The Data Documentation Initiative

The Value and Significance of a Worldwide Standard

GRANT BLANK
American University

KARSTEN BOYE RASMUSSEN
University of Southern Denmark

Effective secondary analysis of social science data requires good documentation. Especially because Internet access has become standard, the problems of reading and understanding the contents of data files have become acute. Resolving these problems requires standards for documenting data, as well as standard formats for both data and documentation that can be read and displayed by computers and software anywhere in the world. To define a documentation standard, representatives of North American and European survey research and data archive organizations have created a Data Documentation Initiative (DDI). This article discusses the value and significance of that effort for the social sciences.

Keywords: data documentation initiative; DDI; secondary analysis; XML

Analyzing data can be difficult and time consuming, so we continue to look for resources that make our work easier and faster. One resource is an emerging standard for data documentation called the Data Documentation Initiative, or DDI. The name sounds prosaic and, perhaps, of interest mostly to data archivists; in fact, the DDI is a core element in a web of rapidly developing scientific infrastructure projects. The DDI will have a major impact not only on data access but also on survey design, data creation, and data analysis itself.

Most social scientists use quantitative data in various forms—including social surveys, psychological test measurements, economic and financial series, and government statistics—but few have thought carefully about documentation. Yet, good documentation is of crucial importance. For files to be useable, documentation has to specify the location and meaning of individual variables. Statistical software system files store this information, but that is all that they can store. It is not sufficient (Blank, 1993; Rasmussen, 1989). Information about variables is useless unless the population, sample, and sampling procedures are described. Many studies use complex sampling designs, and even simple data sets often require weights, which must be documented. Data files can become corrupted during storage or transmission, so good documentation includes complete frequencies or descriptive statis-

AUTHORS' NOTE: Many institutions and individuals have contributed to the collective efforts of the Data Documentation Initiative. The authors want to express their thanks and admiration for all these people. For comments on earlier drafts of this article, we would like to thank G. David Garson, Ann Green, Mary Vardigan, and an anonymous referee. Special thanks to Mary Vardigan for answering numerous questions about the DDI committee and the DDI project.

Social Science Computer Review, Vol. 22 No. 3, Fall 2004 307-318
DOI: 10.1177/0894439304263144

tics for every variable. Some data, such as opinion surveys, can be time dependent, and this requires knowing the dates when data were collected. Long-running or complex studies often store data in multiple files, and the relations between files must be documented. These are only some of the requirements for good documentation. High-quality documentation benefits the social sciences in the following four ways: (a) other investigators can understand and use the data; (b) the original researcher can return to the data long after details have faded from memory; (c) the initial investigator is forced to be more systematic and rigorous in understanding the limits of the data; and (d) it provides a basis for systematic cumulative building on prior knowledge (see Sieber, 1991).[1] The fundamental fact is that data sets will be indecipherable and useless unless they are adequately documented (Waters, cited in Green, Dionne, & Dennis, 1999, p. vii).

We begin this article by describing central problems of social science data-set documentation and storage. Second, we briefly relate the history of prior efforts to provide standard, machine-readable documentation and the origins of the DDI. Then, we describe the DDI standard and indicate some areas where it is in current use. We conclude with remarks on the future development of the DDI and the potential impact of improved access to data. Some of the ideas and capabilities that we discuss below have been implemented in some places and others are planned; none of it has been widely implemented. Our task is to describe the goals of the entire system, its current implementations, and its promise for social scientists.

PROBLEMS TO BE SOLVED

The social sciences have accumulated immense resources of data. Decades of data files, including many long-running studies such as the General Social Survey (GSS; outside the United States, this is known as the International Social Survey Program, ISSP), the Eurobarometer series, and the National Election Studies (NES) are available from data archives in North America and Europe. Data files were stored and distributed in electronic form on magnetic tape beginning in the early 1970s. Data in digital form could be easily converted to be downloadable over the Internet, and almost all datasets are now accessible online. However, as we pointed out in the previous section, data are useless without adequate documentation.

The story for documentation is different. For decades, documentation remained entirely on paper. It required a separate system of processing, storage, and retrieval in parallel with data. Paper documentation has a variety of problems: storage is expensive, it requires careful controls to track it in inventory, and it deteriorates over time.

