

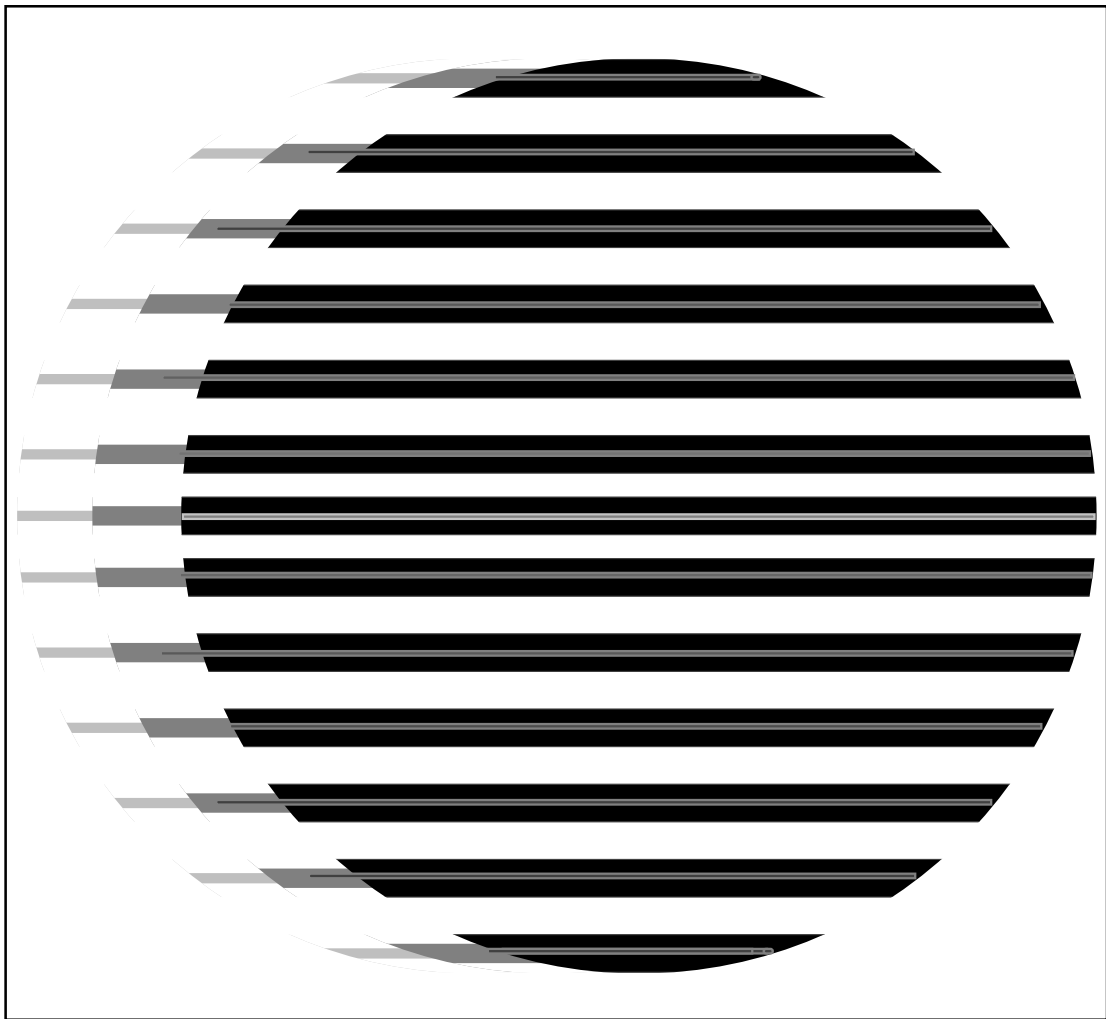
# IASSIST

Q U A R T E R L Y

VOLUME 23

Winter 1999

NUMBER 4



---

Printed in the USA

---

# 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. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu). Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

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. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

## Editors

**Karsten Boye Rasmussen,**  
Eckersbergsvej 56,  
5230 Odense M,  
Denmark.  
Phone: +45 6612 9811  
Email: [kbr@sam.sdu.dk](mailto:kbr@sam.sdu.dk)

**Juri Stratford**  
Government Information and  
Maps Department,  
ShieldsLibrary,  
University of California,  
100 North WestQuad,  
Davis, California 95616-5292  
Phone: (530) 752-1624.  
Email: [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu)

## Production

**William Block,**  
25 Blegen Hall,  
269 19th Ave S,  
University of Minnesota,  
Minneapolis, MN 55455.  
Phone: 612-624-7091.  
Email: [block@socsci.umn.edu](mailto:block@socsci.umn.edu)

