Social Science Metadata and the Foundations of the DDI

by Karsten Boye Rasmussen[1]

Abstract

The work and life of Sue A. Dodd had influence in its own right and her work was adapted and incorporated by others just as her work was influenced by others and part of a general evolution of social science metadata. From her focus on the catalogue description of machine-readable data that made users able to reference and identify data files as a research source, the description of social science data files gained further momentum. This paper centers on the fundamentals of social science data and their relation to metadata. There are levels of metadata in typical social science where the study, the variables of the study and the codes of the variables define a hierarchy. For each level there are many potential descriptive items that can be part of the full metadata. The work was initiated in the US but there was also work carried out in Europe through the described period, mostly centered around 1975-1995. All of this can be considered the foundation of social science data metadata description that later evolved to become the work carried out within the Data Documentation Initiative (DDI).

Keywords: DDI, Data Documentation Initiative, metadata, social science, study description, codebook.

Introduction

Sue A. Dodd identified the vacuum of library catalogue description for the new and growing area of machine-readable data (Dodd, 1979), and provided guidance for using a standard bibliographic format to fill this vacuum (Dodd, 1982). In following years she continued to communicate, discuss and elaborate upon the guidance. Some of the other papers in this collection in the IASSIST Quarterly (IQ) will bring more focus to the writings of Sue Dodd and some papers will elaborate on the work carried out within the Data Documentation Initiative (DDI). This paper sees the influence of Sue Dodd as her work was adapted and incorporated, highlighting some of the European work on the study description for social science data during the period 1975 to 1995.

The term 'machine-readable data' for the materials being described was later to be renamed 'computer files' which comprises more materials than the data file that is the main object of this paper. This paper will bring descriptions and discussions of why and how the formats, processes, and technology developed in collaboration and hopefully demonstrate how this totality articulated the need for a documentation standard and formed the basis for a solution of elements for the Data Documentation Initiative.

For a tour of the developments in data archiving combined with the technical and political – as well as personal - developments I'll recommend 'The Decades of My Life' by Judith Rowe (1999).

The digital age: Useful data are useless without documentation

The digital age signifies that everything is stored as numbers. Obviously when only a number is communicated – '42' is my favorite example - it cannot alone carry any meaning for the recipient. A number has to be wrapped in explanation to convey any meaning. We know we have been fooled - and we find the nonsense amusing - when the computer Deep Thought after seventy-five-thousand human generations of calculating produces the answer 'Forty-Two' to 'the great question of Life, the Universe and Everything' (Adams, 1986, p. 128).

Data, information, knowledge

The machine-readable data file consists of nearly endless series of numbers. Let us visualize the data file as a database table (survey file of a questionnaire) consisting of many attributes (variables as columns) concerning entities (individuals as rows).

Many information scientists have attacked the problem of having more precise concepts for categories or levels of information. In everyday life we might use the term "'data' intermingled with the term 'information' without much bother. I must admit that I personally don't have any problem with the less rigid use of the concepts. However, there are insights to be gained when entering a more meticulous definition of the terms. When working within systems analysis in Britain the researcher Peter Checkland proposed some useful distinctions for his methodology of soft systems development. Data are viewed as an unordered, formless, disorganized pool of facts. (Checkland actually used the term 'cloud' that now carries the meaning of organized access to safe storage always available.) 'Information' is facts situated in context resulting in 'meaningful facts'. In relation to the data file the documentation is delivering the context. 'Knowledge' is 'larger, longer living structures of meaningful facts'. In our analogy this relates to the use of the data file mostly exemplified as the relationship between variables. In getting to knowledge the first issue is to select the individuals and the variables for our data collection. Thus we select the data relevant to us. Checkland proposed the term 'capta' for the selected data (Checkland and Holwell, 1998, p. 90) to be distinguished from the disorganized pool of data. The context in the form of the description of the selection process - concerning both the selection of entities and the selection of attributes - is a necessity for creation of 'information' from 'capta'.