More than a decade after the Internet came into significant use, large quantities of documentation remain only available on paper.[2] Data archives have strong incentives to convert documentation into electronic form. Most obviously, mailing paper documentation for data available via the Internet is slow and costly, for both archives and users. These problems have limited the usefulness of archived data sets.

From the point of view of researchers, another problem is even more important: Searching paper documentation is slow and monotonous. In practice, to use paper, the analyst must know beforehand which studies are relevant to the question of interest. Few social scientists have such detailed knowledge of prior work. In practice, users have been dependent on archive personnel and their ability to find—or remember—the appropriate studies. In short, locating and searching paper documentation is time consuming, tedious, and requires detailed knowledge of prior studies. Without searchable electronic documentation, users cannot easily find populations of interest or studies that have asked questions of interest.

Although users often prefer paper as the medium for reading, but paper has serious limits when it comes to locating appropriate data.

Increasingly, documentation is available in searchable form over the Internet. In practice, the value of searchable documentation has been limited and has never reached its potential. The ability to search across studies for similar items or similar studies has been inadequate because no standard formats existed.[3] The result has been that access to archived data sets for secondary analysis has been limited. The promise of the Internet—to make information easily and widely available—has remained unfulfilled.

A BRIEF HISTORY OF DATA ARCHIVES AND DOCUMENTATION

Archives have been painfully aware of these limitations. Some kinds of electronic documentation have been in use since the mid-1960s when the Inter-university Consortium for Political and Social Research (ICPSR) gained access to an IBM mainframe computer. Early documentation was called a codebook because it documented the codes used in sample surveys. In the 1970s, the ICPSR and other archives developed electronic documentation in the form of the OSIRIS codebook. This was a capable documentation standard. For individual survey data sets, OSIRIS provided documentation at the variable level, such as variable names, variable labels, value labels, and missing values. OSIRIS even allowed limited information about the study itself.[4] Everything about computing in the 1970s was very expensive, so electronic documentation was also expensive. Only major archives and a few large government data producers could see the potential for thorough documentation and could afford the cost. Others decided that supplying electronic documentation was too complex and, in practice, often did not provide any electronic documentation at all.

During the mainframe era of the 1970s, some efforts were made to improve standardization of data archives worldwide. Typically, universities had a connection to an IBM, CDC, or UNIVAC computer, and all three were serviced from national data archives or (in the United States) from distributed data archives and data resource centers.[5] There was limited variation in both hardware platforms and statistical software packages; standards were reasonable because of this relatively simple environment. Furthermore, there was a belief in preservation that supported the development of standards. This was apparent, for example, in recommendations for standard formats to use when exchanging magnetic tape (Rasmussen, 1978). The limits of the OSIRIS documentation were apparent. Archives made efforts to develop a standard that would document the study level; that is, the population, sampling frame, sample, funding agency, contact person, and other information about who conducted the study and how it was done. The National Science Foundation funded a workshop in 1974 specifically to improve "mechanisms for exchange of social data" (Anderson, 1974, p. 153), including the development of machine-readable codebooks.[6] Anderson proposed several further steps, including formation of a Codebook Standard Task Group. There were international workshops as well (Nielsen, 1974), but the standards were only implemented at a few archives, and a study-level standard was not generally accepted.[7]

The 1980s introduced personal computers (PCs), and attitudes shifted. The new emphasis was on quick dissemination of data sets and documentation, and cooperation lagged. With PCs, every data library could develop its own systems, and they did! By the end of the 1980s, much electronic documentation was hidden in nonstandard, individualized, personalized crypts of data. Throughout the period prior to the Internet, dissemination incurred long delays as data and documentation were copied and delivered by surface mail.

By the 1990s, the Internet brought users and archives back into closer contact. The downside of PCs became more obvious: supporting the enormous variety of hardware and software drove up costs. Users wanted access to the original data elements as well as to complex, sometimes fanciful presentation and retrieval systems. It quickly became obvious that these problems could only be overcome by developing standards for data documentation.

In 1993, staff from data archives and members of the International Association for Social Science Information Service and Technology (IASSIST)—the professional association of data archivists—formed The IASSIST Codebook Action Group to work on problems of electronic codebooks. Since 1995, the ICPSR has been leading an international effort to create a worldwide documentation standard to improve access to data.[8] The original committee was appointed by Richard Rockwell, then executive director of the ICPSR.[9] Now called the Data Documentation Initiative, the DDI effort has come to fruition in a standard (see www.icpsr.umich.edu/ddi).

