Arch Sci (2007) 7:55–71
DOI 10.1007/s10502-006-9036-0

ORIGINAL PAPER

The data documentation initiative: a preservation standard for research

Karsten Boye Rasmussen · Grant Blank

Received: 10 March 2005 / Accepted: 15 November 2006 / Published online: 24 May 2007

Introduction

When the answer is the number ''42'' (or actually ''forty-two'') some know instantly that the question is ''The Great Question'' concerning ''Life, the Universe and Everything ...'' (Adams 1986, p. 128). This demonstrates that even when we know the answer to a question its meaning and usefulness are not always obvious. Context is required. This is especially true for quantitative data. More information is needed in order to understand the numbers and transform data into useful knowledge. This further information is itself data, and thus metadata or ''data about data''. The importance of context and metadata are widely recognized in the social science community. This paper discusses a project to provide standardized metadata to document social science datasets: the Data Documentation Initiative (DDI). The two authors worked together on the DDI committee and are authors of a previous article on the DDI (Blank and Rasmussen 2004).

We begin by discussing the fundamental problems of social science dataset documentation. We focus on the value that standardized documentation in the form of the DDI creates for the social science community.[1] To understand the virtues of the new, we look briefly at some historical documentation standards. We describe the fundamental features of the DDI standard by describing a current application. The DDI is a standard still being

[1] Although we focus on social science researchers, other stakeholders have an interest in the DDI: archives, funding agencies, the layman, society as a whole. Methodologically we describe the value for researchers based on the value of high-quality documentation as well as the ability easily to find relevant data.

K. B. Rasmussen (✉)
University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
e-mail: kbr@sam.sdu.dk

G. Blank
American University, 4400 Massachusetts Ave NW, Washington, DC 20016-8072, USA
e-mail: grant.blank@acm.org

developed and we close with a discussion of future development plans and a summary of the value of the DDI for research.

The background

Quantitative research data consists of values and categories describing characteristics of entities, or to use object-oriented systems terminology, the categories and values are descriptions or measurements of the attributes of objects. The object could be an individual (a person), an artifact (a phone, a car etc.) or an aggregation in the form of some social or geographic gathering (an organization, a political party, a city, a region, a nation, etc).[2] Furthermore the object could be a representation of an act like the transactions performed by the former representations (a phone call made on a certain phone, a phone call made by a certain person, a sales receipt from a cash register, a car trip, etc.). Typical social science data contains information about the object (often an individual) using numeric representations of the data attributes. The data are often collected via a questionnaire. The complexity and need for contextual information is clear from the following glimpse of a numeric data file.

Figure 1 illustrates—just like the opening anecdote—that without appropriate context data are meaningless.

```
20911242291000000040000100050706112121011241191010000030001000080305051021211222291000000040001000101001082021311222291020000030100010101005072121511221191000000031000000101004045010111323390000201020000010091001081110911111191000200030001000050606061010611134220010000030100000030105072021611221191000000010000010101002021011141132291000100020101000010205072010821142291000000001100000004090508202011122229100000000100010101010021220112111222910000000101000001010030521203112242200100000200000100910030631.....
```

Fig. 1 Numeric data

Data—information—knowledge

Peter Checkland—the British founder of the soft systems development—and Sue Holwell (Checkland and Holwell 1998) supply a useful starting point. ''Data'' are viewed as a cloud of jumbled, unordered, formless facts. ''Information'' refers to ''meaningful facts'' where meaning is derived from context. The point is important: meaning requires context and meaning emerges from context. ''Knowledge'' is the ''larger, longer living structures of meaningful facts''. How do we get from data to knowledge? According to Checkland and Holwell our first task is to select. Billions of data could be collected on a person; we select the data relevant to us. Checkland and Holwell (1998, p. 90) propose the term ''capta'' for what is actually and actively selected. In order to make use of the data, the analyst has to have information on the process of selecting this ''capta''. This information is the metadata or data description.

Thus, when we have data, we require data description in order to do useful analysis. The data description is the key that supplies meaning to data. The point is that because data are

[2] The basic problems of qualitative data are quite similar. In order to present more easily understandable description this article concentrates on quantitative data and exemplifies the problems associated with it [See the article by Louise Corti on qualitative data elsewhere in this issue; the editors].

useless without the meaning supplied by the information embedded in the description, preservation of the data description is as important as preservation of the data itself.

The use of computers

Data, such as the stream of numbers shown in Fig. 1, are described as ''machine readable'' when stored in a computer and this implies that these data will be processed by computers. Streams of numerical data cannot be used either by computers or by people without additional information. We need to know what each of the numbers represents. For example, assume that the data file in Fig. 1 contains information about individuals, that every person's data record consists of 36 bytes (characters), and that the 18–19th byte in each record contains information about the age-category for that particular person. Further, assume that this is the 12th variable describing an individual. The information documenting what each number represents is metadata.

Metadata must be ''person readable'' because it carries meaning. Ever since data were first stored on computers, person readable documentation has been required. For decades it was stored in paper form. Now metadata is almost always stored on computers, so it is also machine-readable. Whether data is readable in the sense that it conveys meaning is a question of the abilities of the human observer, the data itself, and the software that processes the data. Schematically the metadata for the file in Fig. 1 could be presented in a formal way as shown in Fig. 2.

