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, 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) 291-4349 E-Mail: USERDLIB@SFU.BITNET

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.UC.EDU

Features

4 The Electronic Triumvirate: The Archives; The Data Processors; The New York State Department of Correctional Services Inmate Files, A Case Study
by Hugh W. Shinn

8 Acquisition and use of electronic records in the National Archives of Sweden
by Magnus Geber

11 Documenting Data for Secondary Analysis: the Primary Producer's Role and Responsibility
by Bridget Winstanley

News
The Electronic Triumvirate: The Archives; The Data Processors; the New York State Department of Correctional Services Inmate Files. A Case Study

by Hugh W. Shinn
New York State Archives and Records Administration

Abstract:
The New York State Archives and Records Administration (SARA) was one of the first state archives in the United States to accession electronic records into its holdings and make them available to the public. SARA worked with the New York State Education Department's (SED) electronic data processing division (EDP) to obtain main frame computing services. SARA's dependency on SED EDP for data processing services required the development of a positive relationship with SED EDP.

This case study examines the relationship between a government archival institution working with electronic records and a centralized data processing unit that is completely unfamiliar with the operations and requirements of a data archive. There are two levels of a successful relationship: formal, which includes agreements on hardware, disk space, training, etc.; and informal, including the development of creative solutions to technical or procedural problems. These levels of interaction were necessary because SARA (unlike other SED divisions) developed and executed its own applications rather than use the traditional EDP services.

SARA
The New York State Archives and Records Administration's (SARA) program for the archival preservation of and research services for electronic records had its origins in 1988 with the release of the Special Media Records Project report: Strategic Plan for Managing and Preserving Electronic Records in New York State Government. This plan gave birth to SARA's Center for Electronic Records (CER) in 1990, which is to be the focal point of electronic records program development in SARA. Among its many charges, CER was the unit assigned to bring electronic records transferred from agencies to SARA under archival control and to provide reference services for those data files. In this instance, archival control refers to variable-level descriptions of the data set, explanations of the data set's technical aspects, the arrangement and description of the data set as a records series, and verification of the data and the documentation. In archival terminology, the process of bringing electronic (or paper) records under intellectual and physical control is termed "accessioning".

Archival Administration
The accessioning procedure for electronic (and paper records) is preceded by the records retention and disposition scheduling process in which agencies establish minimum retention requirements for records and determine their final disposition: either destruction or transfer to SARA. Most records are destroyed after they are no longer useful to the agency because they do not have enduring legal, administrative, evidential, or research value.

Before electronic records are accessioned by SARA, they are appraised for archival value. Appraisal involves the examination of electronic records from both a content and a technical point of view. Once the records have been determined to be of archival value, the data set's technical aspects are examined to determine whether SARA has the skill and equipment to preserve the data. Technical appraisal is based on discussions with agency personnel and careful examination of the technical and data documentation of the data set.

Data Processing
Before electronic records accessioning operations could begin in 1990, SARA had to obtain computer and technical services from the State Department of Education (SED) Division of Electronic Data Processing (EDP). SARA initially developed an informal plan that specified CER's requirements and contributions to the joint venture of accessioning electronic records.

Requirements of the Archives
CER operates differently from other units in SED with respect to data processing requirements. Most units work with a specified type of data and a specific set of data files. Conversely, SARA collects data from all Executive Branch agencies and must contend with a wide array of data files, types, and formats. The result is that diagnostic programs must be individually designed for specific data files.

Generally, SED EDP customers have informational products such as reports and publications that must be produced. Unfortunately, the user may lack the skills,
time, or equipment to accomplish the unit's tasks. Typically, at the customer's request, EDP will carry out an examination of the unit's particular information requirements, conduct a needs assessment, and where appropriate, assist in the development of goals for the automated system. After developing these plans, SED EDP designs, produces, and tests the programs that accomplish the project's stated objectives. The data processing shop is also responsible for system upgrades and major changes. The customer determines when the data system will run, EDP determines how it will operate. This type of arrangement is practical for systems that are totally dependent upon EDP for development and support.

SARA contends with data from different systems that contain few (if any) common denominators, designed by personnel with varying skill levels. Each data set requires unique levels of handling and effort to document the data fully. In order to provide this level of service, the SARA analyst must become a hybrid of programmer and end user, successfully combining attributes from both worlds. It is this requirement for multi-faceted operation that sets SARA apart from traditional SED EDP customers.