DATA DOCUMENTATION

We refer to data documentation as the rich, full technical documentation of a data set. Recognizing that much more is being documented than codes in the remainder of the article, we replaced the word *codebook* with the more general term *metadata*, meaning data about the data. Data documentation provides information about the sponsorship, design, history, setting, structure, format, and limitations of data sets. Easy, accurate use of data depends on access to comprehensive, accurate documentation. Both teaching and research uses of data require this kind of documentation.

The DDI is a standard for electronic documentation. The advantages of the DDI result from the following two characteristics: it is a standard and it is structured. We discuss both issues in turn.

A comprehensive, worldwide, electronic documentation standard offers significant advantages for two groups: data archives and researchers. An electronic standard eliminates the problems of paper documentation. Electronic documentation is much cheaper to store and inventory. Although it deteriorates more quickly than paper, it can be easily copied or refreshed almost automatically and inexpensively. A comparison of costs between paper and electronic documentation illuminates these differences. Typical study documentation is around 200 pages in paper form. The ICPSR estimates that the cost of sending paper documentation to a user—including duplicating costs, inventory control, storage, handling, and shipping—totals around $25 per data set. Electronic documentation has essentially no duplication or shipping and handling costs. Inventory control and storage is much cheaper, amounting to less than $1 per data set.[10]

Earlier documentation standards often focused on immediate presentation and use; for example, statistical packages typically include only the metadata required for presentation of their output, such as variable names, variable labels, and value labels. This narrow focus is appropriate for the purposes of statistical output, but it is a very limited view of documentation. As the use of information technology matured, the following four points became clear. First, the value of information lies in use. Thus, the primary goal of a documentation standard is that it stores information to make it understandable and accessible. Second, because archives never know what information will prove valuable to current or future users, the documentation standard should be comprehensive and flexible so that it can contain all possible current information and be open to future information needs. Third, presentation and analysis are the tasks of application programs and should not be a direct concern of the documentation standard. To be usable with different software applications, the standard should support a variety of flexible retrieval mechanisms. The structuring of the documentation supports flexible transformation. Fourth, no single software application will utilize all available information. How various software utilizes information is, again, not a concern of the documentation standard.

Because many of these issues flow from the structuring of the documentation, we turn next to the issues of how the DDI is structured. This requires a discussion of the language in which the DDI is written.

STRUCTURED DOCUMENTATION IN XML

Human beings can easily understand the content of a printed page or computer screen. Once a person learns what information is contained in documentation, that knowledge can be easily generalized. A person can find the appropriate information in other documentation even though it is structured differently and the information is found in a different location. Computers are not tolerant of changes in order, location, or structure. Minor differences in the location or wording of information can create great difficulties for computers attempting to interpret content. This problem is well-known; the solution is perhaps less widely understood.

The solution is to separate content from display of information, and then create a set of rules for identifying content by using highly structured data. This solution is embodied in a language called eXtensible Markup Language (XML). Unlike the widely used Hypertext Markup Language (HTML), which defines only how information is displayed via predefined tags for such components as headers and paragraphs, XML focuses on content.[11] It leaves display issues to other software (such as web browsers), and it defines the content in a highly structured, computer-readable form. Here is how it works. Consider the following example, the documentation for a single question from an attitude survey:[12]

```
VAR 0497 WORRIED CONVENTIONAL WAR?
NAME WCONWAR
LOC 909 WIDTH 1

How worried are you about our country getting into a CONVENTIONAL WAR at
this time, one in which nuclear weapons are not used? Are you VERY
WORRIED, SOMEWHAT WORRIED, or NOT WORRIED AT ALL?
----------------------------------------------------------
     274     1.  VERY WORRIED
     298     3.  SOMEWHAT WORRIED
      96     5.  NOT WORRIED
       5     8.  DON'T KNOW
       3     9.  NO ANSWER
    1809     0.  SKIP, I in Q.0403
```

The example question has been formatted for display on a screen or page. The question has a large number of elements: a variable name, variable label, the location of the variable in the data file, the full text of the question, and others. Although the form above is meaningful to a person, to a computer it is just a string of text. For a computer to process this text, we must identify all the different elements. We do this by using *tags*.