These descriptive entries are the metadata or data description. I did warn you that I might not be rigid and consistent in my use of these concepts. The term 'capta' is in my view a way to understand what we normally call 'data'. When we have data we require data documentation in order to produce information and knowledge.

Multiple benefits of data documentation

During the 1960s and 1970s the research data file became a regular resource available to other users for secondary analysis and the benefits of data sharing and data documentation was addressed consistently in the 1980s (Rasmussen and Grant, 2007, p. 60). A conference in 1979 resulted later in a comprehensive report on 'Sharing Research Data' supported by the National Research Council (USA) with extensive discussions and papers and a leading chapter of 'Issues and recommendations'. The number one recommendation is 'Sharing data should be a regular practice' (Fienberg et al., 1985, p. 25). In the journal American Sociological Review the benefits of data sharing published in the report was summarized as:

> 'reinforcement of open scientific inquiry; the verification, refutation, or refinement of original results; the promotion of new research through existing data; encouragement of more appropriate use of empirical data in policy formulation and evaluation; improvement of measurement and data collection methods; development of theoretical knowledge and knowledge of analytical techniques; encouragement of multiple perspectives; provision of resources for training in research; and protection against faulty data' (Hauser, 1987).

The benefits of sharing of machine-readable data are parallel to the benefit of having libraries sharing human-readable material. Making the sharing possible implies the benefits of metadata describing the data file. The value of sharing data can be accredited to several dimensions of arguments:

Actors

Sharing data primarily implies the use of the data as secondary data when data are reused for other than the intended purposes by other researchers or by students for educational purposes. Data archives have also often experienced the original investigator(s) among the requestors for their own dataset because the archives had not only preserved but also value-added to the dataset by elaborated and accessible metadata.

Resources

Collecting data is a money intensive action. Research data collection is often financed by public money. Sharing the data is the most cost-efficient way to carry out science. Furthermore, some retrospect research initiatives are only possible to realize through secondary analysis often of an extensive kind where several earlier data sources are being used in combination to produce a more accurate account.

Naturally there are also costs involved in sharing data and producing useful metadata. It is possible to argue that not all collected data are worth the extra cost. However, when research projects are financed in competition (e.g. from types of science funds) it would be counter-intuitive if the board would not regard the future collected data as sufficiently valuable. The National Science Foundation (USA), the Economic and Social Research Council (UK) and the Danish SSF (Social Science Foundation) - and probably many more like these funding agencies - all had clauses in the contracts for archiving and documenting research data for reuse when I investigated this in the late 1990s (Rasmussen, 2000, p. 169).

Controls

A popular issue of research data being publicly available is an issue of being able to control that science is not infected with fraud. However, it is also mentioned in the citation above that the public through data access can control administration and evaluation of policies. Furthermore the sharing of data also supports 'multiple perspectives'. This is regarded as fundamentals of having a democratic and free science.

The short Hauser entry above reminds us that there is a distinction between survey data being 'public' and 'usable'. This is a distinction depending upon metadata. Hauser recommended that journals (e.g. the ASA) would be keeping the tabulations related to the published papers in the journals (this was more than 25 years ago and the recommendation was practical and proposed the technical solution of storage on floppy disks). The recommendation also brought an attention for a recommended format. However, the discussion on what to archive and in which formats was in the 1980s already a continuing discussion and system decisions were implemented in data archives around the world and will be addressed later in this paper.

The general improvement and development of research can also be placed under the dimension of control. Metadata description of a research dataset will act as an evaluation of the study, e.g. a survey is described in terms of the population, the selection

IASSIST Quarterly 2013 29

process, the nonresponse, the response rate etc. These items will act as a checklist for new and not so experienced researchers as well as demand consistency and thus easier access to the data.