One aspect of commonality among the data sets that SARA has already accessioned is that they were produced by automated systems that are now largely defunct. Electronic record systems are suspended for many reasons including: loss of funding, migration to new equipment, and the completion of a temporary commission's task. When this is the case, the source code is generally missing, and the original programmers have long since departed. Under these conditions, SARA must become an investigative body to determine the scope and function of aging or superseded electronic records systems. These investigations seldom conform to predictable schedules such as annual reviews or a decennial census. Rather, they may appear at inopportune moments in an astonishing array of formats. SARA is currently working with agencies to replace this practice of sporadic submission of outdated material with the more predictable method of scheduled transfers of current data.

CER's role as an investigator and the necessity for multi-faceted operations, produced a number of data processing requirements. The services required by CER for screening purposes are as follows:

1. Access to magnetic tape drives on a non-scheduled basis.
2. Programs to check the physical condition of magnetic tapes for problems such as parity errors.
3. Non-standard training for SED's Unisys mainframe software tools. (The goal is to develop a degree of independence from the EDP staff.)
4. Training in the use of software packages (such as SPSS and INSYTE) that can be used to analyze agency data sets.
5. On-line access to data files.
6. Ability to manipulate data files from the programmer's point of view. This includes variable-level operations involving the verification of data values.
7. The ability to carry out data verification procedures by comparing data values with existing documentation. (Generally carried out with frequency distributions or displays of value ranges.)
8. The ability to isolate and examine specific columns and records within a data set from the operating system level.
9. Production of copies and extracts of data sets, and a method for distributing the data to customers.

Armed with these requirements, CER met with representatives of SED EDP in search of agreements and assistance. For the most part, the data processing staff were helpful and available for consultation.

SED EDP's Operating Procedures
SED is an enormous and complex organization. It has an excess of 3,000 employees, many of whom use or maintain the Department's fifty diverse mainframe applications or its state-wide data network. The size and complexity of the Department required that SED EDP develop standard operating procedures for most types of data and technical functions. These procedures are as follows:

1. All requests for services will be directed to an EDP staff member in the unit assigned to the customer's organization.
2. Customers will not have direct contact with the technical support staff.
3. Customers that have programs requiring the use of magnetic tape drives will request that their EDP contact run the program and submit a 'run sheet' to the EDP operations staff.
4. Formal training above the level of technical manuals and vendor programs is not available for
software packages or higher-level languages.

Most of these procedures reflect EDP’s customer service orientation. In fact, even the lack of formal training relieves the customer of the tedium of writing code and deciphering arcane documentation. SED EDP’s view is that the customer should simply have to make a request, and the data processing professionals will provide the user with the required product. It is an efficient method for managing requests from a large number of customers who require routine services.

Combined Procedures:
The first attempt to combine SARA’s data processing requirements with EDP’s existing procedures occurred when CER began to examine records for the Department of Correctional Services (DOCS).

The DOCS Inmate Under Custody data files were the testing vehicle for the electronic records accessioning program at SARA. DOCS began collecting data on punch cards in 1956. The data were used to produce reports on the numbers and characteristics of inmates under custody from 1956-1974. The information contains incarceration data including facility name and date received; crime and sentencing data; detailed demographic data including race, religion, nativity, occupation, and education; criminal history data. Each annual file is a reflection of the prison population at a given point in time.

These files are unique in that they contain relatively complete inmate-level demographic data on the population of New York State’s correctional facilities 2.

The initial foray into the unknown and the semi-known began when the first data files from the DOCS were examined. The technical information did not coincide with the data sets, nor did the data documentation match the actual data values. It became apparent that additional tools were necessary if the proper file formats were to be discovered.

The Mystery File
The mystery file appeared with the second shipment of data files from the Department of Correctional Services. This data file refused to coincide with any of the printed documentation. It was difficult to determine the structure, size, record lengths and other important characteristics of this data set.

The tools and skills that could be used to augment SARA’s level of training and experience and bring the mystery file under archival control were distributed throughout SED EDP’s sections. While individual sections often cooperated with each other, staff in one section had only a limited need to understand the activities in the other sections. In this instance, the single contact system introduced an additional level of bureaucracy into the arena. The contact had to transmit SARA’s requests accurately to the appropriate EDP section, and accurately transmit that section’s response. Unfortunately, the mystery file stretched the single contact system beyond its limitations, and the file had to be abandoned.

