korean or other third world data entry personnel, at least until they prove otherwise, but meanwhile they are offering very attractive prices for transcription of original, photocopied or telefacsimile texts - between $.95 and $1.75 per 1000 characters ($950 -$1750 per megabyte or approximately 400 pages of typed material). Discounts are available for larger volumes. This and other similar services, usually explicitly involving off-shore data entry, have dramatically changed the financial realities of major retrospective conversion projects for archives - anyone up for retrospective conversion of finding aids?

IN-BOX

This Department lists a selection of items received by Archives & Museum Informatics in recent months. It is not intended to be a complete bibliography of archival automation.

REPORTS:

American Library Association (ALA Washington Office, 110 Maryland Ave. NE, Washington, DC 20002), Less Access to Less Information By and About the U.S. Government: 2 - A 1985-1986 Chronology, 1987, 33pp. A sequel to the 1981-1984 Chronology. A quite frightening picture emerges from these pages.

Council on Library Resources (1785 Massachusetts Ave., NW, Washington DC 20036), Scholarship in the Electronic Age: A Selected Bibliography on Research & Communication in the Humanities and Social Sciences, Compiled by Anita Lowry and Junko Stuveras, February 1987, 54pp. Exceptionally interesting selection with excellent annotation important to archivists trying to understand the ways in which automation is changing scholarly research and communications.

Also, Costs of Microfilm Preservation at Research Libraries: A Study of Four Institutions, by Paul B. Kantor ($3 prepaid).

OCLC (6565 Frantz Rd., Dublin OH, 43017-0702), Campus of the Future, March 1987; proceedings of a conference sponsored by OCLC and the Johnson Foundation June 22-24, 1986. This conference heard a number of very important papers on the ways in which information technologies are changing academia, including one by Douglas Van Houweling (Vice Provost for Information Technology at the University of Michigan) entitled "The Information Technology Environment in Higher Education" which should be must reading for university archivists.

National Research Council, Preservation of Historical Records, Washington DC 1986. Technical discussion of media characteristics and current research.

University of Waterloo, Institute for Computer Research, Microcomputer Database Management Systems: Application to Cultural and Heritage Information Tasks. A Report Concerning Tests of 24 Commercial Software Packages. by E.M. Avedon et.al., March 1986. The continuing value of this report lies in the five test models, based on different types of cultural management tasks, developed for the evaluation and the discussion of features of products which make them suited to doing specific tasks well. ($5 to: Canadian Conference on the Arts, 126 York St., Suite 400, Ottowa, ON, Canada, K1N-5T5).

NEWSLETTERS:

Humanities Communication Newsletter - ISSN 0269-0519 (Office of Humanities Communication, University of Leicester, Leicester LE1-7RH, ENGLAND) edited by Dr. May Katzen, is the most comprehensive single source on European computerization projects in the humanities. Since 1983 it has grown from a two page flier to a 50 plus page bi-annual journal.

Museum Archivist (Kathleen Robinson, Editor, Museum of Fine Arts, P.O.Box 6826, Houston TX 77265) is the new publication of the museum roundtable of the Society of American Archivists. Its first issue, December 1986, contained a variety of useful reports.

Scholarly Communication (American Council of Learned Societies, Office of Scholarly Communication, 1717 Massachusetts Avenue NW, Suite 401, Washington DC 20036) has reached issue seven and continues to report in depth on a wide range of scholarly publishing and library issues.

ARTICLES & BOOKS:

Bailey, Robert, Archives Automation 1975-1987, For the Record ...Newsletter of the Illinois State Archives, Vol.9(2), Winter 1987 p.1-4. Describes over a decade of automation projects at the Illinois State Archives and includes a section in which the author complains that too much emphasis is being placed on MARC AMC.