Fig. 2 The information for variable extraction

```
RECORD=36
V12="AGECAT",COL=18-19
```

Using a standard format for the metadata, software is able to extract the location of the age category variable. Once it has the location information, software is able to extract age category information for all individuals in this particular file. Extraction of the variables in the demonstration dataset will result in a data matrix with the cases (entities/rows) and variables (attributes/columns) presented in Fig. 3.

209	1	12	4	2	2	9	1	0	0	0	4	0	0	10	0	50	70	61	121
210	1	12	4	1	1	9	1	1	0	0	3	0	1	0	0	80	30	50	510
212	1	12	2	2	2	9	1	0	0	0	4	0	1	0	1	1	0	10	820
213	1	12	2	2	2	9	1	2	0	0	3	1	0	1	1	1	0	50	721
215	1	12	2	1	1	9	1	0	0	0	3	10	0	0	1	1	0	40	450
101	1	13	2	3	3	9	0	0	2	1	2	0	0	1	0	91	0	10	811
109	1	11	1	1	1	9	1	0	2	0	3	0	1	0	0	50	60	60	610
106	1	11	3	4	2	2	0	1	0	0	3	1	0	0	0	30	10	50	720
216	1	12	2	1	1	9	1	0	0	0	1	0	0	1	1	1	0	20	210
111	4	11	3	2	2	9	1	0	1	0	2	1	1	0	0	10	20	50	720
108	2	11	4	2	2	9	1	0	0	0	1	10	0	0	0	40	90	50	820
201	1	12	2	2	2	9	1	0	0	0	1	0	1	1	1	1	0	21	220
112	1	11	2	2	2	9	1	0	0	0	1	1	0	0	1	1	0	30	521
203	1	12	2	4	2	2	0	1	0	0	2	0	0	1	0	91	0	30	631

Fig. 3 The data matrix

The true beauty of the idea of a standard metadata format lies in the fact that the software that can find the location of the age category variable will also be able to find the location of any and every variable in a file. Further, it can read the location of variables in any other file that has been documented using the same standard format. It is a general tool. The same software would for example be able to extract information from a file containing car trips between two destinations, given the metadata in Fig. 4.

Fig. 4 The information for variable extraction for another file

```
RECORD=32
V9="DEPARTURE",COL=16-19
V10="DESTINATION",COL=20-23
V11="TRIP",COL=24-26
```

A standardized format for the metadata allows data in many types of files to be read and processed.

Data documentation initiative

Analysing undocumented data is impossible. But even with documentation the process of analysis is often difficult (e.g., the user must be able to understand the jargon of the documentation), error prone (e.g., the documentation might be imperfect, and/or the user might misunderstand the documentation), and time-consuming (e.g., users have to familiarize themselves with the documentation and the software). Providing a standard format for machine-readable metadata can reduce errors and simplify analysis. From these reasons, the DDI is intended to become the cornerstone of many scientific infrastructure projects.

The standard data matrix—as presented in Fig. 3—is a matrix of entities (e.g., individuals stored as rows) having attributes (columns or variables). Quantitative social science research relies on many similar data matrices: social surveys, psychological test measurements, economic and financial series, government statistics etc.

Basic data documentation

The rationale for the DDI is based upon a multi-faceted view of data documentation that goes beyond the limited variable information available in statistical software. The rationale focuses on the actual information stored and the reasons for storing it. In addition the rationale includes the additional value that good documentation adds to research. This section examines the fundamental elements of data documentation and their accompanying benefits.

The dictionary and codebook

Basic documentation simply describes each variable. Historically basic variable information (like the location or number and short description of the variable—‘‘column 18–19’’ and ‘‘AGECAT’’) is often called a ‘‘dictionary’’. However, the short description or the plain identification (as in ‘‘V12’’), does not supply sufficient information, so textual

documentation of the variable often included both a name and a longer label carrying a somewhat more comprehensive description ''AGECAT 10 YRS 2003''. A longer description of the variable might be needed in order to obtain valid information from the data; for instance age is dependent upon the time the data were collected.

Most variables are numerically coded; this means that even though fancy tabulations can be run from datasets containing only limited variable descriptions, without value-level information the tabulations will not carry information. The ''codebook'' contains further descriptions below the variable level including an explanation of the codes stored in the numeric dataset. For example: numeric code ''1'' stands for ''Male''; ''2'' for ''Female'', another code stands for missing data, etc. Obviously without this information tabulations will be impossible to understand.

The producers of social science statistical software are aware that documentation is important for analysis. Most statistical software allows documentation but it is often limited to the location and name of individual variables (as exemplified in the Figs. 2 and 4 above) plus labels for variables and categories; in short, a dictionary and a codebook. In order to supply meaning, documentation has to specify more than that (Blank 1993; Rasmussen 1989).

Both the dictionary and the codebook document individual variables. For this reason, they are often referred to collectively as the ''variable level'' documentation. There is another level of metadata that we describe next: the ''study level'' documentation.

Further elements of the structured study description

From an analytical viewpoint we are getting closer to meaningful information. A table using the dictionary and codebook can show the distribution of entities on a particular variable but information about variables is not useful if they cannot be precisely linked to the selection of objects. In order to interpret our results we need to know the population, and if this is a sample, we need information on sampling procedures, appropriate weighting, and other information about the study. Futhermore, most data are time-dependent and this requires knowing the dates when data were collected. Data might have relationships with other data; this could be because they contained data about the same objects, or because the same instruments had been used for measurement, etc. Long-running or complex studies often store data in multiple files and the relations between files must be documented. All these elements—and more—have to be described. This is information about how the collector or researcher turn data into ''capta''. This holistic approach can be called ''structured'' when referring to the DDI as ''structured codebook standard'' (Green et al. 1999, p. 31).