The mystery file situation demonstrated that CER’s data processing requirements were not routine and new procedures had to be developed by EDP to assist CER in its attempt to accession electronic records. SARA had the ability to perform much of the work normally provided by SED EDP. With this in mind, SARA requested access to the tools, software, and training that would lead to self-sufficiency.

EDP’s Response to SARA’s Requests
SED EDP’s response to SARA’s requests for services was to set up a liaison system where CER would contact a programmer in the applications section assigned to SARA. The liaison would respond to SARA’s requests for services, or attempt to find someone who could. Negotiations between SARA and SED EDP resulted in the modification of EDP’s procedures with respect to non-routine operations.

1. On-line access to data files from a programmer’s point of view was always available.

2. SED EDP responded rapidly to requests for service.

3. SED EDP willingly participated in the development of tools and procedures to solve unforeseen problems.

4. SED EDP allowed limited access to the technical support staff.

These modifications allowed CER to make use of the old, inadequate documentation, examine the data, and determine the true nature of the DOCS files. In addition, SARA was able produce more useful documentation for the DOCS Inmate Under Custody files.

Attempts at the formal level to resolve the issues and problems raised by the mystery file situation were not entirely successful; and it became apparent that an additional and different type of relationship with SED EDP was required. Fortunately, strong working relationships were developing between CER and several SED EDP sections. This led to the development of informal relationships with specific programmers from selected sections of the data processing shop.

Winter 1992
The basis of the informal relationship was the realization that personal cooperation between specific SED EDP programmers and the SARA analyst would be a more effective than an organizational approach for solving unique problems. One of the mystery file issues was resolved when the programmers and the analyst devised a method for examining specific columns and records in a data set. Admittedly, this was one CER’s formal data processing requirements. It became an informal item when it could only be operationalized by cooperation between the analyst and the SED EDP programmers.

The informal relationship was designed to augment the liaison system by developing contacts in other sections of EDP. The liaison served as a facilitator, providing the SARA analyst with the initial introduction to the appropriate sections. From that point, the analyst and the programmer in that section would work together as needed to resolve the problem at hand. The strength of the informal relationship is its flexibility.

The combination of the informal and formal relationships formats results in the creation of a Relationship Network. This network is based on the premise that formal relationships are effective for repetitive operations; while informal relationships are more suited to solving unique problems.

There are formal connections between SARA and the EDP liaison and between the EDP liaison and EDP’s individual sections. Through these connections flow procedural recommendations, written procedures, and formal requests for assistance. The formal relationship is particularly useful for contending with repetitive activities such as transferring data files. In addition, it provides a structure for communication between CER and SED EDP.

Informal relationships also exist between SARA and the EDP liaison and among SED EDP’s various sections.

Additionally, informal relationships exist between SARA and the individual SED EDP sections. The informal relationship is useful for the resolution of unique problems such as accessioning new types of electronic records. In short, the informal relationship is more effective in resolving cutting-edge problems rather than completing routine tasks. Informal relationships are legitimized by the existence of formal relationships, and cannot function effectively in the long-term without them.

Conclusion
The flexible aspects of accessioning electronic records require that the archivists and the programmers have the ability to adjust to the variable demands of differing electronic records formats. Flexible operations spring from adjustable, non-stagnant relationships.

The development of a Relationship Network between SARA and SED EDP provided the necessary structure for a flexible electronic records accessioning program. The Relationship Network combines the aspects of the formal and the informal relationship formats. As a result, the electronic records program at SARA is flexible, efficient, and has the capability to address technical and logistical problems (e.g. reading data dictionaries and transferring data files).

The development of a strong Relationship Network has provided a stable foundation for cooperation between SARA and SED EDP.

**** a diagram follows (see p.12 of text) ****


2. New York State Archives and Records Administration, Bureau of Records Analysis and Disposition. Appraisal report # 87-28N; May, 1987; p. 3.
Acquisition and use of electronic records in the National Archives of Sweden

by Magnus Geber,

National Archives of Sweden, Stockholm, Sweden

This paper generally reflects the experiences made at the National Archives of Sweden concerning the management of electronic records and describes how we try to solve the new archival problems that have been caused by these media. I will stress what seems to be special for the Swedish development and possible reasons for that.

