

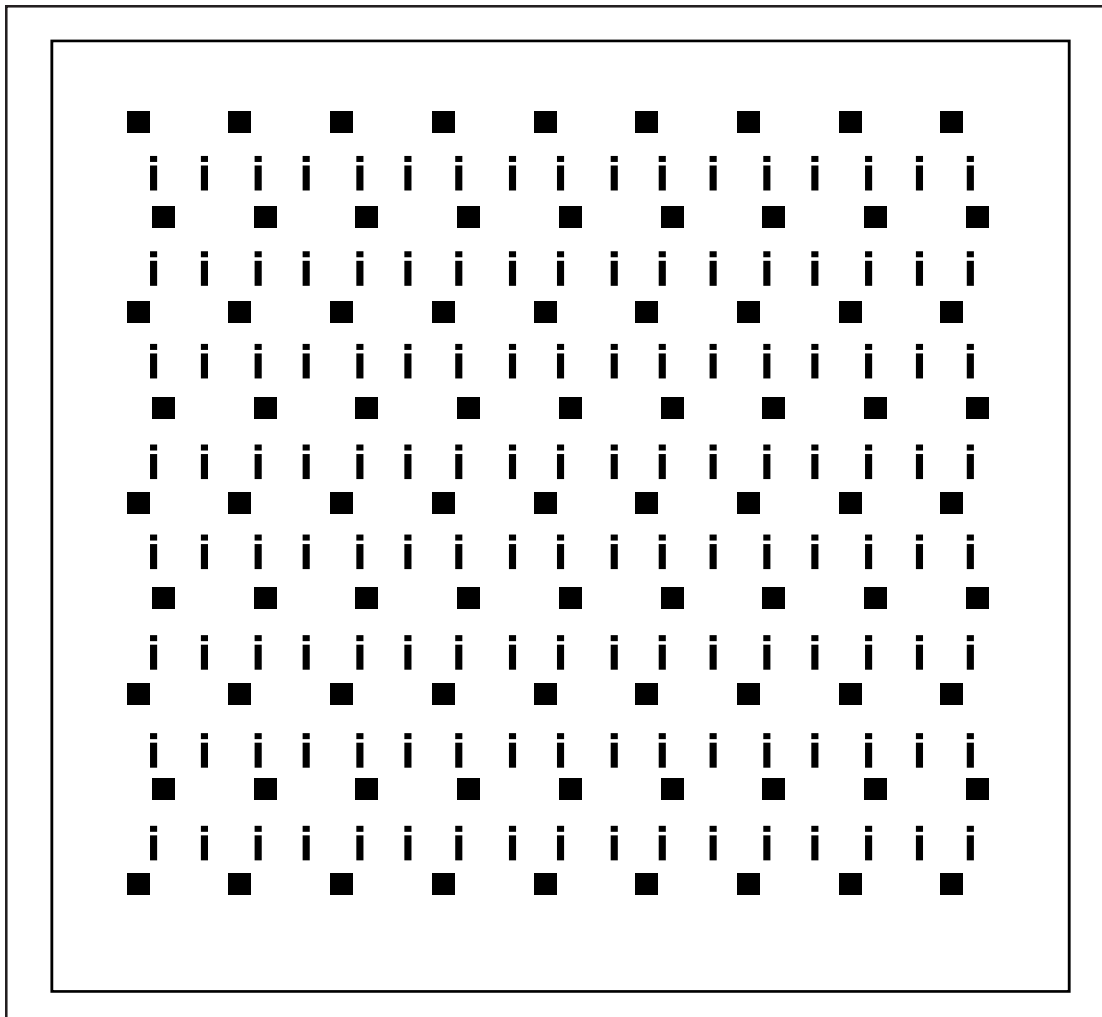
# IASSIST

Q U A R T E R L Y

VOLUME 18

Spring/Summer 1994

NUMBER 1&2





# IASSIST

---

# QUARTERLY



The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

#### Information for Authors

The **QUARTERLY** is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science* 30(2):77-82, March 1979. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Walter Piovesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University, Burnaby, B.C., V5A 1S6 CANADA. (604) 291-5869 E-Mail: walter@sfu.ca Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.EDU

Title: **Newsletter - International Association for Social Science Information Service and Technology**  
ISSN - United States: 0739-1137 Copyright 1985 by **IASSIST**. All rights reserved.

## C O N T E N T S

Volume 18 Number 1/2 Spring/Summer 1994

### FEATURES

- 4** **Setting Up a National On-line Census Data Service**  
*by Virginia Knight*
- 7** **Managing Machine-Readable Archives: Progress with Description and Exchange Standards.**  
*by Michael Cook*
- 11** **Reeling Them In: Accessioning the Electronic Records of the United States Government**  
*by Mark Conrad*
- 14** **Electronic Media and Preservation**  
*by Fynnette Eaton*
- 18** **"The Library of Congress at a Glance": Text Visualization and Reference Rooms Without Walls**  
*by Lee A. Gladwin*

### News

- 23** **ISBD(CF) Report**

---

# Data Services Since the 1960s: Where are We Going?

---

by David Elesh<sup>1</sup>  
Center for Public Policy,  
Social Science Data Library,  
Temple University,

## Introduction

Twenty-five years ago, I wrote a proposal to the National Science Foundation for a Workshop on the Management of a Data and Program Library to promote the establishment of local university social science data archives. Although organized in only a few weeks, the Workshop attracted 100 people from 50 different institutions in the U.S. and Canada. We even had one person from Sweden, indicating the deep roots of Swedish archival data activity. A few institutions already had data archives, and within a very short time, all did.

In preparation for my remarks here today, I went back to the *Proceedings* of the Workshop we published shortly afterward<sup>2</sup>. In the “Preface” to the *Proceedings*, I outlined what I thought were the benefits of establishing a data archive. For faculty, the benefits appeared to be greater productivity, the potential for investigating new kinds of problems, and the ability for scholars with limited funding and resources to access high quality data. Graduate students could be given a greater opportunity to gain research experience, and they shared with faculty the possibilities of examining new kinds of Problems and high quality data.

However, I reserved my greatest expectation of benefits for undergraduates. Because of the expense and time required for empirical social research, undergraduates in the sixties rarely experienced a genuine introduction to the social sciences as research disciplines. Instead they read brief and, often, highly simplified synopses of research that gave little indication of how quantitative social science is done. The ready availability of data and program libraries would, I thought, make possible realistic introductions to the social sciences.

Looking back now more than two decades, I think it fair to say that data archives have made a significant difference to faculty and graduate students and have not had much impact on undergraduates. For faculty and graduate students, data archives have vastly increased the amount of comparative and over-time research published. Better data also fostered greater sophistication in social scientific theories and analytical techniques. Twenty-five years ago, economists aside, most social scientists were content with theories based upon cross-sectional

data, cross-tabulations, ordinary least squares, and path analytic techniques which required no more than ordinary least squares. We now find publications and dissertations dealing with dynamic theories based upon event history analysis, partial adjustment models, dynamic LISREL type models, and more<sup>3</sup>. Most of this work was made possible by the collections of data archives.

However, the impact of data archives on undergraduate education has been modest. I will acknowledge that data archives significantly altered patterns of instruction in undergraduate methods and statistics courses, but in most colleges and universities, these are “required” courses segregated from the substantive foci of their disciplines. In all too many cases, the links between these courses and substantive courses are left to the imagination of the students. The vast majority of undergraduate courses today are taught in a manner little different than decades ago. Faculty lecture about research that students find summarized in their texts, and the investigatory process that produced the results about which students read is about as much an enigma today as it was then. As a result, the skills taught in the methods and statistics quickly grow stale.

Ironically, the achievements of data archives in enriching faculty and graduate student work often has led to an institutional success which works against serving undergraduate education. Twenty years ago, data libraries had little or nothing to do with conventional libraries. They were creations of faculty members seeking to provide research and instructional resources for themselves, their colleagues, and their students, and they were housed in faculty offices or a few rooms down the hall. By and large, librarians did not understand computers and data and often were uninterested. Much has changed. At contemporary meetings of Inter-university Consortium for Political and Social Research Official representatives, there seem to be as many librarians as faculty—certainly they form a substantial fraction of those attending the meetings. Libraries are moving to accept data archives as important parts of their collections—a change that represents institutional commitments to data libraries as scholarly resources. What, then, is the irony? It is that even as libraries take on this new function, university and collegial support for libraries is either static or declining.

From 1975-76 to 1985-86, current fund expenditures for libraries in institutions of higher education fell from 3.1 percent of total expenditures to 2.6 percent<sup>4</sup>; and this occurred during a period in which library costs for books and periodicals increased roughly 60 percent faster than overall inflation.