The electronic transmission of copies of the data file creates a need for additional documentation. During transmission or copying the data may be jeopardized by unintended corruption. Consequently basic documentation also includes record counts and complete frequencies or descriptive statistics for every variable. Such redundancy ensures that the data received are the same as the data sent. Further requirements include a precise description of the storage media and some basic checkpoints.

Scientific benefits of documentation and archiving

Archiving in the digital age always means archiving both data and metadata. The benefits of documentation are linked to the benefits of archiving. The archiving and availability of

high quality documentation will benefit research in the social sciences for the following reasons:

(1) The use of metadata simplifies understanding and reuse of data, thus facilitating secondary analysis. For discussions of the value of sharing data, see Sieber (1991); Hauser (1987) and Fienberg et al. (1985). A summary with some American cases are found in Fienberg (1994); they concentrate on health data, but are built upon experience with social science investigations. A special case of reuse is the case where the person performing secondary analysis is the original researcher. It is obviously formally easy to reuse your own data, but returning to the data long after details have faded from human memory is extremely difficult. Metadata with easy access and comprehensive coverage can be of great value to the original researcher.
(2) Empirical social science research is expensive, and much of the cost is due to the expense of data collection. Original data collection is much more costly than secondary analysis. Consequently, it is wise behaviour to leverage expensive data in as many secondary ways as possible.
(3) Secondary analysis is often the only possible way to investigate a time that is now past. The benefit of strong documentation is that this research would not be possible at all, had the data not been properly documented and preserved in an archive.
(4) A standard for quality metadata provides in effect a checklist of key information. This forces the original investigator to systematically and rigorously understand and describe the data—which may improve both collection and analysis. This raises the quality of the investigation itself.
(5) The social sciences benefit from firm support for systematic, cumulative building on prior knowledge. The addition of secondary data to newly collected primary data is often a source of new knowledge creation. Documentation is crucial in order to be able to ''match'' multiple datasets.[3]
(6) Data acquisition can benefit from the data documentation. For instance the metadata describing headings, questions, codes, and structure of a questionnaire can be used as input to automatically generate the computer software that is used to capture the data; e.g. setting up the progression, the logic, and the screen layouts. This will further improve the rigor of the questionnaire.

In addition to the benefits of standard metadata there are also disadvantages and costs.

(1) The production of the metadata demands resources, primarily work time of researchers and assistants. Time requirements are a fierce impediment. If someone other than the original researcher produces the metadata then the costs will be still greater.
(2) Metadata demands knowledge. Obviously knowledge about the data and the research being carried out is needed, but the further cost is that it demands knowledge of metadata description, the structure and content of metadata and some technical insight regarding the practical arrangement of the metadata.

These are important disadvantages, particularly for small studies. However, individual researchers gain substantial benefits from the checklist as well as from generating the data

[3] We use the somewhat vague term ''match'' here. It does not imply a match-merge of the two datasets at a record level nor is our intention to imply a match on similar variables. The term is used in the broadest sense where similarities and dissimilarities between the two datasets can be noticed and acted upon in the process of further research.

acquisition software. The further benefits for the social sciences in general stem from the improved quality of secondary research in general. These benefits far outweigh the costs.

The accumulation of knowledge in science is supported through peer-reviewed publications. Many of the strengths of empirical social science research derive from the ability of other researchers to have access to data to replicate and validate the original scientific findings, as well as use the data in new and different ways. Some journals now require that data be publicly archived and available to other researchers. An example from the *American Journal of Political Science*: ''AJPS requires that all manuscripts containing analysis of quantitative data contain an initial footnote that states how individuals can obtain the data and the documentation of statistical analysis necessary to replicate the paper'' (quoted from Meier (1995) in Rasmussen (2000)). With the growth of the Internet, journals now assume that an interested party can contact the author for access to the original data.

Most countries have boards for judging scientific dishonesty. In the description of the purpose and work of the Danish Committee on Scientific Dishonesty is stated:

> Scientific dishonesty shall mean intentional or grossly negligent conduct in the form of falsification, plagiarism, non-disclosure or any similar conduct involving undue misrepresentation of a person's own scientific work and/or scientific results'' (Danish Committee on Scientific Dishonesty 2005).

Examples of possible dishonesty include construction of data, selective and hidden cassation of unwanted results, substitution with fictitious data, or knowingly erroneous use of statistical methods. Very few actual breaches are found, as mentioned in reports from a Danish scientific committee (Rasmussen 2000, p. 171, Fig. 119). Data documentation supports the cumulative and ethical standards of science by making validation possible and by improving the quality of the science.

A short history of documentation

To understand the objectives of the DDI it is necessary to understand how data archives and libraries previously dealt with documentation. This clarifies the goals of the DDI standard.

Paper documentation

Long after data began to be stored in digital form, most archives continued to keep documentation on paper. Having two separate media—one for data and another for documentation—required separate systems for processing, storage, retrieval, and dissemination. Paper documentation has a variety of problems: storage is expensive, it requires careful controls to track it in inventory, and it deteriorates over time. In contrast, digital copies can be produced on demand or made directly accessible. 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 upon archive personnel and their ability to find—or remember—the appropriate studies.