Before describing the situation in Sweden I will present some figures about the general situation concerning electronic records at national archives in the world. Since the archival congress in September 1992 I have been trying to gather information about this matter, mainly by distributing a questionnaire. The table below probably contains most of the countries where national archives have received electronic records. The questions might have been interpreted in different ways so the numbers specified shall not be looked upon as exact.

As seen above the National Archives of Sweden has of today received more than 10 000 magnetic tapes containing electronic records. That would be about 1 Terabyte of information totally (1600 and 6250 bpi). This means that although Sweden is a small country the amount of electronic records delivered to the National Archives is amongst the highest in the world.

What might then be the explanation of the amount of delivered electronic records in Sweden? Two main reasons are to be seen, the relatively early and high degree of computerisation among Swedish governmental agencies and the existence of the Swedish Data Protection Act. This is also related to an old and maybe quite unique Swedish tradition of keeping a lot of information about the citizens as governmental personal records.

Swedish governmental agencies expanded especially during the 60s which among other things led to an early introduction of computer systems. Many agencies introduced large administrative systems. With large amounts of the governmental information made machine readable the National Archives became involved in making disposal decisions concerning some of these systems already during the 70s. This was natural as this information in its traditional form on paper constituted basic series of archival records, frequently used by researchers.

The other reason was the Data Protection Act. It was promulgated in 1973 to secure the personal integrity in computer systems, governmental or private. The Act specifies that all computer systems containing personal

<table>
<thead>
<tr>
<th>Number of received tapes/cassettes</th>
<th>files/datasets</th>
<th>Use (annual frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>9 000</td>
<td>12 000</td>
</tr>
<tr>
<td>USA</td>
<td>2 520</td>
<td>11 646</td>
</tr>
<tr>
<td>Sweden</td>
<td>10 100</td>
<td>985</td>
</tr>
<tr>
<td>Danmark</td>
<td>1 200</td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>289</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>1 092</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>2 500</td>
<td>3 500</td>
</tr>
<tr>
<td>Switzerland</td>
<td>304</td>
<td>527</td>
</tr>
<tr>
<td>Italy</td>
<td>(is said to have)</td>
<td></td>
</tr>
<tr>
<td>U K</td>
<td>(plans to have 1995)</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>(is having a project on the matter)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Electronic records at national archives
information of a certain degree of sensitivity, must be permitted by the Data Inspection Board, and have a limited time to exist. This fact originally prevented the National Archives from receiving these electronic records, since the Data Protection Act didn’t support giving permits for archival reasons. But with the change of the Act 1982, the Data Inspection Board could decide that the electronic records from these systems were to be delivered to the National Archives instead of being destroyed. No permission was needed for the National Archives. Disposal decision were also to be made after consultations with the National Archives. Still it meant that the disposal concerning a big part of the governmental information wasn’t decided by the National Archives. However, normally the view of the National Archives is accepted by the Data Inspection Board. Altogether this has led to a high amount of transfers to the National Archives. When a system is ended or parts of the information is getting to old to be stored in the system a decision has to be made. The information must be destroyed or transferred to the National Archives.

Transfers of electronic records had actually started already during the 70s. At that time the records mainly came from different governmental committees which had finished their activities, having all their archival records transferred. After the above mentioned change of the Data Protection Act the transfers increased. The records mainly came from large governmental administrative systems like systems for unemployment, insurance, governmental accounting or applying for university studies. The largest part of the information has come, and still comes, from the systems for taxation where tapes are delivered annually from 25 regional agencies. There is also a certain amount of transfers from research projects, the universities being a part of the governmental sector in Sweden. A transfer to the National Archives offers a possibility for the researcher to have the electronic records preserved with the personal identification even when a study is finished. If there is a need for a follow-up in the future the records may be requested from the National Archives on the condition that a new permit from the Data Inspection Board is given.

Another main part of the transfers, the foremost being from the taxation systems, is the records from Statistics Sweden. Statistics Sweden has of course a large amount of computer systems with personal information. This is regulated by the Data Protection Act. After long discussions between Statistics Sweden and the Data Inspection Board also involving the National Archives, the Swedish Government decided that a large part of these records that had reached a certain age were to be transferred to the National Archives. This was done in 1989. At the same time the National Archives received extra economical resources for this matter. The transfer was partly made voluntary by Statistics Sweden to avoid the costs caused by clause 10 in the Data Protection Act. (That clause enables every person having their personal data registered in a computer system to have an outprint of all that information once a year. But the electronic records transferred to the National Archives are excluded from this plight.) The transferred records are still frequently used by Statistic Sweden as I will mention later.