In XML, all text—including tags and the document itself—is in simple ASCII text. As in HTML, tags are delimited by using pairs of angle brackets, like this: < >. We could begin by defining a variable name tag, call it VARNAME. To indicate the variable name for this question, we would use a pair of tags, one to indicate the beginning of the variable name text string and another to indicate the end. They would look like this: <VARNAME> WCONWAR</VARNAME>. This looks very similar to HTML, but there are three important differences. First, notice that instead of using tags for display, XML uses tags to describe content. There is no information here about how WCONWAR is to be displayed. We are simply saying that it is a variable name. Second, in HTML, the tags are predefined and we cannot add new tags. In XML, we can define any tags that we need. Using XML, the DDI defines tags for every possible attribute of a variable as well as for other parts of the documentation. Third, browsers have a built-in capability to convert the tags to presentation, so text within <B>Bold text</B> will indeed be bold. We can add display instructions to XML to specify how any tagged information will be displayed. This is done through stylesheets, using XSL, the eXtensible Stylesheet Language. This approach yields additional flexibility; by using different stylesheets, we can display the same tagged text in different ways. The full tagged information on this variable might appear like this:

```
<VARIABLE ID=497>
<VARLABEL>WORRIED CONVENTIONAL WAR?</VARLABEL>
<VARNAME>WCONWAR</VARNAME>
<COLUMNLOC>909</COLUMNLOC>
<LENGTH>1</LENGTH>
<VARTEXT>
<PARA>How worried are you about our country getting into a
CONVENTIONAL WAR at this time, one in which nuclear weapons
are not used? Are you VERY WORRIED, SOMEWHAT WORRIED, or NOT
WORRIED AT ALL?</PARA>
</VARTEXT>
<CATEGORY><VALUE>1</VALUE>
<CATTEXT>VERY WORRIED</CATTEXT>
<CATSTAT TYPE=FREQ>274</CATSTAT>
</CATEGORY>
<CATEGORY><VALUE>3</VALUE>
<CATTEXT>SOMEWHAT WORRIED</CATTEXT>
<CATSTAT TYPE=FREQ>298</CATSTAT>
</CATEGORY>
. . . .
</VARIABLE>
```

The newlines have been added to improve human readability; they are not required by XML. The most important thing to notice about the tagged question is that each element of the variable has its own separate tag. The collection of defined tags and their structure is called a Document Type Definition (DTD). The DTD for the DDI can be viewed at www.icpsr.umich.edu/DDI/users/dtd/index.html.[13] It includes not only information on individual variables but also information on the study as a whole. A brief, illustrative list of the tags in the DTD is below, divided into study-level and variable-level tags:

Illustrative study-level tags:

- Principal investigator or agency.
- Official title of the data set.
- Funding source(s).
- Persons or organizations responsible for data collection.
- Sample and sampling procedures.
- Weighting.
- Response rate.
- Date and geographic location of data collection and the time period covered.
- Unit(s) of analysis.
- Restrictions on use of the data.
- Bibliographic citation.

Illustrative variable-level tags:

- Exact wording of the question.
- Item or question number.
- Associated variable name.
- Location in the data file.
- Missing data codes.
- Imputation and other editing information.
- Details of constructed (computed) variables.
- Exact wording of all possible responses.
- Meaning of each response code.
- Unweighted frequency distributions or summary statistics.

Of course, not every tag will be relevant for every study, but collectively they define the universe of information needed to document any study.

The structure supplied by the tagged file is what gives other software the ability to manipulate the file. This division of labor between content and display is the greatest strength of the DDI. By focusing on the single issue of information storage, the DDI can be more comprehensive and flexible. Because it does not define presentation, the display of information can be tailored to the unique needs of every user. The DDI is open to any kind of presentation; it is flexible enough to accommodate the presentation needs of future applications that we do not now envision. Software can read a DDI-tagged file, extracting the information that it needs and displaying it to users or processing it in some other way. For example, included among the variable-level tags is all the information needed to read the data file into the system file format of major statistical packages (e.g., SAS, SPSS, Systat). A second example: All the advantages of searching across files are possible because a computer can use the tags to identify relevant information. As need arises, additional tags can be defined for new elements and added to the DTD, hence the *extensibility.* The possibilities are unlimited.