**Walter Piovesan**  
Research Data Library,  
Simon Fraser University,  
Burnaby, B.C  
Canada V5A 1S6.  
Phone: (604) 291-5869.  
Email: [walter@sfu.ca](mailto:walter@sfu.ca)

**Title: Newsletter - International Association for Social  
Science Information Service and Technology**

**ISSN - United States: 0739-1137 © 1999 by IASSIST. All  
rights reserved.**

# C O N T E N T S

Volume 23

Number 4

Winter 1999



## FEATURES

- 4** **Automated Preservation of Electronic  
Records: A Case Study of the Archival  
Preservation System**  
*Fynnette Eaton*
- 10** **Maximizing the Search Potential of Social  
Science Codebooks Through the Application  
of the Codebook DTD**  
*Wendy Treadwell*

# Automated Preservation of Electronic Records: A Case Study of the Archival Preservation System

The National Archives and Records Administration has had a program for accessioning, describing, preserving, and providing reference service to the electronic records (Machine Readable records) created by Federal agencies and transferred to the National Archives for almost thirty years. Although there have been many changes in the name of the office, its basic mission has remained the same: to preserve and make available those records created by Federal agencies in electronic format that the National Archives has determined to have value beyond the short-term need of the originating agency.

Most people think of the National Archives as the keeper of the Constitution and the Declaration of Independence. Even the most experienced researchers are largely unaware of the growing number of files in electronic format. Since the creation of the Center for Electronic Records in October 1988, the number of files transferred has literally skyrocketed. In 1988 the Archives received 150 files from Federal agencies. In fiscal year 1991, the number was 1500, ten times as many in three years. The numbers jumped again in 1992 to 8730 files. Unfortunately, the Center became involved in a resource-draining court case, which forced Tom Brown and his staff to reduce their efforts in accessioning new files, with a resultant decrease in accessions in FY94 and FY95 to 843 files and 1590 files respectively. Nevertheless this is a vast increase compared to earlier years.

Currently, the Center has accessioned about 23,000 files produced by over 100 bureaus, departments, and other components of executive branch agencies and their contractors. These files range from the American Soldier surveys of World War II to records of the 1980 and 1990 Decennial Censuses. These files include education data illustrating the variety of education programs of the Federal government; health and social science data incorporating both biomedical and sociological information and efforts to measure the effectiveness of a variety of social programs; international data including import-export statistics and USIA-sponsored surveys. The represented military data ranges from Prisoner of War records for World War II and the Korean conflict, and casualty records for the Korean and Vietnam conflicts, to a large collection of data files

*by Fynnette Eaton\**

resulting from the use of computers for military operations, management, and research dating from the 1960's especially during combat in Southeast Asia.

Clearly, as the size of our holdings grew, with the tremendous increase in transfers of files, the Center recognized the need to develop new methods for accessioning and preserving these new files. My paper will discuss the development and implementation of the Archival Preservation System (which I will refer to as APS). Another system, the Archival Electronic Records Inspection and Control system or AERIC was also developed in the early 1990's to provide automated validation of electronic files and to build a base of descriptive data drawn from the data elements of these files. But that is another paper, by a different author.

As I stated at the beginning of this paper, the Center has been involved with the various archival activities associated with electronic records for more than twenty years. But it was only in the late 1980's that the number of files being transferred overwhelmed the staff, requiring reexamination of the methods used to process these files. Permit me to give you a brief overview of how the staff used the resources that were available at the time to perform the basic preservation work at the National Archives.

During the 1970's all computer processing required by the Machine Readable Archives Division was performed at service bureaus. The division had an IBM 029 keypunch, which the staff used to punch cards for the programs to copy or dump tapes. The program card deck was wrapped in a rubber band along with a sheet of instructions to the service bureau indicating which tape volumes were to be used for input and output and any other special instructions. The card decks, instructions, input tapes, and blank output tapes were boxed and sent by courier to the service bureau.

The service bureau staffs usually ran the jobs at night with a twenty-four to forty-eight hour turnaround. The division preservation and reference staff checked the jobs, labeled the tapes, assigned the output tapes location numbers and took the tapes to the storage area in the Washington National Records Center in Suitland, Maryland.

During this period the original agency tapes was kept as the master tape and the NARS-created tape became the reference copy to be used as input tapes to make copies for researchers.

In 1975/1976 the transfer of DOD files written in NIPS (National Military Command System Information Processing System) software required special handling. GSA made available to NARS a copy of the NIPS software at the DC Share computer facility. Generally the staff transported the tapes and card decks to DC Share. Because of the faster turnaround time, staff began to use DC Share to run other jobs as well.

