Historical Methods: A Journal of Quantitative and Interdisciplinary History
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vhim20

Implementing the Data Documentation Initiative at the Minnesota Population Center

William Block [a] & Wendy Thomas [a]
[a] Minnesota Population Center , University of Minnesota
Published online: 30 Mar 2010.

To cite this article: William Block & Wendy Thomas (2003) Implementing the Data Documentation Initiative at the Minnesota Population Center, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36:2, 97-101, DOI: 10.1080/01615440309601219

To link to this article: http://dx.doi.org/10.1080/01615440309601219

Implementing the Data Documentation Initiative at the Minnesota Population Center

WILLIAM BLOCK
WENDY THOMAS
Minnesota Population Center
University of Minnesota

Abstract. The Data Documentation Initiative (DDI) is an emerging international specification for documentation of social science data. Designed as an archival standard to help preserve access to data and codebooks, the DDI is nonproprietary and hardware independent. By treating documentation as metadata—or data about data—the DDI will open the door to the development of general-purpose automated software tools for accessing both data and documentation. The authors provide a background and history of the DDI and discuss the advantages of machine-processable documentation. They also describe several specific applications of DDI-compliant metadata currently under way at the Minnesota Population Center (MPC).

Keywords: Data Documentation Initiative (DDI), documentation, eXtensible Markup Language (XML), metadata, social science data archive

The Data Documentation Initiative (DDI) is an emerging international specification for documentation of social science data. Built on the underlying technology known as the eXtensible Markup Language (XML), the DDI is a nonproprietary, hardware-independent, neutral documentation specification standard. The DDI offers the promise of machine-processable documentation. By treating documentation as metadata—or data about data—the DDI will open the door to the development of general-purpose automated software tools for accessing both data and documentation.

Three of the most recently funded major projects under way at the Minnesota Population Center (MPC)—the National Historical Geographic Information System (NHGIS), the North Atlantic Population Project (NAPP), and the redesign of the Integrated Public Use Microdata Series (IPUMS Redesign)—were designed with the DDI in mind (see articles by Ruggles on pp. 9–19, Fitch and Ruggles on pp. 41–51 in Part One of this issue, and articles by Roberts et al. on pp. 80–88 and 89–96 in Part Two of this issue). Other MPC projects, such as the international data sets produced as part of IPUMS-International, predate the first formal DDI metadata specification and will be retrofitted with DDI-compliant documentation. This article provides a background and history of the DDI, discusses its advantages, describes our experience at MPC with DDI, and concludes with a look toward the future.

Background

The effort to create an international social science codebook standard dates to 1994, when the Inter-university Consortium for Political and Social Research (ICPSR) formed a working group that brought together representatives from various social science data archives, libraries, and production centers in the United States, Canada, and Europe. The original goal was to replace an obsolete codebook and dictionary standard known as OSIRIS with a modern, technologically capable format. At that time, the need for a well-structured, compatible electronic standard for social science documentation was clear. From the 1960s to the 1990s, each social science archive and data producer developed its own way of carrying out these tasks that reflected distinct variations in missions, budgets, technological environments, personnel, and constituencies. The result of this diversity was the fact that archival computer and catalog systems could not speak to one other. With the support of the National Science Foundation, ICPSR formed the DDI Committee to develop a new archival standard.

The committee quickly recognized that machine-processable metadata could offer more than a compatible standard. Data archives could use the codebooks to provide online search capabilities of their holdings or even searches across multiple archives. DDI-compliant metadata could automatically generate data definition files for common statistical packages. Ultimately, software tools could be developed that would locate data and metadata on the Internet, interpret it, and carry out data analysis online.

After considering various technologies, the DDI Committee—with thousands of hours of volunteer effort—published Version 1.0 of an XML Document Type Definition (DTD)

for social science codebooks in March 2000. XML is an electronic publishing and data interchange format developed by the World Wide Web Consortium (W3C). A DTD sets the rules by which a particular type of XML document must be created. Thus the DDI DTD establishes the parameters for writing an XML version of a social science data codebook. It carefully defines both the content and logical structure for the metadata that describe social science data sets.

As shown in table 1, the DDI DTD describes everything that is normally found in a social science codebook, including a description of the data collection, the data files, the variables in those data files, and complete bibliographic and citation information.

The DDI DTD requires that all the information in table 1 be documented in discrete and specific ways. Information about a specific variable, for example, is not presented as a block of text as it would be in a traditional codebook. Instead, a variable in a DDI codebook can be described by additional elements arranged hierarchically beneath the "<var>" element. Examples include the universe statement and the location of a variable in a data file. Additionally, many of these elements are further defined through the use of "attributes." The "<location>" element, for example, consists only of attributes that define location information about a variable. Examples include "StartPos," "EndPos," and "width," which document the starting column, ending column, and the width of a variable in a data set. Altogether, a single variable in a DDI codebook can be described by as many as 50 distinct elements and over 200 attributes, all of which nest beneath the "<var>" element.