Quite simply, the cost pressures universities have faced for more than a decade show no signs of improving soon, despite all the professed concern about the state of American education. This means that libraries are unlikely to have the kinds of staff resources required to help undergraduates use data resources. In fact, there is the real possibility that all users will suffer.

But the picture is not completely bleak. Much is changing in social scientific instruction, and an increasing number of undergraduates are gaining experience in doing quantitative social science. The introduction of microcomputers is slowly—very slowly—transforming undergraduate education. Texts increasingly come complete with analytical software and databases, and independent instructional packages and databases such as ShowCase have been adopted widely. Both types of innovations make it possible to introduce real, quantitative social scientific work to both lower and upper division undergraduates and usually find enthusiastic acceptance. But neither involves or leads to use of data archives. Why?

First, it is important to recognize that data archives were conceived in an era of mainframe computing and were meant to be analyzed by mainframe statistical packages. The fundamental meaning of this statement is that the knowledge base required to use these resources is simply much larger than for a PC. The architecture, organization, and funding of mainframe computing are designed for the researcher, not the instructor, and certainly not, excepting computer science students, the student.

Mainframe operating systems are far more sophisticated than those available on PCs. Mainframe statistical packages, while powerful, are typically intimidating in their complexity, and even those of us who routinely have required undergraduates to learn these packages sufficiently to get through our methods and statistics courses know that we must sacrifice some content to allow time for teaching basic computing skills. The learning curve for mainframe computing is a great deal steeper than for PCs.

In the past, we could defend the loss of statistical or methodological subject matter in the belief that knowledge of SPSS or SAS and the like formed part of the research skills we were trying to impart. Those of us who used the computer in our undergraduate instruction

learned to create program and/or system files that shortcut many of the procedures we expected graduate students to learn. We used archived files because there were few alternatives, and setting a file up for a class was little different than setting one up for our own research use. But the skill level and time it requires to do these things are significant, and many social scientists simply did not and still do not have them.

Nor would they or their students find much help in the organization of computing. Because the machine or machines were located centrally, it was and is almost universally true that consultants were as well. One had to go to the Computing Center to use computers or seek assistance in using them. At the same time, the available consultants were and are typically programmers unfamiliar with statistical software and social science data. To a very large extent, users must learn the consultants' language in order to obtain help; they do not learn users' language. A few consultants might learn SPSS, SAS, BMDP, or the other statistical packages, but the demand for their services always exceeds the supply because social science users of the computer are greatly outnumbered by users in computer science, engineering, and the physical sciences, and the latter have the influence that attends greater external funding; thus central computing budgets favor the latter over the former.

Local data archives often tried to fill the void by offering assistance in use of the computer as well as of the data. Staff became expert in the use of tapes and the manipulation of large and complex files; in some institutions, they provide and have provided the primary consultative assistance in these areas.

Against this background, it is not surprising that analyses of archived data did not spread widely in undergraduate instruction.

While I think there is little doubt that the introduction of microcomputers can transform undergraduate instruction, I have some doubt that data libraries will be significant actors in the transformation. Micros eventually will succeed in changing social scientific instruction because they significantly lower the slope of the learning curve for computing. Students find PC operating systems easier to learn, and unlike the mainframe world, there are statistical packages specifically designed for instructional use which require far less faculty and student time to learn. However, generally these packages incorporate data which has been tailored for them, and the tools necessary to include other data sets are omitted. One can even find software that allows the student to place a disk in the lowliest PC, turn it on, and find him or herself in a menu driven analytical package capable of multivariate crosstabulations on a substantial number of variables

with adequate samples and with virtually instantaneous response.

Microcomputers also introduced a new market structure for computing and data. In the mainframe environments, computing is provided and funded centrally as a university or college function, and instructional costs are, at least partially, borne by tuition. Data files are also provided centrally and usually cost users nothing. However, with the introduction of microcomputers, the cost of the hardware, software, and data are increasingly being borne by the user as direct charges. Where universities or colleges supply microcomputing laboratories, the number of these institutions that have introduced “laboratory” fees to cover these costs grows with each passing year. And, as noted, software and data increasingly come either from text publishers or other third party vendors.

Clearly, publishers are seeking to make it significantly easier for students and faculty to analyze data. Clearly, too, if my history of the past quarter century or so is correct, greater ease-of-use is necessary if data analysis is to spread to substantive subjects. Although I have no hard evidence, I suspect that the effort to produce greater ease of use is producing instructional software that is increasingly valuable for research purposes—the development of analytical graphical displays is one example—which is an interesting reversal of direction for the traditional flow of technology.

It is possible for data libraries to participate in this transformation, but they will have to change their traditional modes of operation in several ways. First, they will have to work with faculty members to identify analytical software that is easy-to-use and capable of analyzing and presenting data in a way that the faculty member finds useful. Second, they will have to create files for that software. Typically, this will mean creating files on a mainframe, exporting them in ASCII, downloading them to a micro, and modifying them for the analytical program. The program may be a statistical package, a spreadsheet, a graphics program, a database program, or something else. The choices are larger in the micro world, and faculty demands are and can be expected to be diverse. Third, data libraries should attempt to develop expertise in exemplars of a number of software types—e.g., statistical packages, databases, spreadsheets, graphics—because it will be necessary if they are to be able to provide support for the files they create and because faculty are likely to ask for recommendations. Fourth, as networks expand and take on some of the functions of mainframes, data libraries will have to learn how to create and maintain data servers for users at all levels of sophistication. Fifth, data archives should look to the creation of display-formatted tables resident as files on disk as reference works for their most heavily utilized files.

While some tables on many subjects will be available on CD-ROMs from a number of vendors, it should be possible to create tables from archival holdings that serve the needs of particular programs at a cost significantly lower than would be required to manufacture a CD-ROM; software exists to compress such files and expand them as they are called by programs.

All of these possibilities for data archives require new investments—albeit at a relatively modest level—at a time when funding for new ventures is difficult. Given the cost pressures higher education now faces and will likely to face during the next decade, it is more likely that funds for these initiatives will come from a more efficient utilization of existing resources than from new ones.

One is supposed to close discussions of the future on an optimistic note, and I will try to do so. The transformation of computing offers substantial opportunities for using the data in our archives more broadly. We can move beyond our traditional support of faculty and graduate student research to make more of an impact on undergraduate instruction. But it will take initiative and a careful marshalling of resources. Otherwise, the past is, at best, all too likely to be prologue.

1. Paper presented at IASSIST 1990 in Poughkeepsie, New York.

2. Workshop on the Management of a Data and Program Library. Proceedings eds Margaret O’Neil Adams, David Elesh, and Alice Robbins. Madison, WI, 1990.

3. I do not wish to re-open old, and typically, fruitless, debates about the relationship between theory and empirical research. I simply wish to note that neither theory nor research was much concerned with dynamic relationships in the 1960s.

4. U.S. Office of Education, Digest of Educational Statistics, Washington, D.C., Government Printing Office, 1990, p. 301.

---

# Managing Machine-Readable Archives: Progress with Description and Exchange Standards.

---

by Michael Cook<sup>1</sup>  
Archival Description Project,  
University of Liverpool

## Background

Archivists have come somewhat belatedly to the idea that there should be formal standards for description and for the exchange of data about their materials. Observing progress made in these fields in North America, British archivists began work on constructing the necessary instruments in 1984. The Archival Description Project was set up at Liverpool University, supported by funds from the British Library Research and Development Department and the Society of Archivists. The Project team has produced two successive texts of a Manual Of Archival Description, affectionately known as MAD. The second edition, MAD2, published in 1990, was published by Gower, and has received a reasonable degree of trialing<sup>2</sup>

The archival community in Britain, however, finds itself in a difficulty as regards the formal adoption of a standard. There is a National Council on Archives, and a working party of this, chaired by Dr. Kitching (who is Secretary to the Royal Commission on Historical Manuscripts), has recommended the adoption of MAD2. In a rather similar way, the Society of Archivists has issued signals of approval, and has asked its Professional Methodology Panel to carry out tests and development work. These measures are somewhat short of a formal endorsement, but they do indicate acceptance at a practical level, and show that there is a will to continue developing the work.

The second edition of MAD contains rules for the description of a number of special formats, commonly found amongst archives. These are:

- title deeds (legal documents transferring land)
- letters and correspondence
- photographs
- cartographic archives
- architectural and engineering plans
- sound archives
- film and video archives
- machine-readable archives

This section of MAD2 must still be regarded as experimental, and it has not yet received adequate trialing. The principles on which the rules and guidelines are based, are coherent over the whole body of MAD2 and will be

discussed later in this paper.

The MAD2 special formats are intended for use in general archives repositories and services, not in specialist institutions. This important restriction should be emphasised.