Baskerville, Peter A. and Gaffield, Chad M., Archives, Automation & Access: Proceedings of a conference at University of Victoria, British Columbia, March 1 & 2, 1985, University of Victoria, no pub.date., 152 p. This conference included papers on automation issues in archives by Terry Cook, Richard Janke, David Bearman, Terry Eastwood, Hugo Stibbe, Susan Rosenfeld Falb, Sue Gravel, W. Theodore Durr, David Mattison, Michel Roberge and the editors. Most of these articles remain timely and are not available elsewhere.

Cook, Michael, The Management of Information From Archives, London, Gower Publishing Co., 1986, 203 p. plus bibliography and index. Cook's intention is to cover the entire life-cycle and all processes. From this side of the Atlantic, his observation on how little has been transmitted on archival automation is powerfully reinforced; at least we can now better understand the English side of the equation.

DeSalvo, Daniel, An Interview on Local Area Networks with Dr. Frank Burke, Acting Archivist of the United States, Telematics and Informatics, vol.3(2) 1986, p.93-100. An informal, slightly ideosyncratic, discussion. De Salvo's aricle on NARA expert systems in the same journal, vol. 3 (1) p.25-38 is well worth reading for the insight it provides into NARA priorities.

Jasco, P., Szucs, A., and Varga, S.; Micro-CDS/ISIS: A Bibliographic Information Management Software from UNESCO, Microcomputers for Information Management, 3(3), 173-198. Describes the microcomputer version of ISIS, a public domain bibliographic system now used in numerous archives (especially film archives) around the world.

Kimball, Margaret, Workflow for Processing Manuscripts in Automated Systems, Rare Books & Manuscript Librarianship 1(2) Fall 1986, p.117-126 is an edited version of a paper given at the 1985 SAA meeting on Stanfords' early RLIN MARC-AMC experience.

Manns, Basil & Swara, Tamara, Books to Bits: Digital Imaging at the Library of Congress, Journal of Information and Image Management, v.19 (10), 1986, p.27-32 report that the Library of Congress Preservation Microfilm office is finding digital scanning more cost effective than microfilm! Other issues relating to the technology are still being explored.

Sermis, Mark & Givner, Christine, Implications of Analog-to-Digital Technology for Social Sciences: A Review & Preview, Social Science Microcomputer Review, v.4 (3), Fall 1986, p.295-309. Yes, it will change the ways we ask questions and the way we answer them too. This and the Archeological Computing Newsletter, which I've recently begun receiving, introduce us to a whole new world of social science methodology.

Unsworth, Michael E., A Review of the National Inventory of Documentary Sources in the United States, Microform Review, 15(4) Fall 1986 p.232-239 is the first extensive review of NIDS I know of. The author is relatively positive overall but doesn't ask how any national inventory which includes only 791 of the 10,000+ collections at the Library of Congress or 196 out of more than 425 published NARA registers can substitute for a comprehensive tool.

Weber, Lisa, Status Report on SAA's Automated Information Program, SAA Newsletter, March 1987, p.7. Reports on SAA workshops on MARC-AMC, the clearinghouse on archival automation, and the results of the January MARBI meeting which adopted changes in the MARC-AMC format and of an SAA convened Thesaurus Study Group which is looking into destinctions between physical format and genre/form of material terms.

EPHEMERA:

Fred Meyer Charitable Trust (Library & Information Resources for the Northwest, 1515 SW Fifth Ave., Suite 500, Portland, OR 97201) Moving Information: Graphic Images on CD-ROM, by David Miller, Nov. 1986, 35pp.

Interagency Committee on Information Resource Management (IAC/IRM), Minutes of Meeting of Nov. 12, 1986. Bill Hooton (NARA) identified three barriers to optical disk projects becoming fully operational - legal admissability of disk stored images, longevity determination for medium, and data recording standards.

National Archives & Records Administration, Strategic Plan for Information Systems & Technology, Fiscal Years 1987-1991, Sept 1986, 168pp.

National Archives & Records Administration, Archival Research & Evaluation Staff, General Information Summary, February 1, 1987, 5pp.

National Information Standards Organization (NISO Z39), Subcommittee response to comments on Holdings Statements for Non-Serial Items, draft standard Z39.57-198x.