The DDI Tag Library fully documents all the elements and attributes of the DDI DTD. The tag library provides clear English-language descriptions of each element, along with clarifying remarks and helpful examples. It also provides an outline of the DTD structure that graphically explains the hierarchical relationship between and among the many elements.[1]

The DDI DTD also defines the *logical structure* for each discretely coded piece of information. This logical structure allows the DDI to take advantage of similarities among groups of variables. For example, work questions on the census are asked only of people of a certain age, so the same universe statement can be applied to the group of work variables. Similarly, groups of variables might have the same valid range of responses or the same code for missing data; the DDI elements can describe whole groups of variables or one specific variable with a unique code.

The combination of discretely marked up pieces of information and the logical relationship between them gives the DDI its power. It is not merely an accompanying document that tells a researcher about a data set; a DDI codebook is itself computable. Indeed, with the

> achievements of the DDI, codebooks can now be created in a uniform, highly structured format that is easily and precisely searchable on the Web, that lends itself well to simultaneous use of multiple data sets, and that will significantly improve the content and usability of metadata. Further, this specification may have far-reaching implications for improvement of the entire process of data collection, data dissemination, and data analysis. (DDI 2002)

The machine-understandable structure of the DDI allows for automated processing by data access software. By forming a computable superset of items fully describing the content and intellectual order of a social science codebook, the DDI provides the infrastructure needed for computers to fully exploit social data and metadata. Users of social science data will be better able to both find and use data sets efficiently and effectively; indeed, the DDI represents a step toward what many are calling the Semantic Web (Berners-Lee, Hendler, and Lassila 2001).

DDI at the Minnesota Population Center

Aggregate Data

The Minnesota Population Center (MPC) has begun implementing the DDI specification for our aggregate data and boundary files project, the National Historical Geographic Information System (NHGIS). Unlike some of the other MPC data projects, the NHGIS is being built from the

TABLE 1. Main Sections of Data Documentation Initiative (DDI) Data Type Definition (DTD)

Section	Description	Content
1	Document	Items describing marked-up document itself as well as source documents (citation, title, etc.)
2	Study	Items describing data collection (title, citation, methodology, study scope, data access, etc.)
3	Data files	Items relating to format, size, and structure of data files
4	Variables	Items relating to variables in data collection
5	Other information	Study-related material not included in other sections (bibliography, separate questionnaire, file, etc.)

outset on a framework of DDI-compliant metadata. The most current production version of the DDI specification (Version 1.2.2, published 8 August 2002), however, does not contain the elements that make it possible to describe aggregate data. Aggregate data, usually represented in table form, are different from microdata, which are typically represented as one case per line. Consequently, the information necessary to make aggregate data machine processable is different from that used for microdata. The extensions to enable markup of aggregate data are currently embodied in the development version of the DDI specification, Version 1.3, and are undergoing beta testing.[2]

The DDI geography working group is discussing a second extension to the DDI that is also important to NHGIS. Aggregate data files frequently use geographic areas (such as blocks, tracts, counties, and metropolitan areas) as the basis for aggregation. The "footprints" or boundaries of these geographic areas often change over time without changing their coding. For example, the Federal Information Processing Standards (FIPS) code for the Minneapolis/St. Paul metropolitan area is 5120; this code has remained unchanged for decades, even though the geographic footprint of this metropolitan area has expanded from a 7-county area to a 13-county area in less than 30 years. NHGIS needs the ability to easily link data from the aggregate files to the boundary files that provide footprints to make sure the user receives data for the correct geographic area. Members of the DDI geography working group are addressing this problem and others created by the need to describe geographic data within the DDI.

Markup of NHGIS aggregate data is proceeding through a combination of machine markup and manual markup. For example, the Summary Tape File 4 for the 1990 census (STF4) is over 150 gigabytes and is accompanied by machine-readable documentation from the Census Bureau. The large size of STF4 makes it extremely costly to mark up by hand. We therefore opted to write software to automatically wrap DDI XML tags around the majority of information in STF4. Researchers using a standard file editor cleaned the small amount of information that did not sort properly into elements. To date, we have completed the automatic markup of Summary Tape Files 1-4 for 1990, as well as the 1990 Equal Opportunity Employee file and PL194-171 (the redistricting file).