The second set of archival description standards which should be mentioned are the international ones. The International Congress on Archives held in Montreal in September 1992, received the text of two new standards:

1. **Statement of Principles regarding archival description** (the Madrid Principles). Since this had been debated by the profession since 1991, this text was adopted.
2. **General International Standard Archival Description (ISAD(G))**. This was received as a draft for dissemination and discussion.

The first of these texts is now will be available. The second, (ISAD(G)), is not immediately available as it is in course of publication. It is intended that there should still be discussion of the topics presented, so that the process of maintenance and development may proceed. ISAD(G) itself has indeed not yet received formal adoption, but since it is in its second draft, and has received considerable discussion all over the world, it must be regarded as being near completion.

The other main standard applicable to archives, which ought to be mentioned here concerns data exchange. This is the MARC format, an archival application which was developed in the USA in 1984. It has become widely used in North America to allow archival descriptions to appear in the bibliographic databases, RLIN and OCLC. These databases are not widely available in Britain as yet, and the resistance of archivists to bringing in a library standard has been such that up to now MARC has been virtually unused for archives. There has indeed been little opportunity for it to be used. This situation appears to be changing, and a version of UKMARC in the archival format (AMC) is due to appear in 1993<sup>3</sup>

Turning now to the management and use of machine-readable archives, few British archivists have yet had

much experience. The ESRC Data Archive and the Edinburgh Data Library have been almost alone in the field in this country. The Public Record Office had ambitious plans to establish a Data Archive department during the mid 1980s. These have not progressed as might have been hoped. Recently there have been signs of life from this quarter, and we are given to understand that the PRO's Computer-Readable Data Archive will be established in 1995, with public access in 1997<sup>5</sup>.

It is important to make clear that there is a distinction between machine-readable files and datasets (which are the material administered by the ESRC Data Archive and other similar services) and machine-readable archives. The latter, like archives generally, are materials produced by, and forming part of the activity of, an organisation of some kind (such as a government). Archives of any sort are therefore unlikely to be one-time studies, or to have enough individual distinctness to allow them to be treated as discrete objects, comparable with books. Archives belong together in aggregations, which owe their character to the administrative system which produced them. Some of the consequences of this distinction are discussed further below.

### The guiding characteristics of MAD2

The work both of the Archival Description Project and of the ICA's Ad Hoc Commission on Archival Description, has shown that certain basic principles underlie all description of archives. International agreement on this, at least as far as traditional records are concerned, is quite remarkable. The description of machine-readable archives, therefore, is likely to require attention to these principles, if only to test their applicability to new materials. The following section attempts to summarise what the basic rules are.

#### 1. Levels of Arrangement and Description.

The idea that there are standard levels of arrangement is not new. The concept was first indicated in Europe at the start of the 20th century, then clarified in the USA<sup>6</sup>. It has been rediscovered and republished in different forms ever since. *MAD2* restates the principle, but also extends it. A table of levels is given which looks at first sight like the hierarchical continuum characteristic of a classification scheme, and numbered like one:

0 Repository level: suitable for combined descriptions covering more than one repository.

1 Management levels assemblies of archival groups brought together on the basis of some common feature, for the convenience of the repository. E.g Official/non-official archives, ecclesiastical archives, private papers. Subordinate groupings may be numbered using decimals of 1.

2. Group or collection level (internationally **fonds**): the archives of distinct entities. Subgroups (functional divisions within the group) are numbered using decimals of 2.

3. Series (within Britain, termed class): physically related sets of archives. Subseries are given decimals of 3.

4. Items: the unit of physical handling (volume, file, box).

5. Pieces: indivisible components; documents. Levels 4 and 5 may be used interchangeably in some cases.

The interesting thing about this table is its universality. Yet it is unlike a general classification scheme because it is tied to observable external phenomena at three points:

**Fonds** (level 2) always relates to the total archival product of a distinct entity (organisation or individual);

**Series** (level 3) are always the physically and systematically related product of an administrative activity, sets that belong together because of the way they were created and used;

**Items** (level 4) are always the physical units of handling.

No level of arrangement is compulsory; though in the Madrid Principles it is stated that the level of the **fonds** is "the broadest unit of description"<sup>7</sup>. Therefore, provided that we accept that the three levels above must always be set to correspond to the appropriate physical entities, any or all of the levels of arrangement can be used, above the fonds, or below the item, as convenient.

There can be problems in identifying what should constitute a **fonds**. *MAD2* advises that administrative or political levels of dependence should be disregarded. Thus an overall or umbrella organisation can be the origin of a fonds, but so can organisations which are administratively part of it. An extreme illustration would be that the Government of a country could be the source of a fonds (provided that it did actually produce records as such); but so could any of its Departments, or even lower subdivisions, sections etc. If any organisation is complete enough in itself to produce its own archives, it can originate a **fonds 1**

#### 2. The multi-level rule

The multi-level rule in *MAD2* states that archival descriptions should normally embrace more than one

level of arrangement. This is fully consistent with the multi-level rule laid down in ISAD(G), and in the Madrid Principles. However, MAD2 has a further elaboration of the principle, which has an important use in the context of finding aids. This is the concept of the 'macro' and 'micro' description.

These two terms do not relate to the specific levels of arrangement which are being described, but to the relationship between them. For example, finding aids frequently contain descriptions at fond, series and item levels. In these, the macro-micro relationship has a triple form:

Fonds description: a macro description governing:

Series description 1: a micro description in relation to the above, but a macro governing.

Item descriptions: micro descriptions of items in series 1, governed by the above.

Series description 2.... etc

In the MAD2 models, guidelines suggest that these relationships of dependence should be demonstrated to the user by the use of narrower margins, left and right; this assumes a hard-copy finding aid using standard pages. That is a common situation but not the only one. The important thing is that in any given case, the macro and micro descriptions may relate to any level of arrangement: fonds/item; management group/fonds; item/piece, etc. It is therefore a misconception to regard the macro description as peculiar to the 'higher' levels of arrangement, and the micro to the 'lower' ones.

Macro descriptions are written from a different standpoint than from micro descriptions. Their standpoint is the aggregate (whichever it is). Micro descriptions give information specific to each case. In the example above, the fonds description will give information relating to the fonds as a whole (probably including provenance information, but this is a separate issue); it also gives all information common to the series which follow, in order to avoid redundancy

The series descriptions which follow have a dual character. In so far as they are micro descriptions, they deal with each series one by one, giving specific information. Each serie description then operates as a macro for the items which follow. As macros they give information which relates to the series as a whole, and common data for the items. Finally, the items give data specific to each case.

This rule has been explained at some length because it

makes it immediately clear that, and why, standards originating in library practice are not suitable for archival applications.

### *3. The data elements table and its structure*

Archival descriptions require data of two different kinds: information about the origin, background, context and provenance of the archive; and information about its content. Descriptions must therefore be essentially structured. The Project team drew up a list of the data elements that can be found in these descriptions, and drew them together into seven 'areas'. Like the levels, most data elements and areas are optional, and are brought into use only when required for the specific case.

MAD2 sets out a number of models which govern the way in which descriptions can be set out, using the data elements and areas. These models accommodate the multi-level rule and allow the dependence of micro upon macro descriptions to be demonstrated so as to be easily perceived by users.

### *4. Access Points and Provenance*

ISAD(G) introduces the concept of access points, which should be subject to authority control. Access points should be provided for provenance information as well as for data from the contents of documents. Work on authority files, sadly lacking in the archive world, is therefore needed.

### **Standards for the description machine-readable archives**

Unless it is true that machine-readable archives are quite unlike any other archive, description standards for them should follow the models and rules for archival description, including the basic principles outlined above. Section 25 of MAD2 deals with this problem.

Although it may be anomalous to speak of levels of arrangement where the materials can never be physically arranged, it is nevertheless true that there must be levels of description. Both the **fonds** (the archive of a whole organisation) and the series still appear to have a real existence.

There is some debate about whether or not machine-readable archives must be treated in a radically different way from other archives<sup>9</sup>. Those who concentrate on the media which carry electronic documents, are conscious above all of its evanescence, its lack of objective existence. Those who look primarily at the origin, context and purpose of the document will have a much more traditional picture. The **fonds** will doubtless also contain descriptions of traditional archives, or archives in alternative forms. The series is normally the dataset which can most be regarded as a complete entity for

description and management purposes. It most clearly resembles the datasets held by the ESRC Data Archive.

MAD2 proposes that there should be short descriptions of the entities at these two levels, written into the main finding aids of the repository. When this is done, separate and specialised descriptions of the machine-readable groups and classes can be established, with a linkage between the two systems. This method allows a generalist repository to have a finding aid system which is an effective intellectual control over its total holdings, while at the same time designing a specialist finding aid which is appropriate to technically different materials.