For a long time there was no special staff for the electronic records at the National Archives. But following increased funding a section dealing with modern media was created in the beginning of the 80s. Today there is an EDP-section as a part of the division for technical matters. This division also includes micrography, conservation and book binding. The EDP-section deals with the internal EDP use and with the development of archival applications like computerised inventories apart from managing transferred electronic records.

For many years there was no computer equipment at the National Archives. All use and copying had to be done through service bureaus. Some years ago we bought a UNIX-computer. But we have had problems finding suitable software for our needs. Today we are examining the possibility to use a DOS-system with special software and different tape-drives, to be able to convert both different physical and logical formats. (The influence of finding such a solution came from visits to the National Archives in USA and Canada last autumn.) Our tapes are stored in a climate archive. They are rotated, rewinded and then copied after 10 years. All tapes are transferred and stored in 2 copies, one being kept in an archives far north of Stockholm.

When electronic records are transferred the National Archives sets certain requirements. This is possible on the basis of the Swedish Archives Act which gives the National Archives the right to regulate the management of electronic records considered governmental archival records. The requirements are mainly specified in accordance with different standards. The aim is to get the electronic records in a form as independent of original hardware and software as possible. We accept ASCII and EBCDIC but the numerical information may not be in any packed format. The files shall be in fixed format and may only contain one record-type. The idea is to have a structure corresponding to and also directly importable to a relational database. This means for example that variable files from a system with a hierarchical structure have to be converted before transfer. We have old transfers with files not satisfying matching our requirements. These files will be converted when we make a new generation of storage copies. (It could mean that a variable file containing 5 record-types would be converted to 5 fixed files.) The transferred tapes shall
contain files labelled with ISO or IBM labels. Today we only accept 9-channel spool tapes, 1600 or 6250 bpi. But in the near future we plan to accept other physical formats. Probably we will choose two types for internal longtime storage, for example 3480 and DAT. We will possibly also accept other types of data-media for transfer and convert them as long as our software can handle them and we can charge an extra conversion fee.

There are also certain regulations about the form of the documentation of the electronic record. Generally the documentation is kept on paper. But we would like getting code tables machine-readable, like an extra table in a relational database. To make it easier to import the files to a relational database system, we wouldn’t mind getting the record descriptions machine-readable. In the future we probably also would like to get the descriptions of how to create a certain archival record from a number of storage files/tables expressed as standardised SQL-commands. It is important to have in mind that the National Archives is collecting public records in machine readable form, which may come from complex administrative systems. The actual record does not have to be similar to what is one data file.

The National Archives also takes part in different fields of the standardisation work.

Primarily, in the standardisation of archival techniques such as terminology, paper, book-binding and storage conditions. But because of our work with electronic records we also take part in the standardisation work on IT, linked with ISO/IEC JTC 1. This has been a good way of getting information in this field. But because of lack of time and technical knowledge the possibilities to contribute to the work and influence the standardisation work has been less then we would have wished.

Finally I would like to mention a little about the use of the electronic records transferred to the National Archives. The basis for this is the Freedom of Information Act which is one of the fundamental laws (constitutions) of Sweden. It stipulates that all governmental information is open to the public, with the restrictions specified in the Act on Security (Privacy Act). This also includes electronic records. However the use of these in the National Archives has been low until the transfers from Statistics Sweden were made. The reason is probably that the users do not know about, lack the technical knowledge or are not interested in modern records.

Today the main user is Statistics Sweden having need of their former records, and doing it so much that there is a transport from the National Archives twice a week. But the use from others, mostly research institutions, also has increased in recent years. The researchers using the electronic records are mostly in the field of medicine and may also ask for some material originally from Statistics Sweden. All this is done in machine-readable form. It will therefore require a permit from the Data Inspection Board if the records shall contain personal information. 99% of the electronic records transferred to the National Archives contains personal information. However it would be possible to get some special records without a permit if the personal information was excluded. There is an example of SSD having received electronic records from the National Archives in that way.

Up to May 1993 it was very rare that someone wanted information as an out-print. But that month we got the national register of private boats as a transfer. The responsible governmental agency had to end it after a political decision. Just now we receive about 25 questions a week concerning these records. Mostly it has been the police, the navy or the public asking about the owners of lost boats. The information are distributed by mail, fax or telephone.