Increasingly over the past 10 years, documentation has been 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 and because an adequate mapping between existing formats could not be constructed. The result has been that access to archived datasets for secondary analysis has been limited. The promise of the Internet—to make information easily and widely available—has remained unfulfilled.

Machine readable documentation

Data archives have been very much aware of these limitations. They have been systematically transferring paper documentation to digital form (as characters or images) since the 1960s.[4] In the 1970s the ICPSR and other archives developed electronic documentation in the form of the OSIRIS codebook. This was a capable documentation standard focusing on the variable level, although OSIRIS allowed limited information about the study. During the 1970s through the 1990s, the ICPSR, and the Danish and Swedish archives developed a series of software preprocessors and filters to add further documentation capabilities to OSIRIS files (Rasmussen 1996, 2000). Everything about computing in the 1970s was very expensive, and electronic documentation was also expensive. Only major archives and a few large government data producers could afford the cost and were aware of the potential savings of thorough documentation. Smaller organizations usually provided no electronic documentation at all.

The other documentation formats developed during that time, including the best known SPSS and SAS system files, described only the content of data files like variables and values. This had serious limitations, even for documenting statistical output. 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).

The 1980s introduced personal computers (PCs) and attitudes shifted. The new emphasis was on quick dissemination of datasets 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 non-standard, heterogeneous, individualized, personalized crypts of data deposits. Archives were supporting hardware and software on microcomputers, minis, and mainframes. Support for all these platforms drove up costs. Furthermore, 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. Users realized they could obtain direct, fast access to data without geographical restrictions. Furthermore, users wanted access to complex search systems to help precisely identify relevant datasets. It became obvious to data disseminators that these wishes could only be met by developing widely accepted, international 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

[4] 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 graphic images of the documentation pages (stored in PDF files). Such images are unsearchable. Even when the archive does optical character recognition (OCR) on the images and produces searchable text, there remain major problems. The most significant weakness is that such documentation is a stream of text that cannot support field-structured searches.

data archivists—formed the ''The IASSIST Codebook Action Group'' to work on the 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. This came to be the DDI.

Data documentation initiative

The term ''data documentation'' includes relevant metadata at both the variable-level and the study-level. The product of the DDI is standardized, highly structured, electronic documentation. ''Standards are nice, as there are so many to choose among''! Even though a standard can be officially certified only by a standardization body (like the International Standards Organization or ISO), we view the DDI as a practical standard because it is widely and increasingly used. The DDI is a de facto standard. The DDI draws upon many prior standards including:

- The *OSIRIS codebook*: This was the most comprehensive documentation format for codebooks and the format in use at the ICPSR as well as at some archives in Europe.
- The *Standard Study Description*: A very elaborate descriptive scheme at the study level. Several different dialects were used in European archives as well as the ICPSR. The work on the Standard Study Description was reported with the Scheme in 1974 (Nielsen 1974). The Standard Study Description Scheme underwent continuous improvement and the latest version is found as a basis of decision in a report for a conference (Rasmussen 1981).

The following standards mentioned are now collectively described in the publication on metadata from NISO (2001).

- The *Dublin Core*: A core of descriptive elements (for Web resources) published in 1995 by OCLC (Online Computer Library Center) and NCSA (National Center for Supercomputing Applications). Several elements including the Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights are part of the DDI.
- The *Text Encoding Initiative*: The ''TEI'' is a standard for marking up texts for research in the Humanities. Like the DDI, the TEI has a strong research emphasis; also the word ''Initiative'' came from the TEI.
- *Machine Readable Cataloging*: MARC is the library record of an object (e.g., a book). Objects themselves can be machine readable and this created special problems that are discussed by Dodd (1982).

Several other standards were influential on the development of the DDI. This list first of all illustrates the point that there are many standards to choose from. Secondly, even though some of the information in the DDI could be stored in existing formats, no other standard attempts to comprehensively define the metadata needed to document quantitative files.

Structured documentation in XML

To support publishing Charles Goldfarb and others developed Standardized General Markup Language (SGML). In 1986, SGML became an ISO standard (ISO-8879). SGML

is a broad, flexible markup language, specifically designed to allow more-specific types of documents to be defined within it. Specific types of documents are defined by a Document Type Definition, or DTD. The standard language used to describe web pages, HTML, is a document type defined in SGML, but HTML is directed towards screen display of web pages: it instructs a computer what text should be bold, what should be italics, what should be a normal paragraph (normal size characters), what should be a header (larger size, bold) etc. The DDI started with the intention of building a DTD for metadata using SGML. However in 1996 the World Wide Web Consortium (W3C) announced XML (eXtensible Markup Language), which was much simpler and almost as flexible as SMGL. XML is designed to support complex documents on the Internet. The DDI soon adopted the XML standard. Compared with HTML the strength of XML is that it separates content from display of information. XML allows creation of a set of rules to identify content by using highly structured data. For a more detailed discussion of the advantages of XML, see Blank and Rasmussen (2004).

To illustrate how XML works, we use a simple question: What is the sex of a respondent?

XML uses printable and person-readable ASCII characters for both content and for tags. There are no non-printing binary codes as in a Microsoft Word document. Tags are identified by pairs of angle brackets. A variable name tag could be '' <VARNAME> ''. The actual content—in this example variable name—is placed within pairs of those tags: one marks the beginning of the variable name text and another marks the end: '' <VARNAME> V58B </VARNAME> '', see Fig. 6. The fully tagged variable might appear in the file like this.