The specialist description may itself be multi-level, or it may be a flat file, according to circumstances. It must clearly contain all the metadata required: the technical information needed to record the internal structure of the file and its software dependence. The data elements needed for this are listed in Section 25.

A final note might be that background, context and provenance information should always be provided, because without it the meaning of the electronic record is lost. Indeed this point is conceded by the practice of Data Archives. This 'macro' information, however, does not necessarily have to be held in the detailed, specialised, file which is the direct finding aid to the machine-readable data. It may be held in the main finding aid system of the repository. In future, this main finding aid may of course be itself held in a machine-readable form; or it may be processed so as to enter it into a national index, or into a data entry system. For these, both cataloguing and data exchange standards will be needed.

1. Paper presented at IASSIST 93 in Edinburgh.

2. M.Cook & K.Grant. Manual of archival description. Society of Archivists, 1986. [Some exemplars of this edition were wrongly marked '2nd edition'] M.Cook & M.Procter. Manual of archival description 2nd ed. Gower, 1990.

3. Copies of texts and current drafts are obtainable from the Secretariat of the International Council on Archives Ad Hoc Commission on Archival Description, National Archives of Canada, Ottawa

4. Alan Hopkinson & M.Cook. Information from the former at the Library and Archive, Tate Gallery, London.

5. Alexandra Nicol and Steven Duffield. Unpublished paper to seminar on electronic records held at the School of Library Archives and Information Studies, University College, London, 10 Dec 1992.

6. Richard Lytle. **Subject retrieval in archives: a comparison of the provenance and content indexing methods**. PhD, University of Maryland, 1979.

7. Statement of principles regarding archival description, First version, revised, section 2.2.

8. Terry Cook. **Treatment of the archival fonds: theory, method and practice**. Bureau of Canadian Archivists, Ottawa, 2. It is interesting that the concept was known to, but misunderstood by, the Marxist regimes of Eastern Europe. They adopted the habit of setting the fonds at too high a level of institutional independence, hence most institutions had to be the originators of sub-fonds. This misuse serves to underline the validity of the concept when used properly.

9. Charles Dollar. New developments and the implication on information handling'. In Information handling in offices and archives, ed. Angelika Menne-Haritz. K.G.Saur, 1993 pp56-66.

---

# Reeling Them In: Accessioning the Electronic Records of the United States Government

---

by Mark Conrad<sup>1</sup>  
Archivist  
National Archives and Records Administration

The Center for Electronic Records is the unit of the National Archives and Records Administration (NARA) that is responsible for, among other things, appraising the electronic records of federal agencies; accessioning those electronic records identified as permanently valuable; preserving the records once they have been acquired; and providing reference services for the records. This paper will offer an overview of the procedures used in carrying out these responsibilities. Particular emphasis will be placed on the steps involved in accessioning electronic records.

First, we need a definition of records. Records, as defined in 44 U.S.C. 3301,

..include all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them. Library and museum material made or acquired and preserved solely for reference or exhibition purposes, extra copies of documents preserved only for the convenience of reference, and stocks of publications and of processed documents are not included.

There are several ways in which the Center for Electronic Records may have its first contact with an agency concerning a dataset. Three of the most common ways are:

First, an agency may contact NARA to schedule the final disposition of records. Under 44 U.S.C. 33, no federal agency may dispose of federal records without the approval of the Archivist of the United States. Agencies usually obtain approval for the final disposition of their records through the use of a Standard Form 115, Request for Records Disposition Authority. This form is often referred to as a

“schedule”, “records schedule”, or “SF 115”. A records schedule contains a description of the records to be disposed of and the proposed disposition. If a records schedule contains electronic records, the Center is contacted.

Second, the agency may contact the Center with a direct offer of unscheduled records. This sometimes happens if the agency has produced a dataset that they consider important to documenting the mission of the agency.

Third, the Center for Electronic Records may initiate contact with the agency in an effort to acquire particular datasets with lasting value. If the Center staff learns that an agency is producing important datasets, we contact that agency in an effort to get them to schedule those records. Unlike textual records which can sit for thirty to fifty years or more without significant deterioration, it is important to acquire electronic records soon after their creation to ensure their preservation.

While there are a number of different ways that the Center for Electronic Records can become involved with the acquisition of electronic records, once the process begins, the same steps are usually followed. The records are first appraised to determine if they have sufficient value to warrant their acquisition and preservation.

It would not be practical, possible, or desirable for the Center for Electronic Records to acquire a copy of every electronic record produced by every agency of the United States Government. Appraisal is the process of determining which records will be retained and preserved and which records will be discarded.

Records are appraised as permanent if they have high legal, evidential or informational value. Records have high legal value if they are required to preserve legal rights of the government or individuals affected by the government. Records have high evidential value if they document how a government agency carries out its mission, develops policy, or makes key decisions. Records have high informational value if they contain data that might be valuable to a researcher for reasons

other than the reason the federal agency gathered the information in the first place.

I do not intend to offer a complete explanation of the appraisal process in this paper. The appraisal of electronic records could be the subject of a paper by itself.

The appraisal archivist must weigh a number of factors related to the electronic records at hand in order to determine whether they should be preserved. Some of the questions to be answered are:

- Where did the data come from?
- How was it used?
- Did this dataset have a major impact on federal policy?
- Is the dataset unique?
- Is the electronic format the most desirable format for keeping the information?
- Is the data available in a software/hardware independent format?
- Is there adequate documentation of the dataset to make the data accessible to a researcher?
- Is this dataset likely to be used by researchers?

Once the appraisal archivist has reached a decision and that decision has gone through an extensive review process, the Center for Electronic Records will begin the negotiations with the agency to accession those datasets that have lasting value. Accessioning is the process of transferring legal and physical custody of the records from the agency to the National Archives.

Standard Form 258, Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States (SF 258), is the document used for transferring title of the records to the National Archives. The agency sends the datasets, documentation for the datasets, and the SF 258 to the Center for Electronic Records to begin the accessioning process.

The Center for Electronic Records requires the agency to transfer the electronic records in a prescribed form. All records must be transferred in a software/hardware independent format. Records must be transferred on 7 or 9 track, 1/2 inch open-reel tape at 1600 or 6250 bpi, or on 3480 class cartridges. Records must be in ASCII or EBCDIC with no internal control characters and blocked no higher than 30,000 bytes. These specifications are

used to ensure that the Center will be able to transfer the electronic records to new media as current media become obsolete.

When the tapes arrive, they are checked for readability. If they contain data errors that cannot be corrected by cleaning the tape, the agency is contacted for replacement tapes. If the tapes can be read with no problems, tape maps and dumps of a limited number of records from each dataset are printed. The accessioning archivist compares the tape map with the documentation to verify that the tape described in the documentation is, in fact, the tape she/he is looking at.

The next step is for the archivist to manually compare the printout of several records with the documentation to ensure the record layout and codebook match the actual records. If, for example, a record layout indicates that columns sixty through sixty-five should contain the "respondent's date of birth" and the archivist finds "Peoria" in those columns, clearly there is a problem. The archivist consults the agency for a new record layout or new records.

This validation can be very time consuming. The archivist usually validates less than ten records per dataset, but an accession may contain several hundred datasets and individual records may be several thousand bytes long. The Center for Electronic Records is currently developing a system to automate the validation process.

The application will allow a staff member to enter information about the record layout and codes for a particular dataset. The dataset can then be checked by the computer, byte for byte, against the entered description. The dataset descriptions can be saved and used for analysis of other datasets with the same record layout and codes. The application should allow the Center to preserve the links between individual files from a relational database system. Researchers may eventually be able to formulate and execute queries against the data. This will enable them to identify diverse files sharing common attributes that could be used to link the files. This capability would allow researchers to perform analyses that were not contemplated by the agency or agencies that produced the files. The application should be able to produce public use files for those files that contain restricted data.

In addition to validating the datasets, the accessioning archivist must process the documentation that accompanies the data. The archivist must first verify that the documentation is adequate. While a record layout, code book, and technical data about the tapes used for transferring the records may be all that is necessary to validate

the datasets, these files are electronic records of the agency that produced them. The archivist must make sure there is enough information to document who used the data, how they used the data, and what impact the use of these records had on the agency. Sometimes this information has not been recorded in any form. In this case the archivist may have to interview agency personnel to gather this information.

The archivist must take steps to ensure the documentation will be preserved for the long term. Documentation in paper form is placed in acid-free folders and boxes. Preservation photocopies are made of unstable documents such as newspaper clippings and fax pages. Electronic documentation is subject to the same preservation regimen as the electronic records themselves.