XML offers significant advantages for an international documentation standard.[14] Specifically:

- XML is extensible.
- XML is not proprietary and requires no license fees. It was developed by the World Wide Web Consortium (W3C).
- XML is an ISO standard.
- XML files are entirely ASCII text, so they are easy to migrate across different computers and across technological changes.
- XML has widespread database support.
- XML has widespread support from software companies, including support in recent versions of Microsoft Office and all current Web browsers.
- XML presentation is flexible through the use of XSL stylesheets.

THE VALUE OF THE DDI FOR RESEARCHERS AND TEACHERS

The effort to define a new comprehensive standard for data documentation has been led by data archives, but the significance of the project is far wider. For researchers and teachers, immediate benefits include the ability of anyone anywhere in the world to instantly obtain a copy of the data and documentation of any study over the Internet at very low cost. Data and documentation are provided in standard formats readable on any kind of computer with any software.

Ease, convenience, and lower cost are valuable, but the most important advantages of standard documentation are in additional capabilities that it makes possible. As the DDI standard spreads, it fosters important emergent features and network externalities. For example, once multiple studies are documented with a single standard, it becomes relatively simple to allow users to search across studies. Because all documentation consists of structured files in electronic form, searches can be conducted for similar variables or for similar text in questions, or for similar response categories or for similar populations. Literally any information relevant to the individual researcher's interests can be part of a search. Because it is impossible to anticipate the needs of future researchers, search capabilities emphasize flexibility. The goal is to make it simple for researchers to discover the studies, variables, or populations relevant to them.

Although not part of XML, the structured form of XML files facilitates construction of hyperlinks between different parts of a file. Thus, once researchers have discovered relevant work, hyperlinks can direct them from individual questions or parts of a study to other documentation. For example, when researchers have found a question of interest, they can jump immediately to examine the population or sample used in the study. Because all documentation is in a standard electronic form, any other relevant information about the study is immediately available.

A major benefit of enhanced discovery is that secondary analysis utilizing multiple studies becomes much more feasible. This simplifies such work as longitudinal analysis of responses of the same population to the same question across time or cross-national analysis of similar questions.

Standardization creates new opportunities for software development to aid users. The advantages of leverage and widespread use are similar to those promised by open-source software. Indeed, a similar development process can occur surrounding the DDI. Software written for the DDI standard can be used by archives worldwide. As one archive enhances existing software, it can be shared with or licensed to other archives, thereby creating a community of software developers around the DDI. This process is already beginning; several projects are currently building on the base constructed by the DDI standard. Projects described in the scholarly literature include the Virtual Data Center (Altman et al., 2001) (www.thedata.org) and Nesstar, the Networked Social Science Tools and Resources project (Ryssevik & Musgrave, 2001) (www.nesstar.com). Other notable projects include the National Historical Geographical Information System project (NHGIS) designed to integrate "all available aggregate census information for the United States from 1790-2000" (www.nhgis.org), the Cultural Policy and the Arts National Data Archive (CPANDA) (www.cpanda.org), the Council of European Social Science Data Archives (CESSDA) project integrating the catalogs of European data archives based on the DDI (www.nsd.uib.no/cessda/IDC), and the European Union funded Multilingual Access to Data Infrastructures of the European Research Area (MADIERA) project based on the DDI (www.madiera.net).[15]

Although good documentation is expensive, the DDI can help reduce the cost. Because XML is widely supported, documentation can be developed using standard software tools, some of which are freeware. The DDI web site contains a current, annotated list of tools, including tools developed specifically for the DDI (see www.icpsr.umich.edu/DDI/users/tools.html). As data-creating software systems—such as CATI and CAPI systems (computer aided telephone/personal interviewing)—proliferate, they already contain much of the information needed to document the data they collect. They can use the DDI directly to generate appropriate high-quality documentation.