Archives and archivist collaboration

Some twenty years before the comprehensive reports on sharing of research data (Fienberg et al., 1985) and of social science data (Sieber, 1990) the sharing of data had already been implemented in institutionalized form by the establishment of the ICPR (Inter-university Consortium for Political Research later ICPSR, Inter-university Consortium for Political and Social Research at the University of Michigan). Before that the Roper Center had been established as an archive for Gallup and other commercial public opinion polls since shortly after the end of World War II. With the eye on comparison of national statistics there were three (UNESCO-) conferences in the early 1960s also addressing data archives (Rowe, 1999).

The concept of data archiving was also institutionalized in Europe by the creation of several national social science research data archives from the mid 1960s and onward. In Germany the Zentralarchiv für Empirische Sozialforschung (later included in GESIS (Leibniz Institute for the Social Sciences)), in UK the UK Data Archive, in the Netherlands the Steinmetz archive (later included in DANS (Data Archiving and Networked Services)), in Norway the NSD, and in Denmark the Danish Data Archives. Many data archives incorporating data files from research in many other European countries followed quickly after.

When the IASSIST[2] was founded there were a sufficient number of individuals within the profession of data librarians, data archivists and other advocates, researchers and technicians who were sharing machine-readable data to form an organization that has lived and has had influence for a span of time that was in older days a lifetime. The acronym IASSIST was created in advance of the name: International Association for Social Science Information Services and Technology (Geda, 2006). Both the acronym and the name continue to capture very well the focus of the association. The sharing of knowledge was institutionalized internationally through the IASSIST conferences - and the included workshops - as well as through workshops and the meetings of official representatives (ORs) under the auspices of the ICPSR.

Soon after the realization of IASSIST, the European CESSDA (Consortium of European Social Science Data Archives) was formed as an umbrella organization for the national European data archives. Issues were trans-border sharing of data and privacy in the different countries. CESSDA was also successful in knowledge transfer between individuals through themed seminars many of which addressed the issue of metadata and discussions on the use of different standards.

Metadata and levels of documentation

The fundamental documentation of a data file as a whole is the identification of the data file as a research source. When we talk about the 'American National Election Study, 2004: Panel Study' others will be able to locate the study[3] (and they will find a shorter form of identification as 'ICPSR 4293'). Similar to citations from literary sources there should be a way to accurately identify the research data source. The documentation of the data source makes it possible for secondary users to give positive credit to the people behind the data source. Among the most fundamental aspects of scientific research is the possibility of inter-subjectivity that is the closest thing to the unattainable objectivity. Documentation has the potential to introduce discussions on the methods used and the research decisions. The researcher describes how the data were created and makes critical evaluation possible. Provided the documentation reveals valid methods, the replication of the procedures described in the documentation should ideally lead another investigator to the same result.

Data equals documentation - documentation equals data - metadata

For some of us the rectangular data file was - and is - the 'normal form' of social science data. The term 'normal form' brings us to relational databases and that is no coincidence. All kinds of complicated database structures are possible with the methodology of relational databases relating rectangular tables. Others may be interested in more exotic examples of data for social scientific analysis, such as artifacts like images and sound bites. Whatever type of data is to be analyzed using a systematic method, you have to define what your specific interest is and define and select your 'capta'.

In other words, the systematic and comprehensive documentation of a collection of machine-readable files can become data for a researcher investigating the collection. For instance a research objective could be to compare surveys carried out at different periods of time or at different geographical locations. The description of data is called metadata as it is 'data about data'. We should note that metadata are not only 'about'. Metadata are also genuine data that can be analyzed.

The levels of data documentation

The primary user of a rectangular social science data file needs information on the variables (columns) as well as explanations of the codes as exemplified in the information:

V6	Sex of respondent'	1. 'male'	2. 'female'

Figure 1. Codebook documentation of a variable with codes (a minimum example).

This codebook with information at the variable level and the code level is elementary yet helpful for the analyst who already is familiar with the overall background for the data file.