Notice first, that the tags in Fig. 6 carry no information about display, they only describe the content. The focus on content frees XML from any restrictions on the number or type of tags. In XML, it is possible to define whatever tags the content requires.

Secondly, XML is not concerned with display. There are no built-in instructions for the display of XML content-based tags. Display instructions can be added to XML 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 and the tagging can concentrate on content and delivering content. It is the choice of the Nesstar software to display the metadata as you see in Fig. 5.

A tagged document would not use the new lines, we added them to Fig. 6 to improve readability. The elements of the variable are represented by the tags, and the structure of the document is represented by the sequence and nesting of tags. The available tags and the structure for the document are defined in the Document Type Definition (DTD) for the DDI. The DTD defines the rules that a particular document has to obey to be a DDI-complaint document. All DDI documents use the same DTD, although any single study is unlikely to make use of all the available elements. The DDI DTD can be accessed from the web http://www.ddialliance.org/dtd/index.html. Both study level and variable level metadata is contained in standard, structured tags, making it easy to identify and retrieve. The DDI DTD thus delivers a flexible, versatile platform for documenting social science data.

The value of the DDI for research

From an archival viewpoint, secondary research is the focus. The process of secondary research can be viewed as a series of stages: retrieval, identification, access, analysis, and publication. The DDI can create value at each stage.

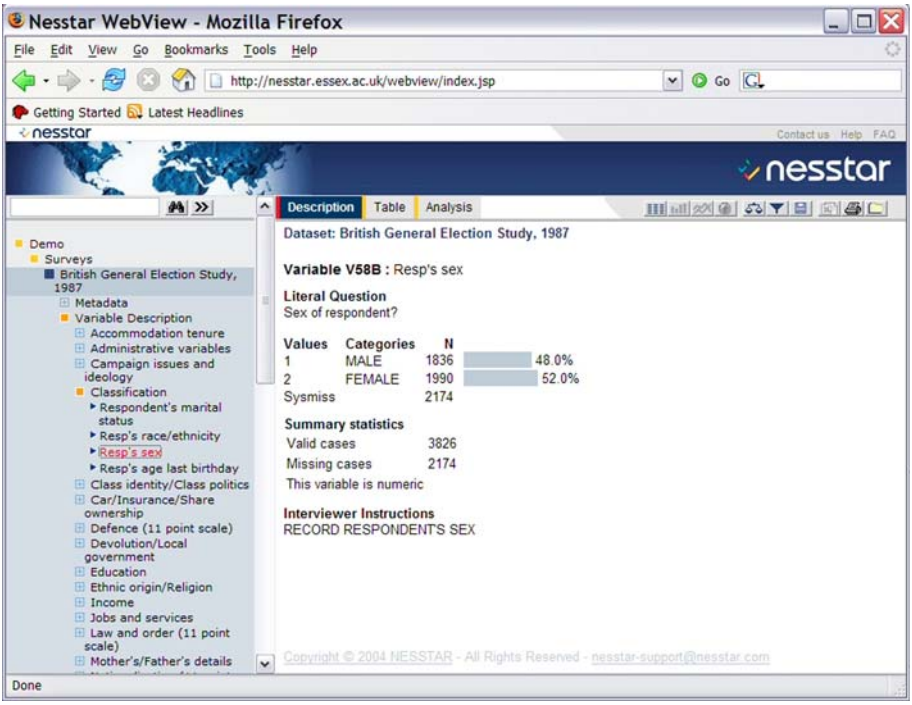

Fig. 5 The Nesstar display of a variable

Fig. 6 An XML-tagged variable

```
<var>
  <varname>V58B</varname>
  <labl>Resp's sex</labl>
<qstn>
  <qstnLit>Sex of respondent?</qstnLit>
  <ivuInstr>RECORD RESPONDENT'S SEX</ivuInstr>
  </qstn>
<valrng>
  <range min="1" max="2" />
  </valrng>
  <sumStat type="vald">3826</sumStat>
  <sumStat type="invd">2174</sumStat>
<catgry>
  <catValu>1</catValu>
  <labl>MALE</labl>
  <catStat type="freq">1836</catStat>
  </catgry>
<catgry>
  <catValu>2</catValu>
  <labl>FEMALE</labl>
  <catStat type="freq">1990</catStat>
  </catgry>
<catgry missing="Y">
  <catValu>Sysmiss</catValu>
  <catStat type="freq">2174</catStat>
  </catgry>
</var>
```

Examples of creating value for research by the DDI

The tagged example in Fig. 6 above makes it clear that software could perform a simple string search on DDI-documents; the tagged DDI structure makes more complex searches

possible. For example a researcher might be looking for a study with (A) more than 1000 respondents, (B) carried out after 1990, containing variables (C) age and (D) highest education. This searches logically for ''A and B'' at the study level as both have to be present. However, at the variable level the search is for the occurrence of ''C or D'', since a single variable containing both—''C and D''—age and highest education—is unlikely. The DDI makes such combinatory complex searches possible.

Not only can individual datasets be located, but searching a large archive may return whole lists of relevant datasets. Refining searches by expanding or contracting the time frame or the geographical area is relatively simple. This directly supports the development of research using more than one dataset for cross-national research or research across time.

