The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Data Documentation Initiative: Toward a Standard for the Social Sciences

Mary Vardigan,
Inter-university Consortium for Political and Social Research (ICPSR),
University of Michigan

Pascal Heus,
Open Data Foundation,
Tucson, Arizona

Wendy Thomas,
Minnesota Population Center,
University of Minnesota

July 2008

Abstract

The Data Documentation Initiative (DDI) is an emerging metadata standard for the social sciences. The DDI is in active use by many data specialists and archivists, but researchers themselves have been slow to recognize the benefits of the standards approach to metadata. This paper outlines how the DDI has evolved since its inception in 1995 and discusses ways to broaden its impact in the social science research community.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The Data Documentation Initiative (DDI) is an international XML-based standard for the compilation, presentation, and exchange of documentation for datasets in the social and behavioral sciences. Documentation, a form of metadata, constitutes the information that enables the effective, efficient, and accurate use of those datasets. Standardized documentation facilitates data access and discovery, improves overall quality, ensures long-term preservation of the information, fosters evidence-based policy-making, and supports the establishment of results-based monitoring.

The most recent version of the DDI, published in April 2008, documents the life cycle of research data and encourages the reuse of metadata for purposes of efficiency and cost savings. While using the DDI is seen as best practice by many social science data archivists and data scientists, researchers have been slow to recognize the benefits of the DDI approach to metadata. Below we discuss the role of metadata in research and the advantages to using the DDI metadata standard, as well as ways to encourage adoption of DDI in the larger social science research community.

The Importance of Metadata

Sharing data is an important ethic in the social sciences, and an international network of data archives exists to facilitate data access and preservation. However, for a secondary analyst to understand a given dataset, he or she must have access to good documentation. In OAIS terms, documentation functions as the Representation Information -- "The information that maps a Data Object into more meaningful concepts." (Consultative Committee for Space Data Systems [CCSDS], 2002) More specifically, "In the social science archive context, a typical example of a Data Object to be preserved would be a numeric survey data file; its associated technical documentation (sometimes called a "codebook"), which is used to understand and interpret the numeric codes in the data file, would comprise the Representation Information. A data file is ultimately just a string of numbers and not understandable on its own; it can only be interpreted and comprehended intellectually through use of the technical documentation, which indicates a variable's location in the numeric data file, the question it was based on, all possible responses to the question, how the population of interest was sampled (for surveys), and so forth. Together, the data file and its documentation make up the Content Information, sometimes called a data collection or a study." (Vardigan & Whiteman, 2007)

Another important scientific norm is replication, which also relies upon having good metadata: "The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." (King, 1995)[1]

Because good documentation is paramount to effective data use, data archives have long encouraged data producers to document their data thoroughly, starting at the very beginning of a research project and in effect creating an audit trail of all variable transformations that take place over the life of the project. In reality, there is little incentive for data producers to follow these guidelines and documentation is often hastily assembled just before deposit into an archive. Furthermore, documentation is

[1] Gary King - Data Sharing and Informatics http://gking.harvard.edu/projects/repl.shtml

most often produced with word-processing software and then rendered into PDF, making reuse difficult.

About the Data Documentation Initiative (DDI)

In response to the need for better documentation, an international group of data archives and producers came together in 1995 to create a documentation standard for the social science research community called the Data Documentation Initiative (DDI)[2]. The DDI is a mechanism for social and behavioral scientists to record clearly and then to communicate to others all the salient characteristics of the empirical data they have collected or compiled. Version 2.0 of the specification, published in 2002, was focused on the elements of a traditional social science codebook and was fairly document-centric, while the new Version 3.0 shifts its focus to the life cycle of the data and metadata. It is designed both to capture information and to present it in a machine-actionable format capable of driving process, data discovery and analysis systems.

DDI and Other Metadata Standards

DDI Version 3.0 provides robust new features and functionality, including expanded alignment with other metadata standards such as Dublin Core, MARC, ISO 11179 (metadata registries)[3], SDMX (data exchange)[4], and geographic standards such as FGDC (Federal Geographic Data Committee) and ISO 19115[5]. The recent proliferation of metadata standards makes it important to be explicit about the contexts in which one might choose to use any given standard. DDI occupies a specific niche and is the most suitable descriptive metadata standard for social science research. The DDI developers have taken the approach that the specification must be compatible with and complementary to the other major standards. For example, in terms of the widely recognized bibliographic standards, there is a crosswalk between Dublin Core elements and DDI,[6] and a mapping between MARC, the library cataloging standard, and DDI metadata records.[7] This supports initial discovery by bibliographic search engines while allowing DDI to describe the richness and complexity of social science data.

The Statistical Data and Metadata eXchange (SDMX) is similar to DDI, but the two standards are actually very different in scope. Whereas DDI documents research across the microdata and aggregate data life cycle, SDMX is concerned with creating efficiencies around the exchange of aggregate data. SDMX is primarily used by the official statistics community for the exchange of time series data. The fact that these two standards are well aligned means that they can be combined in powerful ways; moreover, users of the two standards can move data from one standard format to the other fairly easily (Gregory & Heus, 2007).