However, for the secondary analyst '(i)nformation about variables is useless unless the population, sample, and sampling procedures are described' (Blank and Rasmussen, 2004, p. 307). For this reason, many more precise items were introduced for study level description early in the history of data archiving and a standard for the study level description was agreed upon (Nielsen, 1974 and 1975). It was discussed, refined and presented at several meetings, workshops, and conferences in the 1970s. However, although there was this continued presence around the 'standard study description' the standard was primarily the foundation for local implementations and never achieved the status of an international 'de facto standard'. More important was the international agreement concerning the items described in the standard. Several of these items were also implemented in other archives – in other local formats and systems – and the items were later included in the development of the DDI-standard. They are listed below:

001-	General information documentation level, subject cluster, keywords
101-	Identification and acknowledgements bibliographic reference, archive, primary investigator(s) and other references
201-	Analysis conditions abstract, kind of data, data sources, type of unit, number of units, size of dataset, time dimensions, universe, selection, sampling, data collection instruments, weighting
301-	Reuse of data Data representation, data cleaning and controls, access conditions
401-	References to relevant publications primary publications, secondary publications, analysis results, references to other studies (data files)
500-	Background variables personal, (age, gender, ethnic group ...), household information, employment, occupation, income, education, mass media, ...

Figure 2. Overview of a selection of items included in the Study Description as used at the Danish Data Archives (summarized from Rasmussen, 2000, p. 287-352, and Rasmussen, 1981).

The items of the Study Description were included as material for the DDI development as explained further in the Green and Humphrey paper in this issue of the *IQ*.

Machine-actionable metadata

There was a great effect when documentation of machine-readable data had the positive experience caused by 'taking its own medicine.' Naturally the decision on which items to include into the documentation was a very important and first decision. However, when that was settled and the human could read the documentation there was a revolution in having a well-formed documentation not as typed pages but in machine-readable form.

The revolution happens when the machine-readable documentation becomes machine-actionable. When metadata collections were formed several data archives were looking into retrieval systems for supporting secondary researchers in their quest for finding appropriate data. In Europe the German archive Zentralarchiv in Cologne (now GESIS) was in the forefront of building retrieval systems. Another earlier use of rigidly structured metadata was the translation of OSIRIS codebooks into the - more reduced - control language used by other statistical packages like SPSS and SAS. As system files were dependent upon the configuration platform of both machine and operating system the character-based solution of exchanging old-fashioned card images in the form of lines of text was the most general solution. Furthermore, at the codebook level the machine-actionable documentation meant that a machine - e.g. software on a computer - would be able to read the documentation and be able to automatically interpret and perform calculations and controls on the data file. In addition, the machine-actionable documentation could, with great benefit, be created prior to the data file. The documentation could by a machine be the foundation for data collection, e.g. generating the screens and software for data-entry for telephone interviewing (CATI, computer aided telephone interviewing).

Turning the complete documentation - including the study level - into machine-readable form meant that studies could be searched effectively and with the Internet the accessibility to data files increased tremendously both regarding the number of studies as well as regarding the speed the users could access the selected information. Having documentation in machine-readable form implies that there should be a defined format - a standard.

Standards of data documentation

The old joke about standards goes: 'Standards must be good since there are so many to choose from.' In this section some of the actually used standards for documentation of social science research data before the DDI are briefly described.

Machine-readable Cataloguing - in short MARC - was developed at and institutionalized by the Library of Congress as a computerized 'library card' format to build a library catalogue that was machine-readable in contrast to the paper cards traditionally used in libraries. The latest MARC format (MARC 21) was finalized before the millennium. Some can consider the format as old-fashioned and related to outdated technology. However, the format is still very much in existence in libraries all over the world.

In regard to the legacy of Sue Dodd, the MARC format for MRDF (machine-readable data files) stands central because her 1982 manual provided guidance for applying the MARC/MRDF format to create standard catalog records for MRDF. Her guidance based the catalog record upon file or study-level documentation. Standards for documentation were discussed continuously throughout the 1975 to 1995 period. As mentioned earlier, there were during the 1970s international workshops in Europe on the documentation at the study level. IASSIST accommodated for many years sessions on documentation with presentations, proposals, and discussions at the international yearly conferences. Further information on the study level documentation standards standards including MARC and 'Dublin Core' is found in the Green and Humphrey paper in this *IQ*..