Once a dataset is identified, software can read a tagged documentation file and extract the elements needed for analysis. With the DDI description it is straightforward to transform the tagged information into sets of commands or system files for major statistical packages (e.g., SAS, SPSS, STATA, SYSTAT). The data are accessible in the preferred format or statistical package. And future software formats can be supported as well.

The DDI dramatically improves access to datasets. Datasets become available everywhere as long as the researcher has access to the Internet and is authorized to access the data. In the future it will be possible to hyperlink publications directly to the data documentation. This may encourage further use of the data. XML features much more intelligent linking than that available in HTML.

Fundamentally, the DDI will improve potential access to and utilization of data. The DDI will improve the efficiency of research, and make better use of scarce research money. This is valuable not only for the researchers but also for the society. The hope of DDI designers is that better access to datasets will improve research and thereby transfer greater benefits to society at large.

Applications utilizing the DDI

The DDI is a free standard; there are no license fees and applications can be developed freely. Many major projects are now underway using the DDI standard. Some examples: The Virtual Data Center (Altman et al. 2001) describes itself as ''An operational, open-source, digital library to enable the sharing of quantitative research data'' (Virtual Data Center 2004–2005). The European Union originally funded Nesstar (Networked Social Science Tools and Resources) (Ryssevik and Musgrave 2001). Nesstar has since become a private commercial organization with strong bonds to archives in both Europe and North America (Nesstar 2004). Among census projects are the National Historical Geographical Information System (NHGIS) that intends to ''create and freely disseminate a database incorporating all available aggregate census information for the United States from 1790–2000'' (NHGIS 2004). In the USA there is the Cultural Policy and the Arts National Data Archive (CPANDA 2006). Several other European projects are also of significance. The Council of European Social Science Data Archives (CESSDA 1997) has been involved in projects integrating the catalogues of the European data archives and the European Union funded Multilingual Access to Data Infrastructures of the European Research Area (MADIERA 2005) project is based on the DDI. A more complete and updated list of DDI involved projects can be found at the DDI website (DDI 2005).

Nesstar as example

To give readers a better feel for how the DDI works in practice, we return to Figs. 5 and 6 above. They show how the XML DDI content for one variable (Fig. 6) is presented by Nesstar software (Fig. 5). The variable is a basic element of a social science quantitative dataset, but before a variable can be displayed the dataset itself has to be identified. Nesstar has the ability to search for possible datasets.

Structured searches on either study level or variable level information, or both can be combined. The example shown in Fig. 7 below is the Nesstar-server that supports individual-level data as well as aggregated data and time series. The demonstration shows how to specify a search for selected elements of documentation. The first screen contains a 2-pane window: a list of datasets appears on the left while the right side displays the contents of individual items for the dataset that is currently highlighted. The hierarchical structure of the documentation is duplicated on the screen. The interface allows users to drill down