Today we have no possibilities for the public and the researchers to do on-line work with the electronic records. But we plan to have it in the future. There is actually a cooperation between the EDP-divisions/sections of the National Archives in the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden). That has resulted in a common project called TEAM ( Availability of Electronic Archival Material). Within the project the Norwegians are constructing a relational database system to import data from the storage files and then using some suitable software to present the data. Through this system we hope that in the future the public will be able to get direct access to and print-outs from our electronic records, naturally under the restrictions that are set by the Act on Secrecy and the Data Protection Act.

1 Paper presented at IASSIST/IFDO’93 Conference, Edinburgh, Scotland. Magnus Geber, National Archives of Sweden, P.O. Box 12541, S-10229 Stockholm, Sweden ph +46-8 737 6486 fax +46-8 73736474

2 The figures are taken from a questionnaire answered from September 1992 to May 1993. The numbers describe the total holdings and the annual use and form of distribution of electronic records to researchers and the public.
Documenting Data for Secondary Analysis: the Primary Producer’s Role and Responsibility

by Bridget Winstanley
ESRC Data Archive, University of Essex, U.K.

Background and Acknowledgements
This paper is based on a session and a round table lunch discussion on the same theme which took place at the IASSIST Conference held in Edinburgh in May 1993. The session was convened by Sue Dodd and Bridget Winstanley. Papers by Laura Guy, Joanne Lamb, Marcia Taylor, Paul Child and Kevin Schurer, as well as the numerous participants at the round table discussion have all contributed to the ideas presented here as have the members of a European Committee on Documentation Guidelines, set up by the ESRC Data Archive earlier this year. This committee has representation from the ESRC Data Archive, the Office of Population Censuses and Surveys, Social and Community Planning Research, the British Household Panel Survey and other areas of the British academic research sector and the Steinmetz Archive in the Netherlands.

The Need for Guidelines
We start from the basic premise that the person or persons best placed to document their data are the primary producers of those data. It is axiomatic that their knowledge of the data must be more complete than anyone else’s. Yet many primary producers of data are reluctant to create documentation of a standard which goes beyond their immediate needs for their own analysis of the data. The reasons for this reluctance, when it occurs, are obvious. The creation of documentation of a substantively and physically high standard is time-consuming and expensive. There are apparently few incentives to producing such documentation. The culture of data sharing is still largely in a state of infancy even after at least a quarter century of data archiving. And additionally, it is not always apparent to the primary producer what the secondary analyst requires in the way of documentation.

The primary producers who fail to document their data to an acceptable standard must be balanced by some shining examples of good practice in this field. Some of the most recent of these include The British Household Panel Survey’s two volume user manual (1) and the U.K. Employment Department’s user guide to the Quarterly Labour Force Survey (2), both of which are available as machine-readable text files at a much lower cost to the user, as well as appearing in printed paper form. North America can show many examples of data which are well documented by the producer for public use, including the General Social Surveys produced at the National Opinion Research Center (3). A recent publication by the Steinmetz Archive in the Netherlands demonstrates a dataset put together from a time series of NIPO polls (4) with thoroughness and consideration for secondary users of the data.

Despite these fine achievements, and many others, by individual research projects, there is much more that data archivists and librarians can do to promote good documentation by primary producers of data. The arguments for doing so encompass both the promotion of good practice and necessity arising from financial and economic constraints facing disseminators. We have already stated what we take to be self-evident, that primary producers are capable of producing the best documentation because of the familiarity with the data. The further imperatives for persuading primary producers that they have a role and a responsibility towards the documentation of their own data lies in the decreasing resources and increasing material coming into data libraries and archives. Many can no longer afford to create documentation for all (indeed any) of the datasets which they distribute and in any case the upgrading of poor documentation after the original project is over is frequently painful and unsuccessful: memories have dimmed and in many cases the original investigators have dispersed. Yet datasets which are inadequately documented are of no use at all to the secondary users to whom the data are being distributed.

A further important incentive to the production of good documentation was described by W.J. Bradley at the IASSIST/IFDO 93 Conference (5). The sponsors of major data collection exercises, typically government departments and other policy-making organisations, expect more for their money than data. They expect information. According to Bradley, policy advisors are often quite desperate for timely, relevant information. Given their wide-ranging and often unpredictable requirements, advisors and decision makers are a prime target audience for easy, responsive secondary data analysis services that integrate and draw upon the broadest possible base resources. Bradley and his colleagues have created software which demonstrates how good documentation, when standardised and
structured, can integrate and front-end rapid and easy access to the data resources that have been documented in this way (6). They also describe how such documentation can actually serve to facilitate the creation of information and knowledge products which in turn can be integrated for re-use in information retrieval. The development of documentation guidelines, together with associated methods of standardisation, are keys to the knowledge delivery process.