When documentation was not electronic or standardized, software packages simply provided facilities to input raw data and create system files. Standardized documentation of data files means that it is both feasible and sensible to write software that can read the standard documentation and automatically generate appropriate system files.[16] This can remove a major source of error and a time-consuming task for researchers. Furthermore, the resulting files will be more complete because, by default, they will contain full labels for variables and values, whereas a researcher would often only document the variables that were supposed to be part of the analysis, and only document them as briefly as possible.[17]

Because both the data and associated documentation are in standard electronic form, it becomes possible for statistical software to read them directly. This makes possible statistical analysis online via a web browser without additional software. Although online statistical analysis has been implemented for certain data sets for more than a decade, heretofore it required extensive special programming. As a result of the DDI, it may become a routine, standard offering for any data set in an archive. This can be valuable for instruction because students do not have to buy and install software and local data centers do not have to maintain student versions of software and support classroom use. Statistical analysis has been implemented in software projects such as the Statistical Documentation and Analysis (SDA) software from the Computer-assisted Survey Methods Program at Berkeley (http://sda.berkeley.edu) and Nesstar (Ryssevik & Musgrave, 2001) (www.nesstar.com). For example, Nesstar "allows users to browse distributed data catalogues over the web, examine detailed information about the data (metadata), carry out simple data analysis (e.g., tabulations and graphical displays) and then download data"[18] (www.nesstar.com).

We have described a large set of benefits to users that stem from the use of standardized electronic documentation. Some of these benefits have been available in the past on a limited basis, usually for specialized projects with special funding. Now it will be possible to provide enhanced functionality on a routine basis.

THE DDI ALLIANCE

Social science data come in bewildering variety. The impact of small computers in the 1980s and the Internet in the 1990s fostered development of more complex forms of data. The 2000s show a rapid increase of qualitative data, including text, pictures, audio, and video. There is no reason to believe that the development of more varied and more complex data has ended. For all these reasons, providing a documentation standard is not a goal but rather a process. Standards such as the DDI need to be updated to support changing kinds of data, and changing needs of researchers and teachers.

The ICPSR and the Roper Center for Public Opinion Research have taken the lead in creating a self-supporting organization specifically to continue development of the DDI, the Alliance for the Data Documentation Initiative or DDI Alliance.[19] The Alliance Steering Committee also includes the officers of IASSIST and CESSDA, continuing the close association of the DDI to professional data organizations. The DDI Alliance began operation on

July 1, 2003, with a core of about 25 members. Members of the alliance are data archives, universities, government agencies, and other institutions that would like to participate in further development of the DDI. Members pay yearly dues, send representatives to expert committee meetings, and vote on changes to the DDI.

The alliance has established specific goals for expansion of the DDI, including the following items: extending the DDI so that it can document more complex file types, document groups of related files, and document spatial data; establishing a repository of examples of tagged documentation; developing public domain software tools to help organizations tag documentation according to DDI standards; maintaining a clearinghouse for public domain DDI-related software; simplifying machine processing by developing controlled vocabularies for as many attributes as possible; and facilitating data exchange between the DDI and other bibliographic software (such as the Dublin Core, Giles, MARC, etc.) by developing lists of corresponding elements. For further details, see `www.icpsr.umich.edu/DDI/org/index.html`.

CONCLUSION: THE DDI AND THE SOCIAL SCIENCES

Documentation must serve multiple purposes. It must fully document the details of studies including the instrument used, the sample, the response rate, and other relevant information. It must be friendly and accessible to both novice and expert users. Because the Internet is the medium of choice for information search, documentation must integrate into the infrastructure on the web, including the ability to search across and within studies at the study- and variable-levels. It must be able to support automatic generation of system files for popular statistical software. The DDI can serve all these purposes and more.

The use of the DDI will produce a powerful improvement in access to a vast range of archival datasets. Expanded use of these data has significant implications for the social sciences. As data accumulate over many decades and societies, enhanced access makes new studies possible and may lead to a significant improvement in our understanding of changes across time as well as differences between societies: both longitudinal and comparative studies become more feasible. This is a way of saying that analysis of secondary data is becoming an important growth area in the social sciences. It is enhanced by increasing the availability of high-quality data. Rich, full, easily accessible documentation is a necessity for this to become a reality. The Data Documentation Initiative promises exactly that: by enabling flexible, user-friendly ways to describe studies, flow through networks, and display in new ways the research processes mediated by that documentation could improve. It will make us more productive and enhance our knowledge.

NOTES

[1] For discussions of the value of sharing data, see Hauser (1987) and Fienberg, Martin, and Straf (1985).