North Carolina Department of Cultural Resources, General Description of FAIDS, March, 1987, 22pp.

'Trends in Publishing: The Impact on Library Storage", by Richard R. Rowe (President , The Faxon Co.), presented to the Library Study Committee of the Ohio Board of Regents. Planning a new building? Rowe's reflections on library storage are worth considering.

Society of American Archivists, Automated Records & Techniques Task Force, Minutes of the Meeting of 26 August, 1986, 15pp. & 1987 program proposals, sent by Harold Naugler, dated February 5, 1987.

Society of American Archivists, Automation Program Officer (Lisa Weber), Archival Automation Information Survey, Test Survey Form (December 22, 1986) for clearinghouse on automation in archives. Forms to be distributed in April.

PLANS AND PROPOSALS:

ACTIONS VOCABULARY

The Wiltshire Museum Service Group in England has developed a vocabulary of acquisition, deposit and disposal actions which is proposed as a starting point for a terminology control working group of the Museum Documentation Association. Writing in a recent issue of MDA Information (v.10,#1), Stuart Holm suggests that the list, which contains terms such as bequest, exchange, loan, offer for gift, purchase and transfer complete with scope notes and a syndetic structure, might serve the needs of a broader community, which he invites others to join by writing to the MDA (Building O, 347 Cherry Hinton Rd., Cambridge, CB1-4DH, ENGLAND).

FRONT-END SYSTEMS:

As Lenore Sarasan has rightly noted in a stream of publications over the past decade, automation efforts in cultural repositories frequently break down around data entry before they pay off in retrieval. Good data entry, data validation systems have rarely been integrated modules of the information retrieval systems which attract archivists and museum curators. In a several issues of MDA Information, Richard Light has been reporting on MDA's internal development of a microcomputer package consistent with MDA standards which they hope to release in April 1987. Last summer the Smithsonian Institution issued an RFP for a similar system to serve as a front-end for all its bibliographic systems. *Data Ease* software was selected and Richard Szary and Avra Michelson expect to release a front-end for the S.I.Bibliographic Information System (SIBIS) in April as well.

RESEARCH & EVALUATION :

In a major contribution to the archival profession and an insurance policy for its own continued existence, the National Archives created the Archival Research & Evaluation staff in January 1983. The range of questions which this relatively small group has been able to address and the fundamental character of the answers they are beginning to provide, should be better known. As reported elsewhere, they recommended an optical disk effort in February 1984 which has been contracted. In October 1984 they completed a study of digital raster scanning, optical character recognition and speech pattern recognition technologies. In February of 1986 they introduced bar code labels into a system to control the audio-visual inventory. Their study of the potential of expert systems in archives may be known to the community through the publications of Dan DeSalvo in Telematics & Informatics, but their on-going research into the use of optical character recognition devices for converting typed and handwritten indexes and finding aids are less publicized. Other projects include research on preservation of paper, microfilm and magnetic media which could be of great significance to the field. Office Director William Holmes has agreed to keep me informed about their work, and I look forward to having him and his staff report it here.

OVERCOMING OPTICAL LIMITS:

The Council on Library Resources announces that Gary W. Collins is working with the Library of Congress to test optical disk aging and to determine the conditions which induce degradation of data stored on optical disks.

A committee co-chaired by Robert Wedgeworth (Dean of Columbia U. Library School) and Kurt D. Steele (Standard & Poor's) has proposed standards for copyright protection of materials transferred to optical disks. The standards, which affirm copyright across media, are expected to have a significant impact on publishers. Contact MaryBeth Peters, Register of Copyright, Library of Congress, Washington DC 20540.

STATUS OF THE AAT:

by Toni Petersen

The Art and Architecture Thesaurus, a project of the Getty Art History Information Program, has been in operation since 1981. Its purpose is to provide a standardized vocabulary of art and architecture terms for use by bibliographic, archival, and visual databases and for the inventorying of object collections. It is hoped that such standardization of input and retrieval language will result in greater user satisfaction and create a unified set of concept descriptors for indexers and researchers.