Strategies for Improved User Documentation
There are several lanes in the highway which leads towards the ultimate goal of improved documentation by producers of data. We need to convince data funders of the economic arguments in favour of improvements in the standard of documentation. We need to convince data producers of the value of good documentation to the organisation of their own research, as well as of the recognition of their work which will come from their work being re-used and acknowledged. We need to convince secondary users to afford this recognition to primary producers. Finally, we need to provide support to primary producers by developing and distributing guidelines on the production of documentation.

The case to be made to the funders of data is, as indicated in the previous paragraph, primarily an economic one. Many funding bodies are indeed aware of the wastefulness of funding projects with major data-collection components without ensuring that the data are made available for further research beyond its primary research aims. In many cases they are aware, too, that a major constraint on the re-use of data is the lack of adequate documentation. There is sometimes a perception, however, that the disseminating agency, usually a data archive or data library, will document the data, so the producer does not need to move beyond minimal standards. We must make the case that producers are better placed than archivists to create documentation of a high standard for their own data and that it is more cost effective for them to do so. A certain amount of data processing and standardisation will always be necessary in the archive or data library, but the better the incoming documentation, the better the outgoing data and documentation. Funders are in a powerful position to provide incentives in the form of additional funds for documentation procedures within the original project funding as well as penalties in the form of blacklisting for those who do not document their data adequately. The judgement as to whether the data are adequately documented for secondary research will probably be the archive’s and for this reason we need minimum standards in the form of guidelines.

Data archives and librarians will rely largely on funding bodies to provide the penalties for inaccurate documentation. But they have a major role to play in persuading their depositors or donors of the incentives for providing high quality documentation. Above all, the case has to be made for making their data widely usable. Why should they care? Because usage can be reported back to funding bodies as an argument for more funding; because when data are well-documented there is no need for the constant answering of queries from secondary users; and because usage will bring citation and recognition. Here we, the data librarians and archivists, have a task ahead to ensure that use of data which leads to publication also leads to the citation of the dataset. The rules of citation for datasets are well established (see Dodd (7)) but we can do more to ensure that they are observed. A scan of examples reveals also that there needs to be clarification on whether the documentation or the data, or both, are being cited. Of the examples given above, only the General Social Survey’s documentation (3) gives guidance on both the citation of data with documentation and the documentation alone, although the ESRC Data Archive’s citation guidance does make it clear that the citation shown is for data with documentation. The other two cases assume citation for documentation only. Guidance on citation should be included in all documentation, editors of journals should be approached to try to ensure their co-operation, and a constant stream of reminders published in newsletters and bulletins. Citation has its own rewards in the form of easier identification of data sources for those reading the citation, but also, of course, it ensures the recognition of the achievement of the producer of that dataset in making it publicly available. But citation can only take place when the dataset has a bibliographic identity conferred upon it by its documentation. Guidelines are required to show producers how to document their data in a way which will ensure this.

Existing and Future Guidelines
Guidelines already exist for creating the necessary elements for documentation. Two US examples are Carolyn Geda’s Data preparation manual (8) and Richard Roistacher et al A style manual for machine-readable data files and their documentation (9). Other examples are the U.S. Bureau of Justice Statistics’ Technical standards for machine-readable Data (10) and Patrick Collins and Jane L. Powers The preparation of data sets for analysis and dissemination : technical standards for machine-readable data (11). Excellent as they are, the earlier of these manuals are out of date and need revision while the latest (Collins and Powers) although providing an attractive introduction to the subject, focuses on the practices required by a particular archive (The National Data Archive on Child Abuse and Neglect at Cornell University) and is consequently short on general detail.