[2] Even where archives have completely converted their documentation to electronic form, like the largest archive, the Inter-university Consortium for Political and Social Research (ICPSR), most documentation is only available as PDF files. See below for a discussion of the weaknesses of PDF files as a documentation form.

[3] PDF files are searchable, but only as individual files, not across files. Their most significant weakness is that they are unstructured. See below for a discussion of the problems of unstructured text.

[4] For study-level documentation, OSIRIS provided a single record type, called S records. By use of a reference field, S records could store limited information about the study. This capability was not widely used. The other documentation formats developed during that time, including the best known SPSS and SAS system files, described only the content of data files such as variables and values. For a discussion of the many design weaknesses that limited SPSS's and SAS's ability to produce adequate documentation, even at the variable level, see Blank (1993). OSIRIS

has continued mostly as a way to store electronic documentation. During the 1970s through the 1990s, the ICPSR and the Danish and Swedish archives developed a series of software preprocessors and filters to create and transform OSIRIS codebooks (Rasmussen, 1996, 2000).

5. Hereafter we use *data archives* and *data libraries* as synonyms. Both refer not only to the national data archives common to most countries but also to the American system of regional data centers and distributed data archives.

6. Several 1970s efforts are described in Anderson (1974). Anderson's description of these discussions presciently foreshadows many issues confronted 25 years later by the DDI project.

7. An important exception to this statement is Sue Dodd's remarkable accomplishment in creating a MARC record for studies (Dodd, 1982).

8. A few years earlier, the National Opinion Research Center developed its own electronic codebook, see Blank (1993). Because its designer served on the DDI committee, this project influenced the DDI, especially the variable-level tags.

9. The committee was chaired by Merrill Shanks, University of California, Berkeley, from 1995 to 2002. Bjorn Henrichsen, Norwegian Social Science Data Service, was chair 2002 to 2003. Committee members included Micah Altman, Harvard University; Atle Alvheim, Norwegian Social Science Data Services; Martin Appel, U.S. Bureau of the Census; Grant Blank, American University; Ernie Boyko, Statistics Canada; Bill Bradley, Health Canada; John Brandt, University of Michigan; Cavan Capps, Bureau of the Census; Cathryn Dippo, Bureau of Labor Statistics; Pat Doyle, U.S. Bureau of the Census; Terence Finnegan, National Center for Supercomputing Applications; Dan Gillman, Bureau of Labor Statistics; Ann Green, Yale University; Lynn Jacobsen, Columbia University; Ken Miller, ESRC Data Archive, University of Essex; Tom Piazza, University of California, Berkeley; Karsten Boye Rasmussen, University of Southern Denmark; Richard Rockwell, The Roper Center; Jostein Ryssevik, Norwegian Social Science Data Services; Wendy Thomas, University of Minnesota; Rolf Uher, Zentralarchiv fuer Empirische Sozialforschung; and Bridget Winstanley, ESRC Data Archive, University of Essex. The ICPSR provided staff support by Peter Granda, Peter Joftis, and Mary Vardigan.

10. This information is from an internal ICPSR document.

11. PDF files are electronic and can be moved across the Internet, but this is the limit of their strengths. Although they are human readable, they are not structured. Thus, their contents are not understandable to a computer. The remainder of this section describes the value of structured electronic documentation. Although PDF files are not regarded as the most desirable form of documentation, they are an acceptable electronic migration for paper documentation. Adobe recognizes the limits of PDF files and it has been working to integrate XML into the PDF file format (Udell, 2003). Although this suggests that future PDF files may be used with structured data, it does not offer a solution for documentation that has been scanned from paper into PDF files. Conversion of unstructured PDF files into structured data is extremely expensive.

12. This example is adapted from the ICPSR ANES CD-ROM American National Election Studies 1948-1994.

13. The DDI started with the intention to build a DTD in Standardized General Markup Language (SGML), which was developed by Charles Goldfarb for use primarily by the publishing industry. SGML became an ISO standard (ISO-8879) in 1986. HTML is a DTD in SGML. XML has to a very large extent the same flexibility as SGML for creating DTDs. In 1996, the World Wide Web Consortium SGML Working Group announced XML with the special interest of creating a facility for more convenient document support on the Internet. Subsequently, the DDI was migrated to XML to take advantage of the increased web functionality.