Machine-readable documentation in use

At the IASSIST conference in 1993 an action group for 'Codebook Documentation of Social Science Data' was formed. In 1994 I carried out an investigation by mail questionnaire in order to obtain an overview of the amount and kind of documentation that existed at archives. The investigation also obtained preferences of documentation among the professionals. This was reported as a presentation at the IASSIST conference in 1995 as well as in the *IQ* (Rasmussen, 1995). The paper presented a snapshot of the situation now 20 years ago. The majority of the studies were then without any machine-readable documentation. The reported machine-readable datasets were from 19 answering archives. The distribution of datasets by type of machine-readable documentation is shown in Table 1. The different machine-readable formats or level of standards are explained below demonstrating the development of metadata for social science.

1. Scanning	505
2. Text	1,943
3. Dict	4,520
4. Dict+	3,656
5. Dict + Codebook	2,003
Total	12,627

Table 1. Datasets with machine-readable documentation (Rasmussen, 1995).

Scanned images

A doable and less time-consuming method of archiving and being able to distribute information was in the form of scanned images. The available documentation often existed in the form of a questionnaire and as sheets of paper that could be scanned and saved as images (de Vries, 1992). As the processing often did not include OCR-processing this meant that the production did not deliver a searchable documentation nor was the information usable for input to the statistical packages so scanned images are considered the lowest form of machine-readable documentation. Actually you can contest that they as images are readable by a machine. The information is not structured and an image is not data nor is it metadata. A considerable effort was demanded from the secondary user as information had to be keyed-in in order to analyze the data. Sometimes the scanned images were additional to other formats and then the images were of convenience to the user and a safe storage solution to the archive. The investigation among individuals showed that most people would not be content with scanned images of a questionnaire for secondary use of a data file as they would likely prefer the structured information with the possibility of machine-action and the direct and accurate feed of the data file into a software package for analysis.

Text

This documentation level implies that machine-readable documentation existed as unstructured (i.e. untagged) text. This could for instance be in the form of output from OCR or a questionnaire kept in WordPerfect. The lack of structure implied that there was no easy solution for bringing the documentation into the analysis making it machine-actionable. However, the bulk of text could be searched just as we still often search within a text based file or in a collection of such files.

Dictionary

The availability of a dictionary implied that the structure of the data file was reflected in the proofed dictionary and data could be loaded into standard packages like SAS, SPSS or OSIRIS. This brought down the time involved in setting-up a system for the secondary analysis as well as increasing the accuracy. The data quality could be greatly deteriorated by the hand-to-hand passing on of instructions. With the availability of a machine-actionable dictionary the secondary analyst would spend less time on controls and more on analysis.

At this level - see below the higher level 'dictionary-plus' - only column information existed and often in a very restricted form. For instance variable labels would often only be able to carry 24 characters of documentation which people did not find sufficient. Furthermore, this format did not include any information about the codes and categories. The user would for instance need scanned images in order to find explanation for the codes actually encountered in a variable.

Dictionary-plus

For the category of 'dict+' the plus indicates that the documentation comprises category labels in OSIRIS or in the form of 'value labels' in SPSS or 'user formats' in SAS. Often the documentation was stored in a system proprietary format and often the packages were forcing a 'lock-in' on the users so you could not for example directly analyze an SPSS system file with the SAS package. The situation has improved but at that time even the change of version of system files within SPSS presented a problem with backward compatibility. Having documentation embedded in system files also presented the archives with a load of migration tasks as information could be lost if files were left in old system formats just as they would be lost if they were left on old media. As Rothenberg phrased it: 'digital information lasts forever - or five years, whichever comes first!' (1995).