The AAT has built its vocabulary from such sources as Library of Congress Subject Headings, RILA, and the Avery Index to Architectural Periodicals, augmented by literature of the field drawn from dictionaries, encyclopedias, glossaries and authoritative

monographs. The terms are arranged hierarchically (i.e., conceptually) and alphabetically. There is an attempt to provide access points via all known synonyms or variants of a term. The MARC Format for Authorities is the system used for transmitting the AAT to users online. There will also be a print version which can be periodically updated.

AAT terminology is prepared by an editorial staff according to ANSI and BSI standards, using the National Library of Medicine's MeSH thesaurus as a model. It is reviewed and approved by advisory boards composed of experts in the areas covered.

There are presently about 30,000 terms covering the following categories: Design Concepts, Design Elements and Attributes, Styles and Periods, People and Organizations, Processes and Techniques, Disciplines and Occupations, Materials, Architectectural Components, Single Built Works, Complexes, Settlements and Landscape, Tools and Tool components, Drawings, Document Types and Visual Genre. Further development is underway in areas of the visual arts and material culture.

A new computerized thesaurus maintenance system was installed recently consisting of IBM AT's networked to a Britton-Lee Intelligent Database Machine (IDM), utilizing a thesaurus construction program developed by the Getty Art History Information Program. The IDM enhances the previous system, providing for numerous additional features such as the ability to access any term across all of the hierarchies. A term authority record, containing full information on an AAT term -- its definition, sources, variant parts of speech (e.g., "etching" and its variant "Etched"), and fields for scope notes and related terms are also provided. Work is underway to add both scope notes and related terms to all existing hierarchies. The next project for the AAT is the design of an improved display format for AAT terminology, and an online user system which looks ahead to direct access to the AAT by outside users.

[Toni Petersen is Director of the AAT, located at 62 Stratton Rd., Williamstown, MA 01267]

STANDARDS:

1987 has seen some extremely important developments on the standards front. This Department will regularly report on standards issues which archivists should be aware of.

OPEN SYSTEMS STANDARDS

The National Bureau of Standards has issued regulations which require all government agencies to adhere to the International Standards Organization Open Systems Inconnection (ISO-OSI) model in purchases from the summer of 1987 forward.

Last summer, the Department of Defense committed itself to making its TCP/IP family of protocols converse with OSI.

Major hardware vendors agreed late in 1986 to revamp the Manufacturing Automation Protocols (MAP) and the Office Protocols (TOP) which are major subsets of OSI in release 3 to be available by 1988.

The Linked Systes Project Policy Committee (the project to link the Library of Congress, OCLC, RLG, and WLN) agreed on protocols which are consistent with the ISO-OSI model for transfer of records from networks to local systems. They also announcved agreement on priorities, including project implementation, developing the capability to exchange bibliographic records among participants and assuring that the local linkages used by all four participants will be identical and LSP based. OCLC, the Library of Congress and RLG will implement intersystem shared cataloging and interlibrary loaning by the end of 1987. The Western Library Network did not announce an anticipated completion date.

The effect of these OSI developments is two-fold: First, archivists can assume that ISO-OSI will be the direction of the future and should plan accordingly even though full compatibility based on the model still has many hurdles to overcome. Secondly, as an effective national bibliographic system is welded from OCLC, RLG & WLN participation in one or another network will increasingly be determined by the services which each network decides to offer. It will be easier for archives to part company with their parent institutions primary bibliographic network when such a course is dictated.

FILE INTERCHANGE STANDARDS

The American National Standards Institute recently threw its support behind the International standard ISO-8211. The standard specifies medium independent, hardware independent and software independent file and data formats which allow for the exchange of files between different computer systems with a minimum of external specification. This is made possible because the envelope carries

along with the data, information about how the data is to be mapped into different structures to prevent information loss. While the standard has yet to affect software vendors (see report on DBMS archiving issues by Thomas E. Brown & William A. Reader in this issue), many hopes (especially archival ones) depend on it. The specification (32 p.) is available from ANSI (1430 Broadway, NY, NY 10018 for $9.00).