A new comprehensive set of guidelines, covering both
optimal and minimal standards, taking into account new
media, new formats, new data collection techniques and a
new archival environment, is urgently required. These
should include a recognition of the fact that many social
scientists are using and creating textual data, or mixed
numeric and textual data in their research. It is important
that new guidelines should recognise too, the consider-
able work already undertaken in the humanities and not to
duplicate that work. The work of the Text Encoding
Initiative should be brought to the attention of social
scientists in a way which will be appropriate to their
needs. Although the guidelines should deal with sub-
stance and content, format should not be forgotten. For
many primary producers and the archives or data libraries
which will be disseminating their data and documenta-
tion, the most convenient format in which to produce
documentation will be machine-readable. In addition to
providing a cheap and convenient means of disseminating
documentation on the same medium as the data, machine-
readable documentation opens the way to better informa-
tion systems, allowing the prospective user to examine
and compare documentation online before deciding on the
appropriateness of a particular dataset for his or her
particular research.

Once we have agreed on both optimal and minimal
standards for documentation we need to think about how
to get them accepted. If they have been developed in
consultation with data producers and if they are attractive
and easy to use, this will be easier. A printed paper
version is indispensable but we must also develop
software applications of the guidelines. Work in this area
has already begun, notably by W.J. Bradley and his
colleagues in the Social Environment group of Health
and Welfare Canada. Their work on DDMS (6), a PC-based
package for managing social science dictionaries and
documentation takes into account the data elements
recommended by Roistacher and provides an easy way to
manage data as well as ensuring that these data will be
well-documented. Such easy-to-use software in the
hands of data producers will be an incentive to the
production of complete documentation. The further work
by Bradley, Hum and Khosla on DAIS (Data and Infor-
mation Sharing) (12) shows how easy, end-user access to
data can be provided by documentation that has been
structured and standardised via DDMS. This system
provides a vital incentive to the funders of data who are
themselves able, via this system, quickly to locate
relevant data items from a broad array of datasets and
generate their own analyses using software of their own
choice.

Other work on codebook software has been carried out by
the Swedish Social Science Data Service and further
work on codebook production is under way as an IAS-
SIST Action Group led by Karsten Boye Rasmussen of
the Danish Data Archives. While recognising the
contribution this will make to the sharing of data through
data archives, this paper, because it is concerned only
with the primary producer’s role and responsibility, does
not aspire to enter into the current debate, conducted
largely through the IASSIST listserver, on the desirabil-
ity of replacing OSIRIS as a codebook tool. It is vital,
however, that before we undertake the publicity and
training required for the acceptance of software products,
we are agreed on the substance of the guidelines for the
documentation of data.

Conclusion
Penalties, incentives and support all depend upon the
existence of guidelines for documentation. Funding
bodies have to be persuaded (as many already are) that
the provision of funds for research projects to collect data
at great expense without making provision for the wider
use of these data is intolerably wasteful. For some, such
as large governmental organisations, good documenta-
tion is essential for sharing within their own organisa-
tions, and all that is required is some guidance on how to
do it in a way which has a broader application outside
their own spheres. Other types of funding organisations,
who have traditionally seen a single report as the end
product of their sponsorship, need to be made aware of
how much further their money will go if many reports and
analyses for different purposes and by different
researchers can result from their investment. Their role
with regard to the documentation of datasets which they
have funded should be to withhold further funding if the
data are not sufficiently documented for further research
(stick) and to provide an element of funding sufficient to
ensure that the data are documented (carrot). Primary
researchers have to be persuaded (as many already are)
that the creation of a dataset which can be used by others
is worthy of recognition, acknowledgement and citation
in the course of scientific research and public policy
planning. Secondary researchers, those making public
policy, and the editors of journals should be persuaded to
provide the recognition, acknowledgement and citation.
The wider use of data and the recognition of the primary
producers is dependent on the quality of the documenta-
tion which accompanies the data. The quality of the
documentation will depend on the guidelines which we,
the data librarians and archivists whose task it is to
facilitate the flow between primary and secondary
researchers, can provide to primary producers.

1 Paper presented at IASSIST/IFDO’93 Conference,
Edinburgh, Scotland.

Household Panel Survey user manual. 2v. Colchester:
University of Essex.


(10) U.S. Bureau of Justice Statistics Technical standards for machine-readable data


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)

### Membership form

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 Pawlicki
Treasurer, IASSIST
% 303 GSLIS Building,
Social Science Data Archives, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484

<table>
<thead>
<tr>
<th>Name / title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional Affiliation</td>
</tr>
<tr>
<td>Mailing Address</td>
</tr>
<tr>
<td>City</td>
</tr>
<tr>
<td>Country / zip / postal code / phone</td>
</tr>
</tbody>
</table>