Dictionary plus codebook

Through the period in focus here (1975-1995) the SAS and SPSS packages were the popular tools for analysis although they had no support for documentation at the study level apart from a filename and a short title for the study. The amount of study level documentation was very similar to the restricted documentation of variables and categories: a name or value and a label of limited length (Grant, 1993; Rasmussen, 1989).

The OSIRIS documentation format was early regarded outdated by being tied to a card-image format of 80-characters - an inheritance from physical punched Hollerith cards. When using physical cards placed in sequence it was very important to be able to reconstruct the sequence (with a counter-sorter) in case a stack of cards got dropped on the floor or was mixed into another stack.

The OSIRIS codebook layout was originally mostly a format for storing the electronic information for later printing a more nicely formed codebook. However, the OSIRIS format was remarkable by being able to store unlimited amounts of text because a description of a variable could occupy multiple lines; this was also similar for comments, explanations, and code description. In OSIRIS different types of documentation were identified through an alphabetical "tag" in the first field or column of each card. (See figure 3).

Remarkably, the OSIRIS format included features for description at the study level (S-cards). OSIRIS was developed over a period of time and I'm here referring to the OSIRIS III (ICPSR, 1973). Structure of the description at the study level was available as the format even included further distinctions on the study level for entries as title, original archive, etc. Furthermore some 'meta-metadata' were possible as the general structural principles of the codebook-layout could be described within the standard itself using a meta explanation type card (E-cards). These features and other parts of the OSIRIS format were extended at the Danish Data Archives and the German Zentralarchiv (Rasmussen, 2000, p. 351). These were extensions for more elaborate description, data checking, and retrieval. The harsh backside was that the OSIRIS system in itself in the analysis tasks – including tabulations - totally ignored all available information apart from the limited dictionary information supplying variables with a number and a 24-character label plus some format information as well as information on missing data.

S	study level
E	meta explanation
T	dictionary, sub-structured including missing data and format
Q	variable description. question in questionnaires
K	continuation
X	explanation
C	code value and label, sub-structured
B	for grouping of more Cs (higher level categories)
F	frequencies, attached to the C
J	temporary comments
G	note number
M	note text

Figure 3. The original card-types of the OSIRIS-codebook (summarized from Rasmussen, 2000, p. 340-352; originally in ICPSR, 1973).

Although the OSIRIS format was old-fashioned judged by the card-image layout, it was thus also very farsighted. The original OSIRIS format and the extensions - including some developed outside of ICPSR - had the capability to include a very high degree of the relevant documentation compared to the offerings by other statistical packages like SPSS and SAS. The documentation types brought attention to the levels and structure of documentation and to the comprehensive items that were later developed into the DDI. The letter in the first column of the lines of OSIRIS documentation was a very early markup of documentation through the 'tagging' by first-column letter. Furthermore, the rigid card-image and fixed-columns OSIRIS format was kept as the archival format, but it was generated from freely typed input with software generating the variable numbering and the required different indentions based on the tagging for the relevant card-type ('Q', 'X', 'C' etc.)

Towards a standard

Some months after the formation of the IASSIST action group on codebooks a CESSDA seminar was held in Gothenburg in August 1993 on 'Variable Level Documentation.' The following year another CESSDA seminar on 'Networking and Internet' was held in Grenoble. Both of these workshops also had non-European participation, most significantly the participation by staff of the ICPSR.

The discussions not only collected the sum of identified elements that from the viewpoint of archives were considered important to include in a standard documentation package but also brought to attention the many functions that the documentation should support. They also introduced carefulness towards what could be termed independence. This independence was the guarantee that a standard could evolve and not be locked, as well as being available to all.

Functions of the documentation

This paper has mentioned how documentation first of all delivered the printed study description and the codebook in a well-formatted human-readable form. This is believed to continue to be a relevant use though the human might read the information from another device than paper. Another important function is that collections of documentation – especially in the form of well-structured computer files – are searchable. It has also been mentioned that documentation can deliver input for the validation of data previously collected and also control data being collected. Lastly, the ultimate use of documentation is in the analysis and the presentations from statistical software of the documented data.