DDI's compatibility with ISO 11179 means that any registries (e.g., variable or question banks) built using DDI will be in compliance with the ISO metadata registry standard. DDI ties variables and questions to ISO 11179 at the concept level,

[2] The Data Documentation Initiative (DDI) http://www.ddialliance.org/
[3] The Metadata Registries ISO Standard http://metadata-standards.org/11179/
[4] Statistical Data and Metadata Exchange Standard http://www.sdmx.org/
[5] FGDC and ISO 19115 Standards http://www.fgdc.gov/metadata/geospatial-metadata-standards
[6] DDI: Mapping to Dublin Core http://www.ddialliance.org/DDI/related/dc.html
[7] ICPSR: MARC Metadata Records http://www.icpsr.umich.edu/ICPSR/or/metadata/marc/

supporting both internal and external comparisons. With respect to the geographic standards, DDI developers consulted with geographers and experts in geospatial data to ensure that the DDI captures the core elements needed for resource discovery of social science data without pulling in the bulk of these larger standards.

DDI Features and Functionality

The DDI is now shaped by a self-sustaining member Alliance that brings together data producers, archivists, and users in a collaborative effort. Most of the 32 Alliance members[8] routinely use DDI in their work. In developing DDI 3.0, the DDI Alliance worked closely with the research community to determine the types of content and functionality that the specification needed to support in order to be successful and to meet the needs of all levels of the social science research community.

Metadata Reuse and Comparison

One of the important new features in DDI is extensive support for metadata reuse. The reuse, or secondary analysis, of social science data is an accepted norm in the field, but until recently little thought has been given to reusing metadata. DDI 3.0 is predicated on the principle of reusing metadata to eliminate costly redundancies and support explicit comparison within and between studies. As an example, response categories, concepts and universes can be defined once, and then used multiple times by both questions and variables. Groups of surveys, such as time series, can inherit common information, such as core questions or variables, altering only those items that change in subsequent rounds of data collection. Implicit comparison via inheritance is supplemented by support for describing comparisons explicitly through mapping.

Life Cycle Support

Support for the data life cycle is another innovation in DDI 3.0. Extensive amounts of metadata are generated over the lifetime of a typical social science survey (see Figure 1 below), beginning with an articulation of the concepts to be studied and continuing on to the proposal for funding to conduct the study, the data collection effort itself, the creation of a data file and its documentation, publication and deposit, archiving and dissemination, data discovery, and finally data analysis that leads to new findings and knowledge that enrich the social science literature. The DDI makes it possible to describe the data life cycle and to augment the amount and types of metadata that travel with the data to broaden the context, thereby contributing to the assessment of data quality.

[8] DDI Alliance Structure: Member Institutions
http://www.ddialliance.org/DDI/org/structure.html#members