In 1981 the Machine Readable Archives Division acquired on loan a dumb terminal so that staff could run some jobs for the FBI Appraisal Task Force (another court case that severely strained NARS resources) at the National Institutes of Health (NIH). The Division took advantage of the access to this computer center by performing some of its normal copy and dump jobs at NIH.

Because of organizational changes within the National Archives and Records Service, the Machine Readable Archives Division became the Machine Readable Branch of the Special Archives Division in 1982. Plans to procure a minicomputer for its own use did not materialize and the Branch lost the terminal with access to the NIH computer center. By October 1982 all requests for preservation and reference work for computer files was submitted to another office for processing. Although reference work continued, only a few preservation jobs were successfully completed after this transfer.

In 1984 the Branch acquired a DECWriter terminal that the Motion Pictures Branch was surplus to use for its work on the Catalog of holdings. Beginning in January 1985, the Branch received authorization to use the terminal to access the National Institutes of Health computer center, as it assumed responsibility for preservation copying of its accessioned files. Within a few months the Branch obtained additional terminals for submitting jobs to the NIH computer center. The computer center at NIH served as the computer resources for all of the Branch's requirements.

This acknowledgment that the Machine Readable Branch should perform the work is borne out by the few statistics I could locate in the files. During the period Fiscal Year 82 through 84 less than 100 tapes were copied. The Machine Readable Branch staff copied over 200 tapes in six months of 1985. Another report indicates that in FY 87 104 reels had been copied by May. Yet these numbers were far too low, once the Center began its accelerated accessioning program.

The preservation work, which required making two copies

of each file offered by a Federal agency, was performed using the mainframe computers at the National Institutes of Health Computer Center in Bethesda, Maryland. This mainframe computer center used IBM machines, so the types of outputs that we could produce were limited to the options that were available to us at that Center. In addition, the file formats that we could accept for transfer were limited to those formats that we could process at this site. Our requirements, which were published in the *Code of Federal Regulations*, stated that agencies were to transfer permanent computer files in a hardware and software independent format. Specifically the files were to be written on half-inch magnetic tape in EBCDIC or ASCII, without internal control characters on 7 or 9 track open-reel magnetic tape recorded at 800, 1600 or 6250 bytes per inch or on 3480 cartridges and blocked not higher than 32,000 bytes.

These requirements clearly reflect the use of a mainframe computer in creation of our preservation copies of these files. Even though we had control over the preservation copying of our files, the Center was at a disadvantage. We had to relinquish physical control of both the agency tapes and the blank tapes or cartridges that we would use to make the preservation copies on when they were sent to the NIH computer center to be mounted on tape drives there, but that was the only real choice we had. We were able to streamline some of our copying procedures, but we often had to wait in line for access to tape drives because we were dependent on modems and phone lines to connect to the NIH Computer Center which had thousands of users. We determined that although the staff time required to copy and compare files created at NIH was between 2 1/4 and 4 hours, the time that elapsed from when we prepared the tapes to be transported to the computer center and their return after successful copying, was literally one week.

One of the first priorities enunciated by the Director, Ken Thibodeau, when he joined the Center in the winter of 1988/89, was to reengineer this process by developing an in-house capability for making preservation copies of electronic files sent to the Center by Federal agencies. By examining the processes associated with producing preservation copies of electronic files, the staff developed a statement of work for prospective contractors that defined the basic requirements and outlined additional features we would like to see developed over the life of the contract. Although we defined the processes based upon our current practice, we also recognized the opportunity to streamline some of the work that had almost developed a life of its own. The development of the statement of work took about a year. The National Archives issued a Request for Procurement in March 1992 and selected as the successful bidder, Muller Media Conversions of New York City in May 1992. Although in the statement of work we defined what the processes should be, we did not define how they should be accomplished.

There were four objectives with this contract. First, we wanted to retain control of the media and perform the work in-house, streamlining and saving valuable time and therefore increasing productivity. Second, we wanted to be able to handle a wider variety of file formats from agencies and produce standardized output. Third, we wanted to capture automatically from the processing of the files the technical attributes of the file, such as the logical record length, blocksize, character code and media. In effect, we wanted to automate the collection of technical description during the processing, rather than entering this information into a separate database (TAPES). Fourth, we sought to increase the types of media we could accept from agencies and, as well, increase the types of media on which we could output records.