CATALOGING STANDARDS

The Joint Steering Committee for the Revision of the Anglo-American Cataloging Rules (AACR) 2nd edition agreed to the content of a revision of those rules to be published in mid-1988. A new version of chapter 9, on computer files, will be published later this year.

The Library of Congress Cataloging Distribution Service has issued update 14 to the Marc Formats for Bibliographic Description. Also available is Betsy Betz Parkers' long awaited Thesaurus of Graphical Terms.

The Society of American Archivists has submitted a grant proposal to NEH to enable Steve Henson to revise his Archives, Personal Papers & Manuscripts: A Cataloging Manual.

VOCABULARY STANDARDS

A meeting of the RLG working group on a functions vocabulary (attended by David Bearman, Max Evans, Jamie Ingram, Marion Matters, Alden Monroe, Kathleen Roe, Alan Tucker & Lisa Weber), reached an agreement about how to proceed in the Seven-States Project exploration of the potential of "functions" terminology to provide access to common activities across different bureaucratic organizations.

Two levels of terms were proposed for a test: a generic, "sphere of activity" term will be applied to agency history records and a "function" term will be applied to series level descriptions of records. The participants will be testing these terms in the spheres of Corrections and Education over the next several months.

A subcommittee of the RLG AMC Task Force proposed standards and examples for recording records schedules in RLIN. The intention of the standards advanced by Kathleen Roe and Alden Monroe is to permit the Seven States project to input some scheduling records to test the potential of such information for records managers and archivists.

TECHNICAL REPORT ON OPTICAL MEDIA

Part 2 of the Quarterly Update on Archival Informatics: Newsletter & Technical Reports, is a technical report on optical media in archives and museums. The following is a summary for those readers who receive only the Newsletter. Non-subscription copies of the technical report are available for $45.00 each.

Technologies which employ light to write and read information, while in their infancy, have tremendous potential for archives and museums. Laser-reflective optical videodisc has already demonstrated its value and is now being used quite widely in collections documentation, exhibit and public information applications. CD-Audio has taken the consumer market by storm, but its archival and museum applicability is still uncertain. CD-ROM is now sufficiently standard to support many data distribution applications and we can expect to see new products for the archives and museum market emerging regularly over the next several years. Optical digital disk is so new that until very recently only national institutions launching prototype and demonstration efforts could claim to have demonstrated archival or museum applications, but in the past year a number of other large ventures have been launched, some of which have moved from pilots to production.

The attraction of each of these technologies is different because their capabilities are different. Laser reflective optical videodisc is extremely attractive for publishing images, especially segments of motion images, with sound and data in interactive formats. It is a stable technology which boasts hundreds of successful implementations in archives and museums; the videodisc cost $2000 to master, copies are cheap and playback devices are standard and available for under $1000. CD-ROM is very attractive for publishing traditional databases, as evidenced by the interest of the Census Bureau in using it for the 1990 census, and can accommodate still images; mastering costs are fairly high and playback devices are not yet standard peripherals, but both can be expected to drop significantly soon. Optical digital disk is exciting because it is not limited to being a publishing medium - it can be directly updated by local devices. It also holds enormous quantities of data and provides

(continued next page)

OPTICAL MEDIA REPORT, continued from p. 19

rapid random access. Nevertheless, optical digital disk is a very non-standard technology and the costs of systems are hard for most archival and museum applications to justify.

In this report the technologies themselves are examined in less detail (with references to more complete treatments) than are applications. Criteria for selecting projects for each technology are discussed, as are cost factors associated with each and the staffing and support requirements of projects which use these technologies. Case studies of previously successful archives and museum community projects are assessed. A series of checklists which can be used to determine whether an application is suited to these technologies, and if so which technology and under what assumptions, is provided for management guidance. Lists of current vendors and descriptions of their services are provided along with extensive bibliographies of background studies, case studies and critical reviews.