Once the accessioning archivist is satisfied that the data matches the documentation and the documentation is adequate, two copies are made of each dataset. These copies are compared with the original records to verify that they are true copies of the files. The original tapes are then returned to the agency.

At this point the accessioning archivist prepares a final accessioning report. This report lists all the datasets received and processed, identifies any problems that remain to be resolved, suggests how the records should be described, identifies any restrictions on the use of the data, and recommends that the SF 258 should be signed by the appropriate parties to complete the transfer. Once the SF 258 has been signed and copies distributed to the appropriate parties, the accession is completed.

Agencies are required to take steps necessary to ensure the preservation of all unscheduled or permanent electronic records in their custody. When an agency transfers their records to the Center for Electronic Records, the Center assumes responsibility for their preservation. Under the provisions of 36 CFR 1234.28, unscheduled or permanent electronic records must, among other things, be stored under tight temperature and humidity controls. Tapes have to be rewound under controlled tension every 3 1/2 years. The agency must test a sample of their tapes for data loss annually. All records must be copied onto new media at least once every ten years.

These requirements serve as an incentive for agencies to schedule all their electronic records and transfer their permanent records to the Center for Electronic Records on a timely basis as well as ensuring against loss of data in the agencies. At the same time they provide for the preservation of the records until the agency does schedule or transfer their records.

Once electronic records have been accessioned, the

Center staff takes steps to provide access to the records. The records are described. The Center forwards descriptions of new accessions to Archival Publications and Accessions Control. This unit maintains a centralized database of descriptions of records accessioned by units of the Office of the National Archives.

The Office of the National Archives is currently developing a new centralized computing system. The Archival Information System (AIS), as presently planned, will include a module for detailed description of electronic records. When AIS is fully operational researchers will be able to query the database interactively to find records related to their research by using one or more sophisticated search paths.

The reference staff maintains the "Partial and Preliminary Title List of Holdings". This list contains the titles and some information about the availability of some of the datasets that the Center has accessioned. This list is available upon request from the reference staff. In addition to the "Title List", the reference staff produces finding aids to particular collections of data that may be of interest to a wide audience. The reference staff has in depth knowledge of the Center's holdings and offers personalized service to the records.

The Center for Electronic Records does not presently provide on-line access to its holdings. At this time a researcher may purchase a copy of a dataset of interest on 1/2 inch open reel tape.

The Center plans to make datasets available on 3480 cartridges this summer. The Center is also investigating the possibilities for making datasets available on media other than 1/2 inch open-reel tape and 3480 cartridges. The Center does plan to make some of its datasets available for on-line access within the next three years.

This has been a brief overview of the procedures used by the Center for Electronic Records in appraising electronic records of the United States Government, accessioning those records determined to have lasting value, preserving those records that have been acquired, and providing reference services for the records. These procedures are constantly being re-examined and revised in an effort to better meet our responsibilities.

1. Paper Presented at IASSIST 92, Madison, Wisconsin.

---

# Electronic Media and Preservation

---

by Fynnette Eaton<sup>1</sup>  
Chief, Technical Services  
Center for Electronic Records  
U.S. National Archives and Records  
Administration

The advent of the microcomputer has introduced great uncertainty to those persons responsible for ensuring the viability of information created on a computer. Yes, magnetic tape, round tape, has been recognized for years as a fragile medium. But at least there were recognized standards for this media. If someone orders a file from ICPSR and they sent a tape to the requestor's university, the information would be accessed easily by that university's mainframe.

But oh, what choices have appeared in the last five to ten years. I am willing to wager that each of you have come into contact with people who are using CD-ROM, WORM, Exabyte CD-recordable, diskette—I could go on. Since for most of us as archivists or librarians, the purchase or acceptance of files on these various media imply a commitment to ensure their accessibility to users for a certain period of time, the decision as to what media to use for storing information has a great impact on the computer operations of our institution.

Deciding what types of storage media to use involves determining the length of time your institution expects to have this information available, the computer resources currently available at your institution and your comfort level.

What I would like to discuss this morning are the general steps that should be considered by all institutions in seeking to preserve the information stored on electronic or optical media and then to briefly summarize the preservation requirements for the media currently in wide use. What do we mean when we say we are preserving information? How long will a product, a document or the information be preserved? I believe in many cases, there are underlying assumptions by the various groups that employ terms such as "archiving data" that are not necessarily shared by the larger community.

Permit me to use a personal story. When I began to work as an archivist with the Office of Presidential Libraries, I was responsible for preliminary preservation work for the color negatives from the Carter Administration. Since I had no previous experience in working with still photographs, I attended a workshop on preservation of photographic materials at the Eastman Kodak Institute. The information provided was timely and instructive. My

most vivid memory, however, was the demonstration of what happened to color film produced in the 1960's. We were viewing slides from the movie "West Side Story." As the lecturer explained the dye process used in manufacturing the film and showed a slide taken from one of the stills, the audience audibly gasped at what they saw: the projected image from the slide projector had only one color—magenta. One could almost feel the dismay as the audience realized that this famous cultural icon had deteriorated to the point that the only color on the film was red. Now, I will grant you that the basic information was still recorded, but for the millions of consumers who had been urged to buy this film to record all those wonderful family memories, would a film of children on vacation at the beach mean much to them if the entire photograph was red?

In many ways this is an unfair example, but it points to the fact that for many the concept of preservation is abstract and the question of longevity, unacknowledged.

It is the purpose of this paper to discuss how general preservation practices can and should be applied to electronic media and to suggest that for many institutions or organizations there is a need to carefully consider how long the information currently stored in an electronic format will be retained by that institution. That decision will influence the media used to store the information electronically.

First, let us discuss general preservation activities, as they relate to electronic records. When I began to research the basic preservation requirements for electronic records I was struck by how the requirements for textual materials seemed to mirror those for electronic records. I will be the first to admit that electronic records pose their own unusual problems, but, and it is a large but, the general maintenance and environmental requirements are very similar for textual and non-textual materials.

Many of the preservation policies that are constructed around attempts to prevent deterioration are just as relevant to electronic records as they are to paper based records. In their discussion of implementing an archival preservation program, Norvell Jones and Mary Lynn

Ritzenthaler detail the interrelated factors that cause archival records to deteriorate: the chemical and physical stability of specific materials, storage under adverse environmental conditions, and external causes such as excessive or careless handling, and loss or destruction brought about by human-induced or natural disasters. In every case the factors enunciated on this list are factors that must be considered in the preservation of electronic records as well.

An understanding of the physical properties of electronic records and the environmental conditions that they should be stored under are essential for ensuring that the information stored on these records are preserved.

The seven elements of a preservation program: environment, storage, handling and use; microreproduction and reformatting, exhibition, disaster planning and treatment must be considered by an institution charged with preserving electronic records. The only element that does not have real importance in electronic records is exhibition.

As with other media, perhaps the single most important factor in the preservation of electronic media is the environment. Electronic records, like other audiovisual records require temperatures between 62-68 degrees Fahrenheit, with an optimum of 65 degrees, which is probably within the range required for textual records. The humidity requirements, however, are different for magnetic tape than for paper. Lower humidity between 35 and 45 percent, with an optimum of 40 percent is the recommended level according to the National Institute of Standards and Technology (formerly the National Bureau of Standards), but this is less than the 50 percent recommended for paper records. According to George Cunha, the commonly accepted view currently held is if audiovisual materials (including magnetic tape) cannot be isolated in a mini-environment, then the overall humidity in the building should be kept between 40 percent and 50 percent."

Successfully attaining the optimum environment recommended can be difficult. Most institutions have conflicting requirements for staff and various media. One must recognize the difficulty of creating the perfect environment with competing interests, and take to heart what one conservation authority has learned: "it is far more important to stabilize both temperature and humidity at points as near as possible to the optimum conditions than to strive for optimum conditions with heating and cooling machinery that is unequal to the task and likely to produce constantly fluctuating temperature and humidity levels."

I would like to emphasize this point as well. Studies

indicate that one of the major contributions reducing the life expectancy of magnetic tapes is fluctuating temperature and humidity. Strive for the best conditions possible, but emphasize stability rather than occasional optimum conditions.

The proper storage and handling of archival materials is an essential element in a preservation program, particularly for paper records; but, again, this is also applicable to electronic records. Proper storage includes placing open reel tapes in plastic canisters and storing these tapes or cartridges vertically in shelving constructed specifically for open tape reels or tape cartridges. Unlike paper, which can be stored indefinitely if placed the proper containers, reels should be exercised periodically (there is discussion as to how often this should be done) and there should be a periodic inspection of a random sample of files, to test the readability of the media. An interesting theory proposed by Margaret Adams, who oversees the reference activities at the Center for Electronic Records is that, unlike paper records, reference activity actively promotes preservation in electronic records, because the staff uses the files, thereby determining the readability of that specific file and the media is cleaned and rewound after use, thus ensuring proper tensioning of the media.