Muller Media began developing the software which was the largest component of the contract. The estimated costs for the Archival Preservation System (APS) included the CPU unit, a 66 MHZ 486 IBM Value Point running on OS2, with two 9 track Overland tape drives, 2 Overland cartridge tape drives, cables, and bar code apparatus at a cost of \$203,165. The software development, \$129,500, was more than 63% of the contract cost. It had been our intention to develop a system, operate it for a certain period of time, and if it performed as we expected, to purchase additional systems as money permitted to increase our efficiency in copying files. We had one system, but I had six programmers; so we anticipated purchasing additional systems. However, we did not anticipate purchasing additional systems even before the software was developed, but life overtook plans.

On January 19, 1989--the last day of the Reagan Administration--Scott Armstrong, among others, filed Freedom of Information Act (FOIA) requests for information stored on the computer system in the offices of the President from its date of installation in 1985 until the end of the Reagan Administration. They sued the Government, including the National Archives, asking for the court to declare many of the materials on the system to be Federal and Presidential records. Until the issues could be resolved, the court ordered the Government not to destroy or alter any of the systems' backup computer tapes since they contained the only extant copies of some of the information on the system. The lawsuit carried on throughout the Bush Administration. On the day after the election, when Bush was defeated, the plaintiffs extended the lawsuit to include materials residing on the computer systems in the Bush White House as well. On January 6, 1993 -- two weeks before Bush formally left office -- the court ruled that some materials on the White House computer systems were Federal records and Presidential records and the court directed the Government, specifically, the Archivist of the United States, to "take all necessary steps to preserve, without erasure, all electronic Federal Records generated" by the White House

agencies. This court order was not to be taken lightly as events showed. The Archives took physical custody of all of the tapes from the White House as the Bush Administration was leaving and the Clinton Administration took office. In late May, the court ruled that the Archives had not complied with his order to preserve the tapes, and levied fines which would amount to \$2.5 million if the Archives did not come into compliance within thirty days. When the ruling also referred to "increases in . . . sanctions reserved . . . for any further noncompliance. . ." it meant the threat of possible jail time was real. While the fines were lifted, upon appeal, it was apparent that the threat of fines and possible imprisonment was very real.

Since the National Archives had acquired custody of the computers files and since the records in question were electronic, it was inevitable that the Center for Electronic Records would become involved in this case. This in fact happened in March 1993 when the Acting Archivist, Trudy Huskamp Peterson, transferred the responsibility for preserving these computer backup tapes from the Office of Presidential Libraries, which had had physical custody of these materials from the time they left the White House until late May, to the Center for Electronic Records within the Office of Special and Regional Archives.

While the APS had been conceptualized to expedite routine preservation processing, we had to use it first in the very non-routine processing of the materials from the White House. Responding to the crisis situation, the APS contractor Muller Media Conversions compiled enough software so that the staff with the requisite clearances were able to make duplicate copies of files in most cases. We encountered problems in copying some of the files and noted any problems. But we complied with the court orders and have successfully avoided legal sanctions. If the Center for Electronic Records had not previously developed the concept of the Archival Preservation System and did not have the APS on order, NARA would have been unable to comply with the court orders.

Unfortunately the work required by what became known as the Armstrong v. Executive Office of the President case overtook the development of the APS system for the next full year (1994). Refinements to the software were made so that problems that we initially encountered were either alleviated or at least better documented by the system. Over the next two years, more than 5900 tapes and/or cartridges were copied, using one or more versions of the Archival Preservation System software. I can state with pride that the Center successfully copied more than 99.998% of the media transferred from the White House. Out of 5906 items, only twenty nine unique items had a data error.

Although it had been anticipated that the full development

of the APS would take approximately 150 days, the requirements of the court case overshadowed the development of the full system. Nonetheless, the staff devoted many hours to developing the data elements for the catalog database, which was the only part of the system that had not been well defined in the contract. Basing the database on a preexisting system that was used to track the technical attributes of electronic files, known as TAPES, the staff sought to include the essential elements from the TAPES database, to capture preservation activities that were previously recorded in a second database (PRESLOG) and to capture information about the individual files as the APS processed the new files. These long staff meetings paid off because the database reflected the needs of the Branch in capturing the information necessary for tracking new media, the progress of preservation work, and the technical attributes of files processed on this system.

As I said, the contract was not completed as soon as we had anticipated, because we had to make adjustments to the APS system to enable the Center to meet the requirements of the Court order for preserving the backup files. There were problems that had to be overcome in processing these backup tapes from a variety of computer systems at the White House. In many cases the tapes had been written over, since they were used for weekly backups; so when we attempted to copy the files, there was information beyond the tape mark which meant we did not get a normal end of file mark to cease the copying operation. We were not able to determine what some of these problems were until we obtained greater functionality with the APS system. With the Bush cartridges our greatest problem was the fact that the backup utility had also used compression, so the APS system had to be modified so it could make duplicate copies of compressed files. Under our normal operating procedures, data in compressed formats do not conform to our transfer requirements.