It is difficult to mark up aggregate data into the DDI specification. To maintain a high level of quality in our markup, we have developed a number of software tools to check each of our XML files. The tools are used to confirm that the file conforms to the latest DDI specification and to proof the data contained in the XML file. One of the most common problems with aggregate data is properly defining and describing multidimensional tables. We have created several tools to check the internal consistency of the DDI metadata.

To explain one of these tools, consider a table that reports county populations by race and sex. This table has two dimensions—race and sex. Each of the dimensions is broken into multiple categories; race has five (American Indian or Alaskan Native; Asian; black or African American; Native Hawaiian or Other Pacific Islander; and white) and sex has two (male and female). For each county, the table will therefore include 10 data cells. The DDI element <nCube> contains attributes about the table, including the total number of dimensions (dmnsQnty) and the total number of data cells (cellQnty). The DDI element <dmns> describes or references a previous description of each dimension that includes the number of categories and their labels. One of our tools ensures that the number of <dmns> elements marked up for each table matches the number of dimensions specified by dmnsQnty; in the example above, dmnsQnty should equal two. Another tool checks that the total number of data cells reported in the cellQnty attribute is equal to the product of the number of categories of each dimension. In the above example, cellQnty would equal 10 (the five categories of race multiplied by the two categories of sex). In addition to the internal-consistency checks for the metadata, we are also creating tools to verify that the metadata accurately reflect the physical structure of the data.

The NHGIS data access system is being built to help users discover and access information within the vast amount of aggregate data contained in NHGIS. At the heart of this system is an ontology that reflects the many terms, concepts, and relationships used in aggregate U.S. census data across time. We are building much of this ontology from the marked-up version of the detailed Census Bureau technical documentation that accompanies each new aggregate data release. This documentation has been marked up in DDI format and forms the basis of the information fed to the ontology. We are creating a computable model of the Census Bureau's aggregate data technical documentation that will help guide users to the data they are seeking (Wozniak 2001). For example, without having to scour through reams of online documentation, researchers looking for general information on disabilities in the African American population will find data on "maimed" blacks and mulattos in 1880 and on work disabilities among blacks in 1980 and 1990. The ontology makes this search feasible because it knows that each of the disability variables is a function of health and that the descriptors "mulatto," "negro," and "black" have all been used to describe Americans of African descent.

Microdata

In addition to our work on the NHGIS, we are also working to build DDI metadata into our microdata projects. Much of the DDI work on these projects, at least initially, will involve retrofitting existing documentation to be compliant with the DDI specification. Fortunately, we can automate much of this work because the bulk of our existing documentation is already in the form of highly structured HTML. We have converted most of our IPUMS documentation—

which consists of nearly 3,000 pages of individually maintained static HTML files—to a rough DDI form. The variable descriptions in the IPUMS follow a standard structural format—each variable has column location information, variable name and label, availability across census years, universe statements, and descriptions—so we began with an initial parsing of the information into the DDI. At that point, we discovered that the IPUMS documentation was not as consistently broken into as many distinct pieces as the DDI required. In some instances, the IPUMS documentation combines the variable description and the discussion of temporal comparability issues; in other cases, these two categories of information are separate. To deal with such inconsistencies, we created a simple editing interface for research assistants to cut and paste entries into the appropriate position in the DDI.

Bringing a high level of consistency to the IPUMS documentation via the DDI will pay dividends in a number of ways. The IPUMS comprises thousands of static Web pages and an increasing number of dynamic documents that are created on the fly, and this system has become unwieldy to maintain and upgrade. In addition, the continuous process of correcting and updating the data and documentation has created serious version-control issues for the documentation. When a variable is altered in the current system, for example, changes must be made in at least eight different places: three data-definition files (for SAS, Stata, and SPSS), three tables used to build pages for the documentation and data extraction systems, and at least two static HTML documentation pages. Any discrepancies among these files can lead to system failure or user confusion.

The DDI will reduce the costs of system maintenance and decrease the potential for documentation errors. As part of the IPUMS Redesign, we are modifying the IPUMS data and documentation access system so that it is driven by DDI-compliant metadata. Once the new system is in place, we will be able to modify a variable by changing its specifications in a single location. The software will then propagate that change throughout the system. This approach will increase the flexibility of the IPUMS, reduce its maintenance costs, and greatly simplify the incorporation of new data files into the system.

Other MPC microdata projects are currently creating a large amount of new documentation. (On these projects, the MPC is still experimenting with the best way to create DDI documentation.) We are currently evaluating various XML editing software packages and considering the creation of customized data-entry interfaces for certain metadata creation tasks. For routine documentation creation and markup, a customized data-entry screen may yield efficiencies that far exceed the cost of programming a simple interface.