But do users really need all the bells and whistles delivered by the DDI? When huge commercial companies like SPSS (now IBM SPSS) and SAS deliver only a minimum of documentation facilities for a dataset should that not be taken as a sign of what the user community is interested in – and as a sign of what quality level of documentation the community is willing to pay for? OSIRIS is still in existence and can be found as MircOsiris for MS Windows. However, this micro-version does not support the unlimited and elaborate OSIRIS codebook format. MircOsiris has only the minimum documentation facilities as described in Figure 1 above.

I believe that minimum documentation presents a one-eyed view with a narrow focus on machinery for analysis of your own data. The limited documentation will prove to be a problem for even the primary researcher if and when an older dataset is to be re-analyzed. Naturally the problem will be even greater for the secondary analyst. In commercial settings they know within data management that elaborate - and often very expensive - documentation and management systems are necessary for gaining profit of the data warehouse.

Independent documentation

When developing a new documentation format as set forth by the Data Documentation Initiative it was considered very important that the standard be independent of commercial interests. Public archives and university libraries would not be able to afford to tie themselves to a storage format that could imply a yearly user fee. Independence was also the term used in connection to which systems should be allowed to analyze data described in the DDI-format so there should be no licensing. The DDI-format should also be independent of operating system platforms. It might be possible to obtain financial support for the development if a company could obtain rights for a proprietary format and system. However, data archives regularly service many users who have distinct preferences for this or that system. Therefore the solution should be that the documentation format is open for use by all software developers or vendors. The archives already have great expertise in converting existing codebook/dictionary documentation to the reduced description used by SAS and SPSS. When developers at archives would use the DDI-format the software for conversion to users' preferred software would naturally follow.

Along came the Internet with a markup language

The Internet was early on seen as being of main interest to archives as a media for users searching for data and thus identifying relevant potential datasets. Later the Internet also became a media for direct deliverance of data. And later again the client could get thinner and direct analysis would be performed on the network servers. The Internet was as such very promising for much faster and easier identification and access to data sources as well as much cheaper distribution and easier analysis.

The use of the Internet was accelerated with the popularity of the World Wide Web. For a coming standard of data documentation the display of documents through the use of HTML (Hyper Text Markup Language) was very stimulating. Further stimulation came from the Text Encoding Initiative (TEI) that used SGML (Standard General Markup Language) for marking up documents. Consequently the proper move would be to make a document type definition (DTD) of data documentation using SGML. Early on some were experimenting with HTML for their documentation but realized that if HTML was to become the standard they would tie the documentation to the presentation and thus commit an offence against the general principle of independence. Quite similar to the reduction by software of documentation to SAS or SPSS format - or any other format - it would be easy to reduce a complete standard data documentation like the DDI to an HTML page or to several selected forms of preferred HTML presentations.

Another technological development from the Internet further paved the way for an effective solution for data documentation. The introduction of XML (Extensive Markup Language) implied easier and more flexible work than SGML and with a strong connection to the Internet. The slogan from Jon Bosak from Sun Microsystems - one of the founders of XML – was that 'XML gives Java something to do' (1997). Internet, Java, XML – all worked together.

Conclusion

Hopefully this paper has demonstrated that the origins of the DDI evolved from work related to social science data documentation issues by several institutions and people in the decades before the emergence of the Data Documentation Initiative. During the period described in this paper several principles and levels of documentation were identified and further refined. Re-using data is a community effort and the community encases the world. This was getting more and more attention with the fast expansion of the Internet, with the spread of the World Wide Web. Local inventions were put together and further developed in a distributed effort. That further story is another paper!