Without APS the National Archives could not possibly have met the requirements set by the court, duplicating all of the backup tapes, thus ensuring preservation of whatever is found on these backup tapes. But the cost was the delay in full implementation of the APS for the "normal" processing for which this system was designed and purchased. In fact, the Center only accepted the system as meeting the basic requirements as outlined in the statement of work in Spring 1994. We are still working with a system that clearly is evolving. Currently we are moving the catalog database to a network, which will make the information available to the Center staff and increase the functionality of the system by being able to use any number of drives for performing copy jobs. Unfortunately for us most of the additional development in the APS system up until last September had been in refinements to address additional problems encountered in copying the White House system backup tapes.

Does that mean that APS is not meeting the needs of the Center in making preservation copies of electronic records? Absolutely not. We have been processing "normal" files on APS since the fall of 1994, whenever we were not processing files associated with the PROFS case. We have copied over 1,375 accessioned files using this system. It has greatly increased our ability to handle a wide variety of file formats, which we could not handle previously. One example is the 1990 Decennial Census Public Use Sample files that the Bureau of the Census has been transferring to us for more than two years. Although they are in a format that is hardware and software independent, user labels and the blocking factor used with these records, prevent us from being able to process these tapes at the NIH computer center.

Perhaps even more importantly, the APS provides the Center with a mechanism to accept files on wider variety of media. In using the NIH mainframe, we had two choices: records on 9 track tape or 3480-class cartridge. We have both of those options with APS, but we have ordered a CD-ROM drive to be installed, so that we can begin to copy scheduled electronic files transferred to us on CD-ROMs. We also want to have the ability to copy files from diskettes which was never possible previously. And, there are other forms of media which we might want to be able to access, which will be possible, by attaching drives to the system. For the classified PROFS files, for example, we had to install both a 4mm and 8mm drive, because some of the files were recorded on that type of media.

Perhaps even more importantly, the Center anticipates using the Archival Preservation System to make copies of files to fill reference requests. We have always used the National Institutes of Health Computer Center to make reference copies of electronic files. We have had to limit the choices of output to 9 track tape or 3480 cartridge. But most users now use personal computers or are attached to networks. In many cases researchers do not have access to tape drives. So, one of our goals is to use the APS to make copies of reference requests and to output some files on diskettes when appropriate, and possibly other media such as CD-ROM as well. We have just received the funding to purchase the system with a CD-R drive attached to the system. Again, it will mean that we will not lose physical control over our records. The tapes and/or cartridges will not be exposed to poor environmental conditions, while they are in transit between our vaults and the external computer center, and we can tailor the output to meet the needs of our customer base.

Has the Archival Preservation System been a success? It literally saved us from a contempt ruling in a contentious court case. We are employing it to make preservation copies of files. Our current objective is to improve the functionality of the catalog database, by moving it from

a Faircom server to ORACLE on a RISC6000 computer and integrating the two systems that are currently used to make copies of accessioned files. And, recognizing the utility of this system, we want to secure another system to make copies of files, and possibly extract of files, to fill reference requests. Do I regret that its development was delayed by the court case? Yes, but there is a silver lining. The APS system was able to deal with nonconforming media and software dependent files. The success we had in overcoming the problems posed by the backup computer tapes has given us greater confidence in being able to use the APS system beyond the narrow confines for which it was developed. The Center will be able to modify this system to meet the demands posed by the newer media being employed by Federal agencies. The APS is a viable system for preserving information into the twenty-first century.

\* Fynnette Eaton Smithsonian, Institution Archives. [This paper was presented at the May 1996 IASSIST meeting in Minneapolis, MN and reflects the situation at the National Archives through 1997. The author left NARA in May 1997 to join the Smithsonian Institution Archives, where she is currently employed.]



# Maximizing the Search Potential of Social Science Codebooks Through the Application of the Codebook DTD

by Wendy Treadwell\*

Data libraries and archives have been working with digitized materials longer than most general libraries and archives. However, we have been slow to develop a means of making our collections searchable in an electronic manner. This has been due primarily to the fact that our metadata (codebooks and data dictionaries) have not been available in a machine processable format. The complexity of the material and the need to be able to identify specific structural elements and their contents made even well enhanced bibliographic records inadequate to the task.