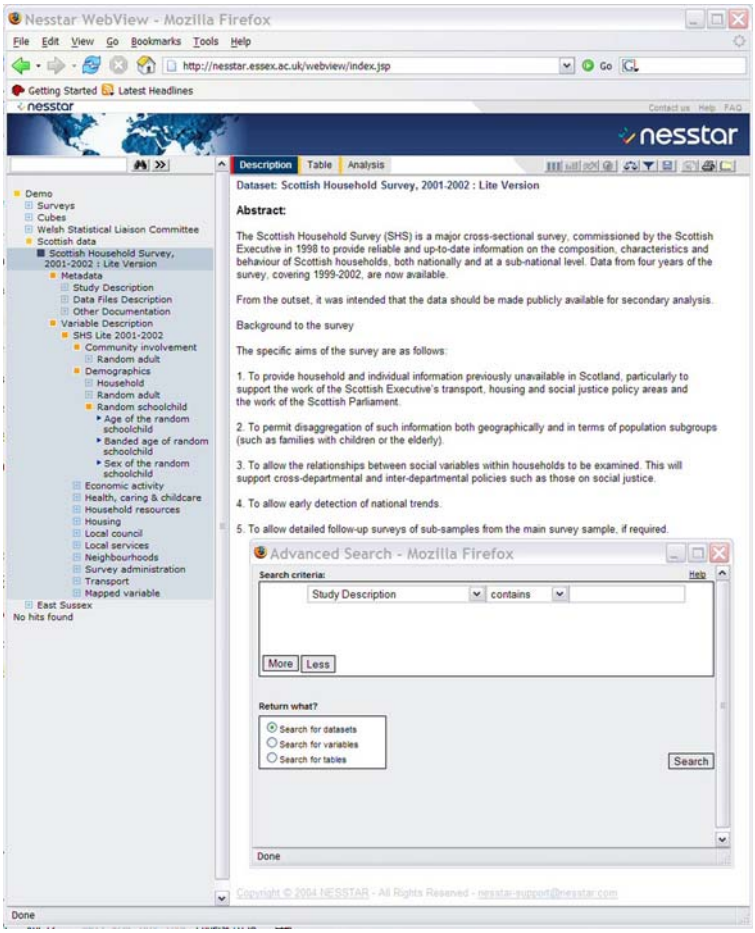

Fig. 7 The Nesstar server

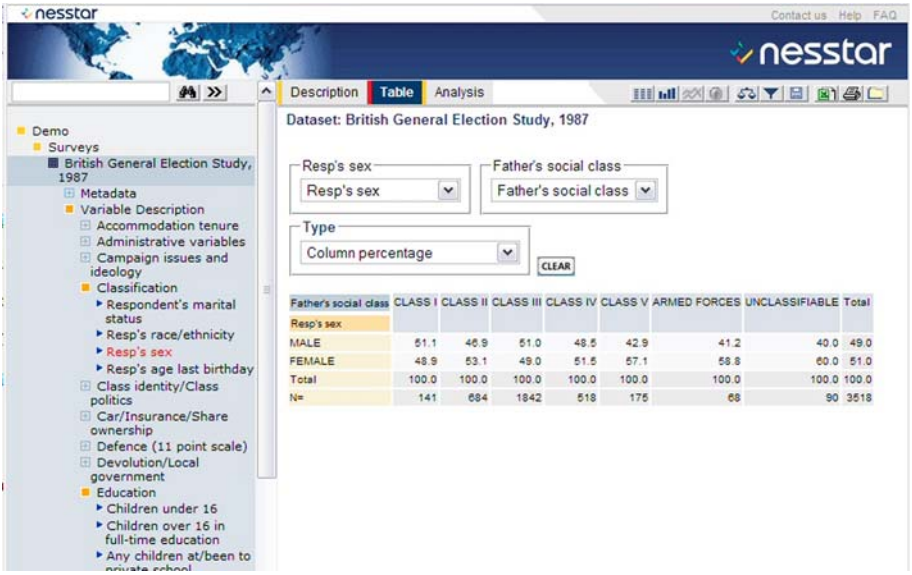

Fig. 8 The Nesstar tabulation

to obtain detailed information about specific topics of the study as well as selecting individual variables.

The ability to retrieve datasets via the Internet has been available from single archives (e.g., ICPSR) as well as at the level of archive associations e.g. CESSDA (Council of European Social Science Data Archives) developed a web-page (''The integrated data catalogue'') for searching European archives at the study level. However, only recently has it been possible to carry out serious analysis over the Internet. Figure 8 shows a tabulation created by Nesstar over the Internet.

The importance of this demonstration is to show that Nesstar or similar software can be used not only for searching datasets, but also for statistical analysis. The benefits of DDI documented datasets are not just a promise for the future, they have already been implemented in pioneering applications.

Applications producing the DDI

One of the major advantages of a standard is that software tools developed for specific projects by one archive can be reused for other projects in other locations. These network effects are among the most compelling features of the DDI project. Several software tools to assist in the creation of XML documents are listed on the DDI Alliance website http://www.ddialliance.org/DDI/related/tools.html. Among the tools is a program to convert data definition information stored in SAS, SPSS, and STATA files into tagged DDI XML elements. Nesstar Publisher software can, for instance, import an SPSS file and automatically create DDI documentation. Researchers can use Publisher to add DDI-complaint metadata without knowledge of XML. The Nesstar Publisher handles the technical details.

Instead of producing documentation by typing information at the end of a research project, it is also possible to use other systems to generate the XML elements of the DDI.

Data-creating software systems—like CATI, CAPI, and CASI systems (Computer Assisted Telephone/Personal Interviewing/Self-Interviewing)—often have their own specification language. Such specifications can be directly converted to DDI XML tags. For example, a Blaise-to-DDI converter, supports the widely used Blaise software for interviewing and survey processing. This capability means that large portions of DDI documentation can be generated automatically. Obviously it is also possible to generate the Blaise specifications from a DDI tagged file.

The impact of the DDI

We can summarize the preceding discussion of the impact of the DDI in the following points:

- Data preservation: Without documentation the preservation of data alone is not worthwhile.
- Access to the stored data: Elaborate metadata is necessary for an archive to deliver the precise information that a user finds relevant.
- Data collection: The DDI can serve as a checklist to improve the design of surveys, and also help with data collection through interfaces with special software.
- Data publishing: Governmental agencies may use the DDI as a means to publish data.
- Data analysis: Strong metadata improves the speed, accuracy, and reliability of analysis.
- Research community: The DDI facilitates easy replication and extension of research results, as well as simplifying secondary analysis of existing datasets.
- Development of tools: Standards facilitate the development of tools for creating, storing, manipulating, and presenting metadata to users.

Further development of the DDI

The original DDI was designed primarily to document the final dataset. In this design the documentation was largely a monolithic entity much like a paper codebook. It was not easy to document parts of a dataset over its lifecycle. The next version of the DDI, version 3.0, will introduce a new data model. A dataset can be thought of as having a lifecycle with multiple stages. The lifecycle begins with the initial concept of the study, it moves to the design stage, then the creation of instruments and a sampling frame, followed by data collection, cleaning, analysis, and preparation of a file for public release. The new model will facilitate documentation during each of the various stages of the lifecycle. The major change introduced will be a new modular structure. Researchers could select particular modules that document the study design, and only later document, say, the sampling frame, and still later individual variables. Such a modular structure makes the DDI more useful because it is easier to document each stage of a research project as it happens. Thus the DDI can serve as the core repository of information even in the early stages of a project. It will require reorganizing the DDI so that individual modules can stand alone but are easy to combine and still remain compatible with the DDI as a whole.

The introduction of smaller computers and PDA's combined with CATI and CAPI systems has fostered the development of more complex forms of data. There is no reason to

believe that the trend toward more varied and more complex data has ended. Thus, a documentation standard is not a goal, but rather a process. Standards like the DDI need to be updated to support changing kinds of data, and changing needs of researchers and teachers.