Actually the development of the DDI is discussed in several papers in this special issue of the IQ. Ann Green and Chuck Humphrey describe the early years and Mary Vardigan offers a paper on the later years of development as well as a very useful DDI-timeline

References

Adams, Douglas (1986) The Hitch Hiker's Guide to the Galaxy. London, Heinemann.

Blank, Grant (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(1): 63-83.

Blank, Grant and Rasmussen, Karsten Boye (2004) The Data Documentation Initiative. The Value and Significance of a Worldwide Standard. Social Science Computer Review 22(3): 307-318.

Bosak, Jon (1997) XML, Java, and the future of the Web. (<http://www.xml.com/pub/a/w3j/s3.bosak.html>).

Checkland, Peter and Holwell, Sue (1998) Information, systems and information systems: making sense of the field. John Wiley & Sons.

de Vries, Repke and van der Meer, Cor (1992) Exchange of scanned documentation between social scientists and data archives: establishing an image file format and method of transfer. IASSIST Quarterly 16(1-2): 18-22.

Dodd, Sue A. (1979) Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines. Journal of the American Society for Information Science, March 1979.

Dodd, Sue A. (1982) Cataloging Machine-Readable Data Files. An Interpretive Manual. American Library Association, Chicago.

Fienberg, Stephen E.; Martin, Margaret E. and Straf, Miron L. (eds.) (1985) Sharing research data. National Academy Press, Washington, DC.

Geda, Carolyn (2006) Recollections of the Formative Years of IASSIST. IASSIST Quarterly 30(3): 15-18. (<http://iassistdata.org/downloads/iqvol303geda.pdf>).

Green, Ann and Humphrey, Chuck (2013) Building the DDI. IASSIST Quarterly 37.

ICPSR ISR (1973) OSIRIS III Volume 1: system and program description. University of Michigan.

Hauser, Robert M. (1987) Sharing data: it's time for ASA journals to follow the folkways of a scientific sociology. American Sociological Review 52(6): vi–viii.

Nielsen, Per (1974) Report on Standardization of Study Description Schemes and Classification of Indicators. Copenhagen, Danish Data Archives.

Nielsen, Per (1975) Study Description Guide and Scheme. Copenhagen, Danish Data Archives.

Rasmussen, Karsten Boye (1981) Proposed Standard Study Description. The SD as a basis for On-Line Inventories of Social Science Data. Odense, Danish Data Archives.

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

Rasmussen, Karsten Boye (1995) Documentation - what we have and what we want: Report of an enquete of data archives and their staff. IASSIST Quarterly 19(1): 22-35. (<http://iassistdata.org/downloads/iqvol191rasmussen.pdf>).

Rasmussen, Karsten Boye (2000) Datadokumentation. Metadata for samfundsvidenskabelige undersøgelser (Data documentation: metadata for social science research). Odense Universitetsforlag, Odense, Denmark.

Rasmussen, Karsten Boye and Blank, Grant (2007) The data documentation initiative: a preservation standard for research. Archival Science 7(1): 55-71.

Rothenberg, Jeff (1995) Ensuring the Longevity of Digital Documents. Scientific American. January.

Rowe, Judith (1999) The Decades of My Life, IASSIST Quarterly 23(1). (<http://iassistdata.org/downloads/iqvol231rowe.pdf>).

Sieber, Joan E. (1991) Introduction: sharing social science data. In: Sieber J.E. (ed) Sharing social science data: advantages and challenges. Sage Publications, Newberry Park, CA, 1–18.

Vardigan, Mary (2013) The DDI Matures: 1997 to the Present and 'Timeline'. IASSIST Quarterly 37.

Vardigan, Mary (2013) Timeline. IASSIST Quarterly 37.

NOTES

1. Karsten Boye Rasmussen is an associate professor of IT and organization and a data scientist at the University of Southern Denmark. He worked in data archiving from 1974-1998. Email: kbr@sam.sdu.dk.
2. <www.iassistdata.org>
3. <https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4293>

Keywords from Vol 37 N0. 1-4. Courtesy Tagxedo.com