14. Plans for the next version of the DDI include converting it to an XML schema. Schemas offer significantly more powerful capabilities than DTDs, including ability to define local element types (in DTDs, all elements are global), type inheritance, and namespaces. For information on XML schemas, see www.w3.org/XML/Schema.

15. For a list of other projects, see www.icpsr.umich.edu/DDI/codebook/projects.html.

16. Although this is reasonable for large statistical software packages, there exist hundreds of specialized statistical programs that do not command the large markets or extensive staff of the major packages. Again, tools have been developed to extract data from DDI documented files so that it can be read into other software without great effort. See the DDI web site for current details.

17. The problem areas as well as different software for converting data and documentation are described in Rasmussen (2000, pp. 356-362).

18. The triple-s should also be mentioned as a standard designed for survey interchange. Its debt to the DDI is acknowledged in Hughes, Jenkins, and Wright (2000). Compared to the DDI, the triple-s is much less broad. It focuses on the description of a few elements of the complete data documentation.

19. Prior to forming the DDI Alliance, the DDI project was funded mostly by the ICPSR. It received external support from the National Science Foundation and from Health Canada.

REFERENCES

Altman, M., Andreev, L., Diggory, M., King, G., Sone, A., Verba, S., et al. (2001). A digital library for the dissemination and replication of quantitative social science research. *Social Science Computer Review*, *19*, 458-470.

Anderson, R. E. (1974). Reducing incompatibilities in social science software and data: Current and proposed effort. *Social Science Information*, *13*, 147-160.

Blank, G. (1993). Codebooks in the 1990s; or, Aren't you embarrassed to be running a multimedia-capable, graphical environment like Windows, and still be limited to 40-byte variable labels? *Social Science Computer Review*, *11*, 63-83.

Dodd, S. A. (1982). *Cataloging machine-readable data files: An interpretive manual*. Chicago: American Library Association.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing research data*. Washington, DC: National Academy Press.

Green, A., Dionne, J., & Dennis, M. (1999). *Preserving the whole: A two-track approach to rescuing social science data and metadata* (Technical Report 83). Washington, DC: Council on Library and Information Resources. Available at www.clir.org/pubs/reports/reports.html

Hauser, R. M. (1987). Sharing data: It's time for ASA journals to follow the folkways of a scientific sociology. *American Sociological Review*, *52*, vi-viii.

Hughes, K., Jenkins, S., & Wright, G. (2000). Triple-s XML: A standard within a standard. *Social Science Computer Review*, *18*, 421-433.

Nielsen, P. (1974). *Report on standardization of study description schemes and classification of indicators*. Copenhagen, Denmark: Danish Data Archives.

Rasmussen, K. B. (1978). Technical standards for magnetic tape exchange. *IASSIST Newsletter*, *2*(3), 76-77.

Rasmussen, K. B. (1989). Data on data. *Proceedings of the SAS European Users Group International Conference 1989* (pp. 369-379). Cary, NC: SAS Institute.

Rasmussen, K. B. (1996, May). *Convergence of meta data. The development of standards for the description of social science data*. Paper presented at the 1996 Population Association of America Conference, New Orleans, LA.

Rasmussen, K. B. (2000). *Datadokumentation. Metadata for Samfundsvidenskabelige Undersøgelser* [Data documentation: Metadata for social science studies]. Odense, Denmark: Universitetsforlag.

Ryssevik, J., & Musgrave, S. (2001). The social science dream machine. *Social Science Computer Review*, *19*, 163-174.

Sieber, J. E. (1991). Introduction: Sharing social science data. In J. E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 1-18). Newbury Park, CA: Sage.

Udell, J. (2003). Acrobat challenges InfoPath. *InfoWorld*, *25*(32), 36.

Grant Blank is an assistant professor of sociology at American University, Washington, D.C. His special interests are in the sociology of culture, the influence of computers and electronic networks, and analysis of qualitative and quantitative data. He has been a member of the DDI committee since 1995. He may be reached by e-mail at grant.blank@acm.org.

Karsten Boye Rasmussen is an associate professor in IT and organization at the University of Southern Denmark. His special interests are in data warehouse and data mining. He was a member of the DDI committee from 1995 to 1999. He chaired the IASSIST Codebook Action Group and is editor of the IASSIST Quarterly. *He may be reached by e-mail at* kbr@sam.sdu.dk.