Researchers seeking data for secondary analysis have a distinct set of needs. They need the ability to:

- **Search across multiple collections in multiple locations.** With bibliographic records they are able to achieve this at the level of basic information, but cannot do so consistently or at the level of information needed.
- **Search heterogeneous collections.** In other words, they do not necessarily wish to search one system for data and another for related materials.
- **Drill down into individual collections and documents for more detailed information (in particular, detailed information regarding the variables in the data set).** The importance of being able to search at the variable and variable response category level is made clear in the following example. A great number of data sets, particularly those aggregated to small geographic levels, use age cohorts. Data published prior to 1980 frequently used upper age cohorts of '65 years and over'. This practice made them unsuitable for researchers examining the relationship between age and socioeconomic factors within the over-65 population. This piece of information was available only by looking at the response categories for the variable age. When most codebooks were only available in hard-copy, researchers would lose valuable time obtaining codebooks and data only to find that the data set was unusable for their purposes.

- **Search both the metadata and the object.** In the case of text documents this means the full-text of the document as well as its bibliographic or other metadata material. In terms of data this means examining both the data file documentation and the data itself.

- **Obtain or manipulate the file contents.**

The goal of the Data Documentation Initiative (DDI) group was to address the needs listed above. They needed to develop a machine readable and machine processable codebook which would fulfill both archival requirements and serve as a source for inquiry. The XML tagged codebook developed by the DDI addresses each of the issues noted above.

Searching across multiple collections becomes possible using a uniform configuration for the codebook. Creating centralized depositories for codebooks or search engines that can search multiple locations now become options.

By using XML tags, the DDI has adopted a tagging scheme commonly used in text documents of various types. Systems which can parse an XML DTD can search through often familiar layers of information. Many attributes of higher level metadata were retained, such as descriptive bibliographic fields. These were then mapped to commonly used schema like the Dublin Core. This makes searching across types of material more efficient. By providing links between the DDI tagged document and related materials, the codebook can also become a central hub through which other materials are identified and obtained.

Of course, the most important feature of the DDI DTD is that it identifies specific structural elements and their attributes. This allows the searcher to drill down into individual collections and documents for more detailed information. The extent of the tagging provided makes it possible to create specialized search engines which can address the eccentricities of both the researcher and the materials being searched.

The merger of the data and the metadata of the document (codebook) into a single unit results in the entire document becoming a resource for discovery. Researchers are no longer as dependent upon the descriptive skills of the cataloger or archivist to capture the concepts important to the individual researcher in a controlled language. External tools can become the driving force for relating past terminology with future terminology and past conceptual structures with future use and perspective.

Finally, the DDI tagged codebook provides all the information needed to create systems to obtain and/or manipulate data file contents. The identification of elements and attributes in a structured tag provides for both machine understanding and processing. This is a feature that has been absent from many earlier attempts to make codebooks machine readable.

The availability of tools such as the Generalized Record Structure 2 (GRS2) within Z39.50 protocol make DDI tagged codebooks potentially accessible through the same tools used for searching other tagged documents with DTD's. The GRS2 is designed to pass information regarding structure of materials using DTD's and structured tags within Z39.50 compliant systems. It provides the ability to map information such as a query from one set of tags to another. For example, the DDI DTD includes mapping information to Dublin Core elements (fig 1).

The GRS2 would be used by one system to inform another system that it was using the DDI DTD instead of the Dublin Core DTD and that information contained in a specific Dublin Core element should be dumped into the search parameter for the following DDI DTD element.

In addition, the parent, sibling and child nodes of the identified element could be obtained and transferred along with the contents of the element based on the hierarchical information available through the DTD. This would allow for the transfer of variable information with