To continue development of the DDI standard, the ICPSR and the Roper Center for Public Opinion Research have taken the lead in creating a self-supporting organization, the Alliance for the DDI or DDI Alliance. The Alliance Steering Committee includes officers of IASSIST and CESSDA, continuing the close association of the DDI to professional data organizations. The DDI has come a long way since the start in 1995. 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 decide on changes to the DDI.

Conclusion

The challenge of documentation is that it must serve multiple purposes. It must explain the details of studies including the instruments used, the sample, the response rate, and other relevant information. It should be friendly and accessible to novices, yet have the depth to meet the needs of experts. Since 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. Finally, it should support automatic generation of system files for popular statistical software. Some of these purposes work against each other; for example, the simplicity that is important for novices often conflicts with the depth desired by experts. The beauty of the DDI is that it is sufficiently flexible and adaptable to serve all these purposes and more.

The use of the DDI will produce a striking improvement in access to a vast number 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. Analysis of secondary data is a rich source of social science knowledge. It is enhanced by increasing the availability of high-quality data. Detailed, easily accessible documentation is necessary for this to become a reality. The DDI promises exactly that: by facilitating flexible, user-friendly documentation, improving its flow through networks, and enhancing our ability to link and display data in new ways, the research processes mediated by that documentation could improve. Every social scientist who uses secondary data can become more productive and create higher quality research. This benefits all of us.

Acknowledgements The authors have participated in DDI development, but the DDI was developed by people representing many academic institutes, governmental agencies, and other institutions. We are thankful to these people and the many participating organizations. We are particularly grateful to Mary Vardigan and Ann Green for conversations about the DDI and for their help in finding documents describing the DDI. More information is available at the DDI website (http://www.ddialliance.org/). Development of the DDI continues: initiatives can enjoy long lives when many people enthusiastically support the effort.

References

Adams D (1986) The Hitch Hiker's guide to the galaxy. Heinemann, London
Altman M, Andreev L, Diggory M, King G, Sone A, Verba S, Kiskis DL, Krot, M (2001) A digital library for the dissemination and replication of quantitative social science research. Soc Sci Comput Rev 19:458–470
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? Soc Sci Comput Rev 11:63–83
Blank G, Rasmussen KB (2004) The data documentation initiative: the value and significance of a worldwide standard. Soc Sci Comput Rev 22(3):307–318
CESSDA (1997) Council of European Social Science Data http://www.nsd.uib.no/cessda/IDC. Cited 05 Nov 2006
Checkland P, Holwell S (1998) Information, systems and information systems: making sense of the field. John Wiley & Sons
CPANDA (2006) Cultural Policy and the Arts National Data Archive http://www.cpanda.org. Cited 05 Nov 2006
Danish Committee on Scientific Dishonesty (2005) Executive Order No. 668 of 28 June 2005. (The official translation to English)
DDI (2005) Data documentation initiative. http://www.ddialliance.org. Cited 05 Nov 2006
Dodd SA (1982) Cataloging machine-readable data files. Chicago, American Library Association
Fienberg SE (1994) Sharing statistical data in the biomedical and health sciences: ethical, institutional, legal and professional dimensions. Annu Rev Public Health, vol 15 Annual Reviews, Inc., Palo Alto, CA
Fienberg SE, Martin ME, Straf ML (eds) (1985) Sharing research data. National Academy Press, Washington, DC
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. http://www.clir.org/pubs/reports/reports.html. Cited 05 Nov 2006
Hauser RM (1987) Sharing data: it's time for ASA journals to follow the folkways of a scientific sociology. Am Sociol Rev 52(6):vi–viii
MADIERA (2005) The MADIERA Project. http://www.madiera.net/. Cited 05 Nov 2006
Meier KJ (1995) Replication: a view from the streets. PS: Political Science and Politics, XXVIII(3):453–459
NESSTAR (2004) Publish your data on the web with Nesstar 3.0 http://www.nesstar.com. Cited 05 Nov 2006
NHGIS (2004) National Historical Geographical Information Center. http://www.nhgis.org. Cited 05 Nov 2006
Nielsen P (1974) Report on standardization of study description schemes and classification of indicators. DDA, Copenhagen (intern)
NISO (2001) Understanding metadata. PDF at http://www.niso.org/standards/std_resources.html. Cited 05 Nov 2006
Rasmussen KB (2000) Datadokumentation. Metadata for samfundsvidenskabelige undersøgelser. [Data documentation: metadata for social science research] Universitetsforlag, Odense. (In Danish)
Rasmussen KB (1996) 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. May
Rasmussen KB (1989) Data on data. In: Proceedings of the SAS European Users Group International Conference 1989. SAS Institute, Cary, NC, pp 369–379
Rasmussen KB (1981) Proposed standard study description. The SD as a basis for on-line inventories of social science data. (DOC00227) DDA, Odense (intern)
Ryssevik J, Musgrave S (2001) The social science dream machine. Soc Sci Comput Rev 19:163–174
Sieber JE (1991) Introduction: sharing social science data. In: Sieber JE (ed) Sharing social science data: advantages and challenges. Sage Publications, Newberry Park, CA, pp 1–18
Virtual Data Center (2004–2005) http://thedata.org/. Cited 05 Nov 2006