Improper handling can have disastrous effects on magnetic tape. Dirt can create read errors. If the tapes are not tensioned properly, stretching can occur, which would create misalignment, leading to the inability of the computer to process the tape. Any distortion of the data due to improper tension or shrinking or expansion of the tape, or erasure of the tape can lead to the loss of the information stored on the tape. Improper handling of magnetic tapes or tape cartridges can cause edge damage as well. Thus procedures for ensuring the proper handling of electronic media must be an integral part of a preservation program for electronic records.

Reformatting, the next element in a preservation program is absolutely essential with electronic records. The requirement of moving electronic records to new formats is to keep up with the ever-changing technology. As the National Research Council pointed out in their study "Preservation of Historical Records" and the National Institute of Standards and Technology (NIST) has confirmed, the recording media in use may well outlast the hardware, thus making it necessary to recopy the electronic file every 10 to 20 years to ensure access to the information. This recopying process simply reformats the information to avoid obsolescence. The information is not changed in any way.

Disaster planning must be a part of any preservation program. Electronic records are susceptible to water and

fire damage. The best way to protect the information in electronic format is by making a second copy of any file and storing it offsite. The costs of a second copy are minimal compared to the expenses that would be incurred in trying to recreate the data. It is highly recommended that there be a second copy of any file stored in any electronic format, even ( and I would say particularly) diskettes.

Treatment, the last element discussed by Jones and Ritzenthaler, does not figure as prominently with electronic records, although the National Archives recently encountered problems with some of its older tapes and is working with the National Media Lab, in Minneapolis, Minnesota to find a way to salvage as much of the information from these tapes as possible. Generally, the best method of treatment is prevention, recopying electronic files before serious problems develop.

These then are the basic elements of a preservation program for archival materials. I have focused on their relevance to magnetic media. But what about the other media available on the market? Do CD-ROMS, optical disk systems and diskettes require the same type of program? Generally, I would say the answer is yes. Optical disks and CD-ROMs have been touted as being extremely durable. Perhaps yes, perhaps, no. There has been little empirical testing performed on these media. What you have heard are vendor claims and some horror stories. In certain cases, the seal on the CD-ROM was not perfect, so oxidation occurred and information was lost. There are indications that information stored on the outer layers of optical disks tend to have greater proportions of errors. Nothing is failsafe. Clean environments should be required for any media. Temperature and humidity should be controlled for best possible results. Disasters must always be planned for, so there should be a backup copy of any file that you are required to preserve. I must admit however, that there is no requirement for cleaning and rewinding of optical media.

Is the data permanent on these media? No. But the reason is not necessarily the medium. Some of these disks could well last 100 years. It is the technology that will fail. As I was preparing this paper I received a publication entitled *Government Imaging*, which claims the title of "The National Newspaper for Government Imaging Technology." In an article about standards for optical disk storage systems there is the clear acknowledgement that optical disks are not necessarily the best media for archival storage. In discussing the various standards used within document imaging systems, the author (Harvey Spencer) states the problem being "... that we are relying on these disks being available to us in twenty or maybe more years time and it is highly likely that the drives, and formats, that we are writing in will no

longer be supported. . ." He goes on to explain that the only standard that has survived from the 1960s is the 1/2" magnetic tape. The reason for its durability was the domination of the computer industry by a handful of suppliers that everyone used; the amount of information stored on these tapes is so great that manufacturers can not abandon this format. For optical disk systems, this situation does not exist.

CD-ROMS look more promising, because of the number of files being published on this media, and the acceptance by the library community as a means of information distribution. There have been questions for a number of years about the longevity of the polycarbonate CD media. An organization which is interested in promoting the use of CD-ROMs by government agencies, SIGCAT or Special Interest Group on CD-ROM Applications and Technology, is trying to collect information on this issue. One member, Ron Kushnier, a storage specialist with the Naval Air Warfare Center in Warminster, Pa reported to SIGCAT members last spring about his extensive environmental tests of CD media from about 100 manufacturers. What he found was that "All CD-ROMS are not created equal." Some disks came out of high-humidity, high temperature chambers in as good shape as they went in; others failed miserably. SIGCAT is continuing its efforts to determine longevity for this medium. Yet there is again the issue of standard, or I should say, the lack thereof.

Charles Dollar, a member of the Archival Research and Evaluation Staff at the National Archives has argued that disk longevity takes second place to "a much more important and pervasive issue—how to deal with technology-dependent records," Dollar used as an example relevant for CD's—data compression. Although there is an international standard for data compression, many vendors use proprietary compression techniques "that in essence become an encryption tool that only one vendor's software can open or close."

The point that I am trying to make with this discussion of the limitations of various media, is that you and your institution should consciously decide how long you intend to preserve information in an electronic format and base the decision of the which format on the length of time you will need access to the media. If it falls within ten to twenty years, then optical disk or CD-ROM is a valid choice, although you must monitor changes in technology in the marketplace and the condition of the equipment you use to access this information. If the requirement is for longer-term preservation, you can still use CD-ROMs and optical disks, but you must plan to reformat the information onto a technology that can be accessed in the future. There is no panacea for electronic media. It is a very small cost of migrating files to newer

format, to preserve previously unimaginable amounts of information and making this available to a world community.

1. Paper presented at IASSIST 93 in Edinburgh.

### Sources

Norvell M.M. Jones and Mary Lynn Ritzenthaler, "Implementing an Archival Preservation Program," in Managing Archives and Archival Institutions, ed. James Gregory Bradsher (Chicago: The University of Chicago Press, 1988), 188.

George Martin Cunha, "Current Trends in Preservation Research and Development," The American Archivist, vol. 53, no.2, Spring 1990, 195. Sidney B. Geller, Care and Handling of Computer Magnetic Storage Media, National Bureau of Standards Special Publication 200-101, (Washington, D.C.: National Bureau of Standards, 1983), 86.

Cunha, p. 195.

Jones and Ritzenthaler, 191-2.

Bruce I. Ambacher, "Managing Machine-Readable Archives," Managing Archives and Archival Institutions, ed. James Gregory Bradsher (Chicago: University of Chicago Press, 1988), 124-5.

National Research Council, Preservation of Historical Records, (Washington, D.C.: National Academy Press, 1986), 61-2.

Harvey Spencer, "Standards for Optical Disk Storage Systems," Government Imaging, vol. 2 no. 3, May - June 1993, 15.

Florence Olsen, "Experts debate longevity of polycarbonate CD media," Government Computer News, June 8, 1992.

Ibid.

---

# **“The Library of Congress at a Glance”: Text Visualization and Reference Rooms Without Walls**

by Lee A. Gladwin<sup>1</sup>

*Center for Electronic Records,  
National Archives and Records Administration*

You should “be able to see the Library of Congress at a glance” declared Ben Shneiderman, Head of the Human-Computer Interaction Laboratory at the University of Maryland. He was one of many presenters at the Advanced Information Processing & Analysis Symposium which was held at Tyson’s Corner, Virginia between March 22nd and March 24th, 1994. The symposium was organized by the Advanced Information Processing & Analysis Steering Group which represents the intelligence community. A primary goal of this organization is to provide liaison between intelligence analysts and potential contractors. Analysts are responsible for assimilating and synthesizing information from a variety of sources and producing digests of the request in timely fashion. In effect, they face the same overwhelming flood of information that all researchers do. They need some technology which allows them to visualize the search environment at a glance, filter out irrelevant information and focus in on what is critical to their tasks. Text visualization is one of the technologies being explored.

Shneiderman provided an example of how the University of Maryland Library holdings could be represented by their Macintosh-based TreeViz program as a series of 100 boxes on a computer screen. Each box represented a Dewey decimal-based classification, such as History, Science or Philosophy. The size of the box was proportional to the size of the particular holdings. Boxes were color coded to indicate rate of use. Boxes were arranged in hierarchies, so that one could descend from a general box, such as History, to a more detailed box-representation of books in various fields of history. Not all intuitive user interface designers agree that more is better. The screen design suggested by Shneiderman could be overwhelming for some users. Other designs and retrieval technologies are examined below.