**Figure 1**

DC ELEMENT	DDI Codebook Element
<b>Title</b>	1.1.1.1 titl (Title of Documentation)
<b>Creator</b>	1.1.2.1 AuthEnty (Authoring Entity)
<b>Subject</b>	2.2.1.1 keyword (Keywords) 2.2.1.2 topcClas (Topic Classification)
<b>Description</b>	2.2.2 abstract (Abstract)
<b>Publisher</b>	1.1.3.1 producer (Producer) [NOTE: The Dublin Core specifies that the publisher should be "the entity responsible for making the resource available *in its present form*" (emphasis added). For a DDI codebook the publisher should be the entity responsible for making the *electronic* version available.]
<b>Contributor</b>	1.1.3.2 othId (Other Ident. & Acknowl.)
<b>Date</b>	1.1.3.3 prodDate (Date of Production) [NOTE: Theoretically, the DC Date element should refer to the date the electronic resource (e.g., the DDI version of the codebook) was created, not any preceding paper version.]
<b>Type</b>	DOES NOT MAP TO ANY DDI CODEBOOK ELEMENT Suggested DC Type: "Text.x-Codebook"
<b>Format</b>	DOES NOT MAP TO ANY DDI CODEBOOK ELEMENT Suggested DC Format: "text/xml" [NOTE: use of MIME type text/xml based on Internet Draft by E.J. Whitehead, Jr. of U.C. Irvine, and M. Murata, of Fuji Xerox Info. Systems.]
<b>Identifier</b>	Suggested DC Identifier: URN for DDI Codebook, if applicable. Alternatively, use the IDNo element within the Document Description citation element.
<b>Source</b>	[NOTE: If a DDI electronic codebook has been produced as the *original* documentation for the data from a study, the DC source element does not apply. If the DDI electronic codebook has been derived from a pre-existing version, then the DC Source refers to bibliographic information regarding this previous paper version. In this case, Source would map to the MARCURI on the docSrc element, or alternatively, to the IDNo element within the docSrc element. [NOTE: Use of the DC Source element is deprecated. The DC Relation element is now preferred.]
<b>Language Relation</b>	xml:lang attribute for codeBook element partially maps to 1.4 docSrc (Documentation Source). No mapping currently exists for the relation type component.
<b>Coverage</b>	2.2.3.1 timePrd (Time Period Covered) 2.2.3.2 collDate (Date of Collection) 2.2.3.3 nation (Country) 2.2.3.4 geogCover (Geographic Coverage) 2.2.3.7 universe (Universe)
<b>Rights</b>	1.1.3.2 copyright (Copyright)

*Dublin Core to DDI DTD mapping suggestions created 7/1/98 by Jerome McDonough, U.C. Berkeley Library Systems Office.*

datafile, question and location information attached as a structured unit of information.

```
<dataDscr ID=da8425>
<var ID='V25' name='empl'>
  <location StartPos='45' EndPos='45' width='1'>
  <labl>Employment Status</labl>
  <qstn ID=Q20>What is the current employment
  status of this person?</qstn>
```

DATA FILE: da8425

Variable: Start End Width

V25 empl 45 45 1

Employment Status

What is the current employment status of this person?

All of these features provide opportunities to develop a range of tools without creating a specialized or unique infrastructure of information. The potential benefits to the researcher are enormous. Possible system features could include:

- Multiple search systems addressing different levels of searches
- The ability to pass information from one level of search to another
- Multiple templates for displaying search results
- The ability to move from the metadata to data manipulation and/or display
- The ability to follow independent tangents of searching through internal links to related materials

The similarity between the DDI DTD and the DTDs of other format types allows for a certain level of cross searching between heterogeneous document types. Specific search engines could be developed which exploited this upper level metadata, making it possible for the researcher to cast a wide net for related information. This could be an upper level of a tiered search approach.

The development of special search engines within specific collections or object types would allow libraries and archives to exploit the unique features of their holdings. Special relationships between collection pieces and unique terminology could be featured. Tools such as a dynamic thesaurus or customized dictionaries could be incorporated. Because both the generalized tool and the specialized tool are addressing the same underlying collection multiple search engines which exploit particular research approaches

could become common. We would no longer be limited to trying to create one tool that works for everyone.

An excellent example of this potential is found in the following three search systems: NESSTAR, ILSES, and GESINE. All address, or eventually intend to address, data collections held at the Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln (ZA) in Köln, Germany. NESSTAR<sup>1</sup> is a search engine, combined with a data manipulation (basic statistics) and extraction tool. It accesses data held at various archives whose metadata has been tagged to the DDI standard. Using structured metadata, NESSTAR allows the user to identify appropriate data sets by querying the variable descriptions, questions, and study description material. It provides options for running real-time calculations on selected variables to further determine applicability and then allows for data extraction according to the access rules of the governing archive. NESSTAR makes searching across archives, exploring data sets and obtaining data on-line a one-stop operation.

ILSES<sup>2</sup> addresses the collection of data and related materials at ZA. ILSES currently does not address the collection through the DDI compliant metadata. There are plans to use this approach in the future. DDI compliant metadata will be accessed directly by the search system or it will serve as a transport format for entering new materials into the system and exporting information from the system to the end-user. ILSES provides access to related literature as well as the data sets and metadata files. The user can approach the collection from either direction. The user has the option of downloading complete data files or customized extracts within the limits of the archives access restrictions. The focus of this system is narrower than NESSTAR in that it addresses only a single collection of data. However, providing the context of related materials and publications provides a better conceptual appreciation of the unique features of ZA's complete collection of materials.