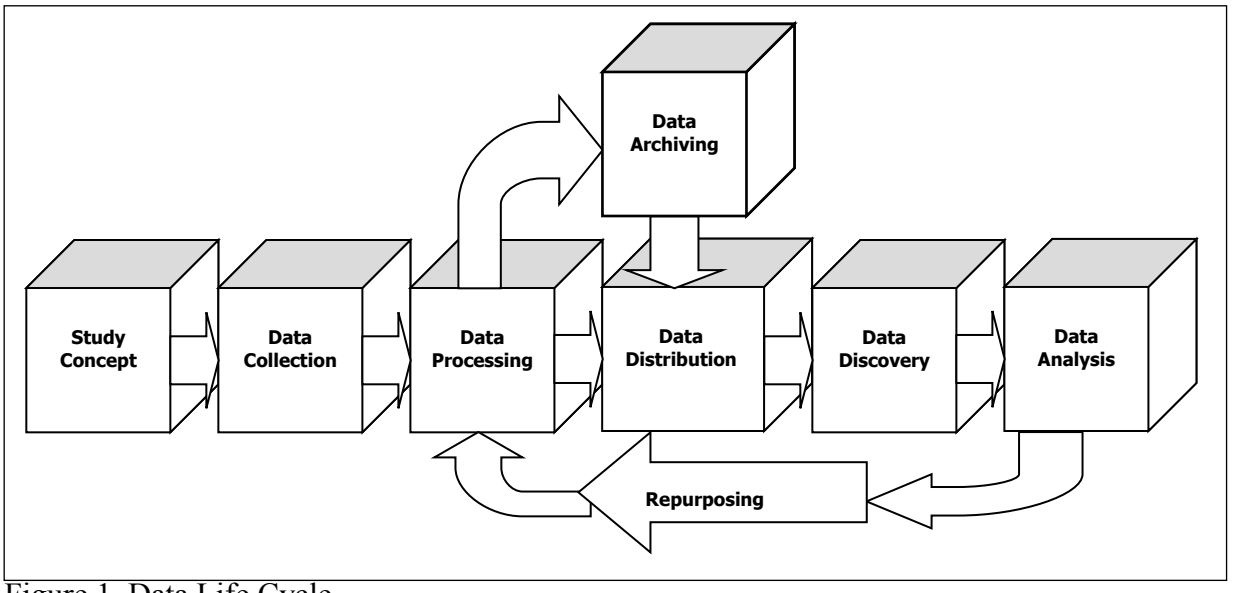

Figure 1. Data Life Cycle

Other DDI 3.0 Features

DDI 3.0 provides support for multiple languages and flexible grouping of studies to support archival organization and secondary research. In addition, it has added the capability to carry data inline as part of a DDI instance along with enhanced coverage of external data storage structures. DDI also facilitates the use of local extensions or overrides so that one may add information to a DDI instance without violating the standard.

Use of DDI

Uptake of the DDI standard has been fairly rapid, especially among the larger social science data archives in the U.S. and Europe. The International Household Survey Network (IHSN)[9], whose members include major international organizations, has adopted DDI as a best practice to improve access to survey datasets in developing countries. Moreover, through the World Bank Accelerated Data Program[10], the IHSN is providing technical assistance to statistical offices in establishing DDI-based national data archives. The DDI is being used by several other projects and organizations around the world as well.[11]

It makes sense that data archives have been the traditional champions of the DDI. In their data curation and support roles they have first-hand experience with the kind of information that is necessary to adequately describe and explain a dataset and they understand that "metadata provide the bridges between the producers of data and their users and convey information that is essential for secondary analysts." (Ryssevik, 2001) Another important driver of data archives' use of the earlier DDI versions is the existence of a software tool called Nesstar[12], created by the Norwegian and UK social science data archives to mark up documentation in DDI format and to analyze and visualize the corresponding data. However, focus on the "codebook" structure meant

[9] International Household Survey Network http://www.surveynetwork.org/
[10] The Accelerated Data Program (ADP) http://www.surveynetwork.org/adp/
[11] DDI: Projects Using the DDI http://www.ddialliance.org/codebook/projects.html
[12] Nesstar http://www.nesstar.com/

that this process was still seen as end-user support.

Social science researchers themselves have been slow to see the advantages of using DDI to document their data. Researchers seem to perceive few incentives to produce high-quality, structured metadata over the life cycle of their research. This is true of both individual researchers and research organizations. The cost of creating metadata without a corresponding benefit to the researcher has been a major stumbling block to the adoption of DDI by this community. DDI 3.0, by facilitating reuse and supporting machine-actionable structures, provides new incentives to the researcher for creating high-quality metadata and using DDI. Research organizations have already noted that facilitation of their production process through use of DDI has been a major factor in their renewed interest in the specification. Researchers must be provided with good tools and a strong business case for the use of DDI as best practice. With a focus on facilitating the full research process, improving data reliability through documentation, and providing wider data access and visibility, DDI offers the potential to create a rich context for the final dataset and thus lead to better science.

The development of tools to integrate DDI into the research process will be a key factor in capturing the attention of this group. DDI is beginning to make inroads into the computer-assisted survey software vendor community, enabling XML markup to happen at the source as an output of the interviewing process. Tools that could receive XML, work with it through their processing and analysis steps and then deposit the revised XML documentation along with their dataset for publication, would streamline the entire process. XML-based tools can make this happen.

Final Thoughts and Conclusions

The DDI has the power to be transformative in terms of the conduct and analysis of social science research. Its robust features enable data comparison while an expanded context for data can make a difference to science, and support a global data environment in which systems of trusted digital repositories are using standards and exchanging data (Vardigan & Whiteman, 2007). This will result in greater access to data for everyone, regardless of where the data are stored, and will showcase the value of a shared infrastructure built on high-quality metadata. What is needed is a concerted effort to present the DDI in terms that make sense to social science researchers so that they become part of the solution for long-term digital preservation. Marketing the DDI to this community will require a sustained campaign to educate them and to persuade them of the value of standards for digital curation efforts and long-term retention of information.

References

Consultative Committee for Space Data Systems (CCSDS). (2002). *Reference model for an open archival information system (OAIS)*. CCSDS 650.0-B-1 Blue Book, January 2002. Retrieved July 9, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Gregory, A., & Heus, P. (2007). *DDI and SDMX: Complementary, not competing, standards*. Open Data Foundation, July 2007. Retrieved July 9, 2008, from http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf

King, G. (1995). Replication, replication. *Political Science and Politics*, *Vol. XXVIII, No. 3 (September 1995)*: pp. 443-499.

Ryssevik, J. (2001). The Data Documentation Initiative (DDI) metadata specification. Paper prepared for *MetaNet 2001*, *Voorburg, Netherlands*. Retrieved July 9, 2008, from http://www.ddialliance.org/DDI/papers/ryssevik.pdf

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: Applying the OAIS reference model to the social science archive context. *Archival Science 7(1)*: pp. 73-87.