## Text Retrieval and Visualization Systems

Since the majority of information requested by users is textual in nature, a great deal of research is being done to find improved ways of clarifying the researcher’s information needs (queries) and comparing them with text content in the database. The effectiveness of any text retrieval system is the degree to which it satisfies “an information need” [Croft, 9]. Does it retrieve most of the relevant documents from the prospective document pool? To satisfy these criteria, any system must be capable of “representing a user’s information problem or need, representing the content of text documents, and comparing these representations to decide which documents should be retrieved” [Croft, 9]. There are two approaches: statistical and knowledge-based. The former is based upon the notion that words occurring less frequently in documents are more important than those frequently found. Knowledge-based approaches are more concerned with the human aspects of information retrieval [Croft, 10]. These text retrieval approaches are fundamental to text visualization.

Before a document collection can be visualized, each document must be retrievable. To be retrievable, each must be represented within the retrieval system. Conventionally, documents are represented in an inverted index file. Each keyword is individually indexed and cross-referenced with the documents containing it. An inverted file is the “set of indexes for all allowable terms and attribute values” [Salton, 232]. This file might be imagined as a spreadsheet with column headers consisting of document identifiers and intersecting rows of keywords. At each column-row intersect would be a “O” if the terms does not appear in the document or a “1” if it does. A further refinement is the addition of term locations to the index giving the frequency of keyword occurrence in a document, and the precise location of the term; e.g., Document 350, paragraph 11, sentence 3, word 7. Since not all terms are created equal, keywords are weighted by their importance within the document. Frequently occurring words such as “the” or “and” receive lesser weights than infrequently occurring keywords such as nouns. Keyword weights, therefore, are also added to the index [Salton, 231 - 239]. Indexing may be done manually by human experts or automatically by a program such as those described below.

When indexing is done automatically, keywords are identified, their weights computed and combined to produce a document vector that may be compared with other document vectors to compute the degree of similarity and provide a

basis for graphical mapping of document relationships [Salton, 275 - 290, 304 - 308]. Following the description of both documents and queries in terms of their vectors, retrieval is executed on the basis of a "computation of query-record similarities" [Salton, 275].

TASC's TEXTVIZ program follows the knowledge-based approach. Multiple documents are first scanned via an optical Character Recognition (OCR) device into the database from which "features" (symbols, keywords, phrases such as "leaders, "house", "Bogota") are then extracted using a natural language processor. Features are then converted to "feature vectors" or unique identification codes which allow the system to compute "the degree to which a specific concept is correlated with a document" [Textviz, 2]. Text features or concepts may then be "mapped to points in a graphical space (text map)" where documents dealing with similar concepts are clustered together. Dissimilar documents are spaced farther apart on the screen. In TASC's TEXTVIZ program, key words appear on the map. In other systems, documents or concepts may be represented as alpha-glyphs, tadpole-like icons with attached lines or "tails" pointing in the direction of similarity.

HNC, Inc took a statistical approach. Their MatchPlus program employs neural network technology to learn database vocabulary by first discovering "similarity of usage at a word level, in a language-independent manner, without the need for external dictionaries, thesauri or semantic networks" [Caid & Carleton, 2]. The neural network generates word vectors which represent document content. Words learned in a given context point to or cluster with other related terms used in such contexts as weather, finance or government [Caid & Carleton, 3]. The next step is to represent documents in terms of "the weighted sum of the context vectors associated with words in the document" [Caid & Carleton, 5]. Documents dealing with similar topics are clustered or indexed for easier scanning. Document retrieval is accomplished by converting the user's natural language query into a context vector and searching for the nearest matching document vectors [Caid & Carleton, 7].

A statistical approach was also adopted by the National Security Agency's ACQUAINTANCE program which employs a "language-independent n-gram method of sorting and retrieving documents by language and topics. N-grams refers to "sequences of n consecutive characters" [Damashek, 39]. PARENTAGE, a visualization program, is used to explore the retrieved documents (See below).

Although not exhibited at the conference, it should be noted that GE Research and Development Center's NLDB text-based information system is perhaps the first to employ a hybrid approach to text retrieval. NLDB imitates human indexers and "automatically assigns categories to news stories for dissemination, retrieval, and browsing" [Jacobs]. Based on recall and precision criteria (retrieval of a high proportion of all relevant documents), their tests show that a combined knowledge-based and statistical approach to term categorization is superior to using either method alone.

### Intuitive User Interfaces (IUI)

Regardless of the technology used to process, index and extract information from textual sources, it is the interface with which the user must interact. James A. Wise, Battelle Pacific Northwest Laboratory, stated that an IUI is only "intuitive" to the degree that it exceeds the bounds of syntax". He emphasized that it must capture and communicate "the essence of meanings in messages through both the verbal and nonverbal domains", utilizing "analogs of the kinds of things that inform intuitions in everyday life". This interface must allow the user to instantly grasp the breadth of the information environment, easily filter out irrelevant material and focus search upon the most promising areas.

Several approaches to displaying the information domain were described above: proportional-colored boxes, concept maps and alpha-glyphs. Starfields, appearing like explosions of multicolored confetti, may be used in conjunction with a

legend at the bottom of the screen to indicate what the colors represent. In a library setting, for example, “blue” could indicate a mystery novel or film.

Viewers of the National Security Agency’s PARENTAGE program first see a screen covered with various concentrations of dots. This is the overview. To filter out some of these points, the user may click on a label in a menu beside a given dot cluster. This results in a cross-sectional view resembling a concept map in which related document nodes are linked by lines of varied thickness depending upon the strength of the relation. Document clusters may be searched by using a query template containing one or more labels and setting that label equal to a specific value, such as Profession s Engineer (See Figure 1). This results in a list of documents or a display of objects from which to make final selections [Cohen, 115].

Logicon’s BROWSER is designed for the user who says, “I don’t know what the evidence is likely to be. I’ll know it when I see it.” A document folder metaphor is used to aid researchers in grasping quickly how the system works. The user begins by looking at a list of folder names. Opening a folder will display a list of subject lines. Clicking on a subject line provides information about a specific document. The user may select a document and then click on a word highlighted in a text in order to see “a list of folders that include documents containing the term, bring up a list of documents in any folder on that list, and open any document”.

All of these approaches sounded terribly futuristic to attendees until they stepped into the exhibit hall and saw actual demonstrations of the systems discussed in the presentations.

#### A Paradigmatic Shift in How We View Information

For those using these new interface designs, a major shift in how we think about information is required. In the Gutenberg galaxy, information was organized linearly. Text and pictures followed in an orderly succession and was searched from beginning to end. In the electronic age, paper, pictures, movies, and sound recordings are collections of objects adrift in cyberspace which can be manipulated by individuals. Information is reduced to related data chunks which may be viewed contextually through text visualization [Hilbing, 54]. Order is initially imposed by various algorithms, saving the user great time and effort. In a way, the concept of information as inter-related chunks of data is analogous to the organization and storage of information in the human brain’s neural network hierarchies. Learning is the formation and modification of neural synapses in the process of forming more complex structures. HNC, Inc’s MatchPlus forms text associations in this manner. Visualization techniques transform associated data chunks into a cohesive visual representation that can be understood by the user.

#### A Paradigmatic Shift in Reference Support Services

Text visualization and associated intelligent systems will transform not only how the researcher locates information, but traditional reference support services as well. Since users will be able to search library holdings and documents on their own, only the most difficult reference work will need to be performed by staff [Hayes, 4].

Through text visualization and related technologies, researchers will be able to view vast amounts of data at a glance, focus their search and explore areas relevant to their interests. This will allow them greater time to interpret, analyze and report information than they have ever had. Reference staff will be free to assist researchers with more challenging reference problems. The intelligent technologies under development for the intelligence community today will be in our reference rooms tomorrow. While there are still problems to be solved before these systems become generally available, the day of the electronic reference room without walls is closer than we may wish to think.



**References.**

Caid, William R. and Joel L. Carleton. "Context Vector-Based Text Retrieval" (HNC, Inc. nd). Working paper supplement to paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Carlotto, Mark J. "Text Visualization" (TASC, 1992). Paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Cohen, Jonathan, "Parentage". Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994.

Combs, Nathan H. "Large Text Database Visualization" (TASC, 1992). Paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Croft, W. Bruce. "Knowledge-based and Statistical Approaches to Text Retrieval", IEEE Expert (April, 1993).

Damashek, Marc. "ACQUAINTANCE". Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994--

Hayes, Phil. "Knowledge-based Systems for the Information Industry", IEEE Expert (April, 1993).

Hilbing, Capt. John F. "Electronic Production in the Post Paradigm Shift World". Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994.

Jacobs, Paul S. "Using Statistical Methods to Improve Knowledge-Based News Categorization", IEEE Expert (April, 1993).

Meadow, Charles T. Text Information Retrieval Systems (NY: Academic Press, 1992).

Proceedings. Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994. Sponsored by Advanced Information Processing and Analysis Steering Group, Intelligence Community.