GESINE<sup>3</sup> is not a data extraction engine nor does it currently address data collections. GESINE provides access to the collection of social science information found at IZ which, like ZA, is part of GESIS (Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen). It currently addresses descriptive information housed in an ORACLE database and performs full-text searches on the documents in their database. There are plans to include ZA study descriptions in this database. This would position them to include options for fully searching DDI compliant metadata files. The value of including full-search capabilities for ZA metadata files would be great. Currently the two other systems, NESSTAR and ILSES, take a data user's approach to the data discovery process. Linkages to related literature within ILSES move from the data collection out to works that are based on the analysis of the

data or related to its collection. Expanding the search capabilities of GESINE to include DDI compliant metadata files would allow the user to search for data within the broader context of social science research. Two of the specialized tools within GESINE, the person/institute search and the graphical search and display system, would provide major enhancements unavailable elsewhere. Access to ZA data collection through GESINE would bring these data sets to the attention of a wider audience, those not aware of the separate systems available data searching. If all of these systems were capable of accessing the DDI format, the user may be able to switch systems, without reentering the search parameters in the new system. For example, the ability to switch systems would allow GESINE users to extract data found initially through GESINE through the ILSES or NESSTAR systems. ILSES or NESSTAR users would also be able to expand their search to a broader range of related materials through GESINE.

The use of a standard underlying structure provides the option for integrating multiple search approaches. A researcher would begin his or her search with a general search engine. Later, as a subset of material or a specific collection was identified, the researcher could switch to a more specific search engine that exploited the features of a certain type of material, area of study, or research approach. All researchers should have the option of choosing the search engine that he or she prefers and that most closely matches their own approach to inquiry.

The ability of GRS2 to pass search parameters between systems means the researcher would not have to be limited to using a single search engine during their inquiry. They should be able to move search parameters between systems which can map from one structure to the other.

Reaping the full benefits of the DDI DTD requires adherence to a set of both design and application principles. First, a level consistency in the development of DTDs across heterogeneous document types must be maintained. This is particularly important for the upper level metadata that would be searched in broad cross collection systems. Second, there needs to be some level of structured language developed and maintained within similar document types or disciplines to identify implied information. Third, there needs to be consistent application of the DTD and tagging nodes within the data community. Finally, we must create the tagged codebooks in the DDI DTD format. Without them, there is nothing to warrant the development of specialized search engines and the ability to address these documents in generalized search systems. This means that producers in the data community need to commit to the DTD and produce documentation in this format. This does not preclude production in other formats, but commits the producer to providing a DDI DTD tagged codebook as one of its format options. Data librarians and archivists must

also find a means of translating their existing collections of legacy documents into the new format. Given the variety of documentation in terms of format, layout and quality, this is a massive undertaking. It should be viewed as a means of preserving not only the codebook information, but of preserving and in many cases creating access to the data.

<sup>1</sup> NESSTAR (Networked Social Science Tools and Resources) Developed by the Norwegian Social Science Data Service, the Data Archive at the University of Essex, and Danish Data Archives <http://www.nesstar.org>

<sup>2</sup> ILSES (Integrated Library and Survey-data Extraction Service) A product of the ZentralArchive and NIWI.

<sup>3</sup> GESINE (Integriertes sozialwissenschaftliches Informationssystem) A product of the Informationszentrum Sozialwissenschaften (IZ), Bonn, Germany, <http://www.bonn.iz-soz.de>

\* Wendy Treadwell, Coordinator, Machine Readable Data Center, University of Minnesota, 2 Wilson Library 309 19<sup>th</sup> Avenue South, Minneapolis, MN 55455. 612-624-4389, [wendy@mrhc.lib.umn.edu](mailto:wendy@mrhc.lib.umn.edu)





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 reduced 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:**

Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:

<http://datalib.library.ualberta.ca/iassist/mbrship2.html>

- \$40 (US) Regular Member
- \$20 Student Member
- \$70 Subscription (payment must be made in US\$)
- List me in the membership directory
- Add me to the IASSIST listserv

**Please make checks payable, in US funds, to IASSIST and Mail to:**

**IASSIST,  
Assistant Treasurer  
JoAnn Dionne  
50360 Warren Road  
Canton, MI 48187  
USA**

**Name:** \_\_\_\_\_

**Job Title:** \_\_\_\_\_

**Organization:** \_\_\_\_\_

**Address:** \_\_\_\_\_

\_\_\_\_\_

**City:** \_\_\_\_\_ **State/Province:** \_\_\_\_\_

**Postal Code:** \_\_\_\_\_ **Country:** \_\_\_\_\_

**Phone:** \_\_\_\_\_ **FAX:** \_\_\_\_\_

**E-mail:** \_\_\_\_\_ **URL:** \_\_\_\_\_

**Return Undelivered Mail To:**

**IASSIST QUARTERLY**

c/o Wendy Treadwell

1758 Pascal St. North

Falcon Heights, MN 55113

USA