Demographic Data Cooperative

The DDI provides the underlying framework for the Demographic Data Cooperative—a collaboration of the MPC, the University of Michigan Population Studies Center (PSC), and ICPSR. The DDI allows us to take advantage of our collective holdings and knowledge, reduce redundancies in effort, provide a formal mechanism for sharing information about data, and capitalize on advances in Web-based resource discovery and dissemination. We are creating a distributed data archive, with human and electronic resources located at all participating organizations. Data access software designed to work with DDI-based machine-understandable metadata will reduce the need for human intervention in the distribution of data to researchers. These innovations will simplify access to data and documentation, allowing data archivists at each center to focus on local support.

The initial holdings of the Demographic Data Cooperative are a combination of the data resources of the PSC, MPC, and ICPSR. Although the strong and complementary data holdings of the current members provide a rich starting nucleus, we hope other institutions will join and augment the collection. An electronic catalog of data co-op resources will soon be published on the Data Co-op Web site (http://www.popdata.org).

Future Directions for the DDI

The DDI has significant support from many of the world's leading social science data archives and statistical agencies, and current members of the DDI Committee include representatives from these leading North American and European institutions. (See table 2 for a complete current list of DDI member institutions and beta testers.) Archives of marked-up documentation are currently growing. Both ICPSR and the Council of European Social Science Data Archives (CESSDA) are converting their data catalogs to DDI specification. Several institutions are working on tools using the DDI. Networked Social Science Tools and Resources (NESSTAR) belongs to a joint project involving members of the U.K. Data Archive, the Norwegian Social Science Data Services, and the Danish Data Archive. Programmers are building software based on the DDI specification at the Virtual Data Center of Harvard-MIT, Counting California at the University of California's Digital Library, the U.S. Census Bureau's DataFerret, and the Web DAIS/DDMS system at Health Canada.

The DDI specification is a work in progress that will continue to evolve. Top priorities include extending the DTD to handle complex file types, creating ways to present DDI codebooks that resemble current written documentation, and developing tools to assist data producers and archives in marking up documentation. The DDI is also moving toward becoming a self-supporting organization known as the Alliance for the Data Documentation Initiative. As the technological environment changes, so will the DDI. The DDI Committee is already looking toward the next generation of

TABLE 2. DDI Committee Member Institutions and Beta Testers

Institutions	Beta testers
Norwegian Social Science Data Services	Centre for Comparative European Survey Data
Harvard University	Danish Data Archive
The American University	U.K. Data Archive, University of Essex
Statistics Canada	Harvard-MIT Data Center
Health Canada	NIWI-Steinmetz Archive
U.S. Bureau of the Census	Norwegian Social Science Data Services
University of Michigan	Survey Research Center, University of California-Berkeley
U.S. Bureau of Labor Statistics	DFD Dokumentationsberatung LLP Arbeits- Berufs- und Wirtschafts-pädagogik, University of Gießen
Inter-university Consortium for Political and Social Research (ICPSR)	Social Science Data Archive, University of Ljubljana
Yale University	Harlan Hatcher Library, University of Michigan
ESRC Data Archive	Minnesota Population Center, University of Minnesota
University of California-Berkeley	Institute for Social Studies, University of Warsaw
University of Southern Denmark	Data and Program Library Service, University of Wisconsin-Madison
The Roper Center	
University of Minnesota	
Zentralarchiv für Empirische Sozialforschung	

Web technologies. The most promising of these include XML Schema, which provides a more rigorous and comprehensive facility for automated processing of XML documents and the Resource Description Framework (RDF), a language for representing information on the Web to enhance exchanges between applications without losing meaning.[3]

Because leading international social science data archives and statistical agencies offer a high level of support and involvement, we can be confident that the DDI will not become obsolete in the foreseeable future. Eventually, new metadata standards will emerge. When they do, the large base of existing DDI-compliant documentation ensures that software will be created to migrate DDI-compliant documentation to these new standards.

NOTES

1. The DDI Tag Library is available at http://www.icpsr.umich.edu/DDI/CODEBOOK/codedtd.html [accessed 11/24/2002].
2. The proposed aggregate extensions to the DDI are described at http://www.icpsr.umich.edu/DDI/CODEBOOK/index.html#00 [retrieved: 24 November 2002].
3. For more information on RDF and XML Schemas, see http://www.w3.org [retrieved: 24 November 2002].

REFERENCES

Berners-Lee, T. J. Hendler, and O. Lassila. 2001. The semantic web. *Scientific American* 284(5): 28–37.

Data Documentation Initiative (DDI). 2002. About the organization. http://www.icpsr.umich.edu/DDI/ORG/index.html [retrieved: 24 November 2002].

Wozniak, R. 2001. Emerging from the quagmire: Building expert systems technologies for the social sciences. *IASSIST Quarterly* 25(4);15–18.