Salton, Gerard. Automatic Text Processing: The Transformation Analysis, and Retrieval of Information by Computer (Reading, MA: Addison-Wesley, 1989).

Sasseen, Robert V. and William R. Caid. "Docuverse: A Context Vector-Based Approach to Graphical Representation of Information Content" (HNC, Inc. nd)

Textviz, "Visualization of Large Text Document Databases: Suggested Statement of Work for Text Visualization (TEXTVIZ) Development and Demonstration (27 February 1992). Paper accompanying presentation at AIPASG Symposium, 1994.

NOTE: Descriptions of exemplars of these systems should not be construed as endorsement of any particular system or retrieval-visualization methods by either NARA or the author. All opinions expressed are those of the author and do not reflect NARA policy or programs.

## ISBD(CF) REVIEW GROUP Meeting of April 24-26, 1995

---

### *~ Summary Report ~ John D. Byrum*

The ISBD(CF) Review Group meet at the Library of Congress April 24-26, 1995 to consider a revised version of the text of the International Standard Bibliographic Description for Computer Files (1990) prepared at the chairman's request by Ann Sandberg-Fox who is serving as principal editor of the Second Edition. In attendance at this meeting were Group members Sten Hedberg (Uppsala Universitetsbibliotek); Catherine Marandas (Bibliothèque nationale de France); Ms. Sandberg-Fox (Colchester, Vermont); chairman John Byrum (Library of Congress) as well as corresponding members Laurel Jizba (Michigan State University Libraries) and Lucy Evans (British Library) as well as observer Claire Vayssade (Bibliothèque nationale de France). The meeting was made possible by a subsidy from IFLA and a grant from the Research Libraries Group (RLG).

The first day was devoted to discussion of several issues-papers which Ms. Sandberg-Fox had prepared. These covered the topics most in need of reconsideration in the light of the rapidly developing technology which has influenced the creation and dissemination of Computer Files: Interactive multimedia; the General Material Designation (GMD); Sources of information; Reproduction and multiple versions; Designation of file; and, Published versus unpublished remote texts. In addition other aspects, such as Preliminaries, Type and extent of file, Physical description and notes were thoroughly discussed, as were a number of proposals received by the chair prior and subsequent to the formation of the Review Group. On the second and third days, the members focused on a close reading of the revision prepared by Ms. Sandberg-Fox, with the result that an agreed upon text emerged from the meeting. The draft will now be updated to incorporate decisions taken at this gathering and, with permission of the Sections on Cataloguing and on Information Technology, presented for world-wide review on or about September 1, 1995. Following a six-month comment period, a final version of ISBD(CF) Second Edition will be readied for IFLA approval and publication; in addition, the text will be shared with the authors of national and international cataloguing codes, such as the Joint Steering Committee for AACR.

Following is a brief summary of the most important outcomes of the April 24- 26 meeting and which will be reflected in the revised ISBD(CF), presented in terms of the objectives that were set out to guide this project:

(1) To take into account the emergence of interactive multimedia, a new and still developing technology that combines and stores products of audio and video technologies, together with text and graphics, on optical discs.

Regarding interactive multimedia, the Review Group concluded that all such resources be incorporated into the new version of CF. This conclusion was reached because no existing ISBD covers these materials (which entered the mass market beginning in the mid-1980's), and because user-manipulated, non-linear navigation using computer-controlled technology are hallmarks which characterize interactive multimedia. (These materials are distinct from multimedia/kits that are covered by the stipulations of ISBD(NBM).) As a result, the new version of CF will add or amend provisions regarding sources of information (0.5), edition (area 2), type and extent of file (area 3), dates (area 4), physical description (area 5) and the notes (area 7) to show treatment of interactive multimedia as a subset of computer files. Examples will be added to illustrate such files.

(2) To consider the impact of developments in optical technology, as new and improved optical discs are replacing magnetic disks as primary storage devices.

The Review Group decided to improve CF to cover not only CD-ROMs (compact disc read-only memory) but also CD-I's (compact disc interactive), and other emergent forms such as photo-optical compact disc. As a result, the new version of CF will add or amend provision regarding sources of information (0.5), edition (area 2), physical description

(area 5), and notes (area 7). The term “disk” (spelled with “k”), currently used throughout area 5 to describe both optical and magnetic devices, will now apply only to magnetic devices, while “disc” (spelled with “c”) will be used in relation to optical manifestations.

(3) To provide for the availability of remote electronic files on the Internet, a global network of networks that allows users access to a vast wealth of remote electronic files, including books, journals, articles, reference sources, and even library catalogs.

Since, at the time CF was first formulated, this was a new area especially designed to treat these files, caution was exercised as to the kind and amount of detail to be given. Designations of the type of file are limited to general terms only—“Data” and “Program” and their combination “Data and program.” The Review Group decided that these terms are not adequate for the purposes of identifying the many different types of data files and software on the Internet. Indeed, the whole treatment of the Designation of file was thoroughly reworked and developed, with area 3 emerging as the one most thoroughly changed in revised CF. Consequently, the Second Edition of CF will propose several levels of specificity as appropriate. The current terms “Data” and “Program” will continue to be authorized, but Data files can alternatively be indicated as “Numeric”, “Text”, “Pictorial”, “Representational” or “Sound”, while Programs can be identified as “Utility”, “Application” or “System”. Most of these categories are further delineated for more specific designation when appropriate; for example, a Bibliographic database may be so identified, as may be a Game. As before, the combination “Data and Program(s)” will continue to be used when applicable. However, alternative identification as to particular types of data and program(s) may be taken from the authorized listing and be used in conjunction with the following terms: “Interactive multimedia” or “Online service.” These latter terms also function as designations when terms from the authorized listing are not appropriate. Where, in the case of combinations, the program or the data may be incidental to the whole, the primary term only is to be given. As for the General Material Designation (GMD), the Group decided to retain “Computer File” in the absence of a better alternative.

Further addressing Internet resources, the revised CF will provide better treatment of the networking environment where an electronic file may be accessed by several methods, reside in many directories, and require more detailed information, enabling users to locate and retrieve these files. Specifically, CF will be updated to include provision for URL’s, gopher and FTP sites.

c (4) To deal with bibliographic problems arising from reproductions of computer files such that many CF titles are now available in a variety of physical formats.

Although such problems are not easily resolved, the CF Review Group did authorize changes to areas 2 and 5 to better distinguish between an “original” and other versions thereof. Reformatting changes were moved from inclusion in the definition of edition to inclusion, instead, into the definition of what would not constitute a new edition. Also, output medium and display format are newly reworked phrases to better reflect CF technology.

In addition, the Review Group agreed to significant modifications of the provisions concerning Sources of information (0.5). Area 4 (“Publication”) will be amended to require treatment of all remote CF as published materials. In addition, the glossary and examples will be updated and increased.

In the course of its meeting, as requested, the Group considered the Official Draft Proposal of the IFLA Division of Bibliographic Control Study Group on the Functional Requirements for Bibliographic Records, with Barbara Tillett, one of the three consultants to that project, present for part of the discussion. It was decided that as a medium, computer files would provide a good test of the draft, and the Group agreed to undertake an in-depth study. Specifically, 1) the use of the words “item” and “work” in the Functional Requirements document will be examined in relationship to related terminology in the ISBD(CF); 2) an experiment will be conducted to apply the suggested model using several types of computer files in several library environments; 3) the results of the experiment will be analyzed; and 4) a summary document, including any potential recommendations for the ISBD(CF) will be written. Laurel Jizba will coordinate this study for presentation by November 1, 1995.







INTERNATIONAL ASSOCIATION FOR  
SOCIAL SCIENCE INFORMATION  
SERVICE AND TECHNOLOGY

• • • • •  
ASSOCIATION INTERNATIONALE  
POUR LES SERVICES ET  
TECHNIQUES D'INFORMATION EN  
SCIENCES SOCIALES

## Membership form

The International Association for Social Science Information Services and Technology (**IASSIST**) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from re-

duced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:

Regular Membership. \$40.00 per calendar year.

Student Membership: \$20.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription:

\$70.00 per calendar year (includes one volume of the Quarterly)

**I would like to become a member of IASSIST. Please see my choice below:**

- \$40 Regular Membership
- \$20 Student Membership
- \$70 Institutional Membership

**My primary Interests are:**

- Archive Services/Administration
- Data Processing
- Data Management
- Research Applications
- Other (specify) \_\_\_\_\_

**Please make checks payable to IASSIST and Mail to :**  
**Mr. Marty Pawlocki**  
**Treasurer, IASSIST**  
**% 303 GSLIS Building,**  
**Social Science Data**  
**Archives, University of**  
**California, 405 Hilgard**  
**Avenue, Los Angeles, CA**  
**90024-1484**

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone



