Documenting the research life cycle: one data model, many products

Mary Vardigan,[1] Peter Granda,[2] Sue Ellen Hansen,[3]
Sanda Ionescu[4] and Felicia LeClere[5]

Introduction

Technical documentation for social science data is produced by a range of technologies in a variety of environments. While statistical agencies generally standardise the content and the look and feel of the documentation they produce, each agency approaches the task differently. Documentation produced by individual researchers shows even more variety. This heterogeneity in documentation comes at a cost – for data archivists who have had to adjust to new documentation styles with each data deposit in order to curate the collection, and for secondary data analysts attempting to understand datasets with which they were not familiar.

The idea of standardising documentation has gained traction in the past few years. There is much to be gained from such standardisation, particularly with respect to automation of related processes, because standardised content provides a consistent structure to build upon and to programme against. If data archives receive standardised input with each data submission, they can tune their ingest processes according to the structure of the documentation. Structured, machine-actionable documentation can also be used to drive systems, including search and browse, data analysis and subsetting, and data visualisation. Converging on a set of standard elements also facilitates data exchange.

The Data Documentation Initiative (DDI) is an effort to establish an international standard in XML for the content and exchange of metadata describing social science data. Version 3.0 of the specification covers the research data life cycle, from inception of a project, through questionnaire design and data collection, to deposit of the data in an archive and beyond. To leverage the potential of this standard, two organisations at the University of Michigan have undertaken a project designed to produce a shared DDI-compliant data model leading to several different data products. This paper reports on the project, which suggests an approach that statistical agencies may want to consider to increase flexibility and to facilitate data and metadata sharing and reuse.

[1] University of Michigan, Inter-university Consortium for Political and Social Research (ICPSR), PO Box 1248, Ann Arbor MI 48104, USA. E-mail: vardigan@umich.edu.

[2] University of Michigan, Inter-university Consortium for Political and Social Research (ICPSR), PO Box 1248, Ann Arbor MI 48104, USA. E-mail: peterg@umich.edu.

[3] University of Michigan, Survey Research Center, PO Box 1248, Ann Arbor MI 48104, USA. E-mail: sehansen@umich.edu.

[4] University of Michigan, Inter-university Consortium for Political and Social Research (ICPSR), PO Box 1248, Ann Arbor MI 48104, USA. E-mail: sandai@umich.edu.

[5] University of Michigan, Inter-university Consortium for Political and Social Research (ICPSR), PO Box 1248, Ann Arbor MI 48104, USA. E-mail: fleclere@umich.edu.

Brief history of the data documentation initiative

The Data Documentation Initiative (DDI) began in 1995 when the Inter-university Consortium for Political and Social Research (ICPSR) convened an international group to begin work on a specification. The standard began as SGML, and was then converted to Web-friendly XML. The project is now directed by the DDI Alliance (http://www.ddialliance.org), a self-sustaining membership organisation whose members have a voice in the development of the DDI specification. A small steering committee provides governance. The governance structure of the Alliance is based on the World Wide Web Consortium (W3C).

In 2000, DDI version 1.0 was published as an XML DTD. This version of the specification was mainly document- and codebook-centric, following closely the traditional codebook models with which social scientists were familiar. In 2003, the scope of the DDI specification was extended to incorporate aggregate data coverage and geography into version 2.0.

Version 3.0 (http://www.ddialliance.org/ddi3/index.html) of the standard was published in the spring of 2008 after vetting and review by the Alliance members and the general public. This version is a full implementation of XML schemas and emphasises the reuse of metadata through modularity and the use of persistent schemes that can stand alone and be referenced. DDI 3.0 also provides for grouping and comparison of datasets as well as multilingual support. The standard supports other metadata standards, including MARC, Dublin Core, SDMX (statistical data and metadata exchange), ISO 11179 (metadata registries), FGDC (digital geospatial metadata), and ISO 19115 (geographic information metadata). Support for PREMIS (preservation metadata) and METS (metadata packaging) is also being built in. Arguably the most interesting thing about DDI 3.0, though, is its coverage of the research data life cycle, from inception of a project to archiving and secondary analysis.

More generally, creating documentation in DDI and XML provides several advantages. Documentation tagged in XML carries “intelligence” about the content of the data. Since it is ASCII at its core, it will remain usable into the future, unlike proprietary word processing software. In addition, it can be repurposed. A DDI codebook contains all of the information necessary to produce several different types of output, including, for example, a traditional social science codebook, a bibliographic record, or SAS/SPSS/Stata data definition statements. Changes made to the core DDI document will be passed along to any output generated. DDI also lends itself to fielded searching on the Web.

Project stakeholders

The two organisations conducting the demonstration project to show the advantages of using DDI 3.0 are both units of the Institute for Social Research, University of Michigan:

- ICPSR, a large social science data archive
- Survey Research Operations (SRO), a data collection centre

The two organisations had worked together previously on the National Survey of Family Growth (NSFG), sponsored by the National Center for Health Statistics, to create an interactive codebook. They partnered again on the Collaborative Psychiatric Epidemiology Surveys (CPES) (www.icpsr.umich.edu/CPES), sponsored by the National Institute for Mental Health. This involved a harmonisation of three datasets and interactive documentation featuring question comparison and five languages, as shown in Figure 1.

Figure 1

Multilingual variable display in CPES

Together, SRO and ICPSR cover the life cycle of research data and they are natural partners. They both need a rich, high-quality metadata structure and both have the desire to comply with metadata standards – in particular, DDI, since it is the relevant standard in this space. Additionally, passing data easily from SRO to ICPSR without information loss is important to both organisations.

Life cycle metadata tools

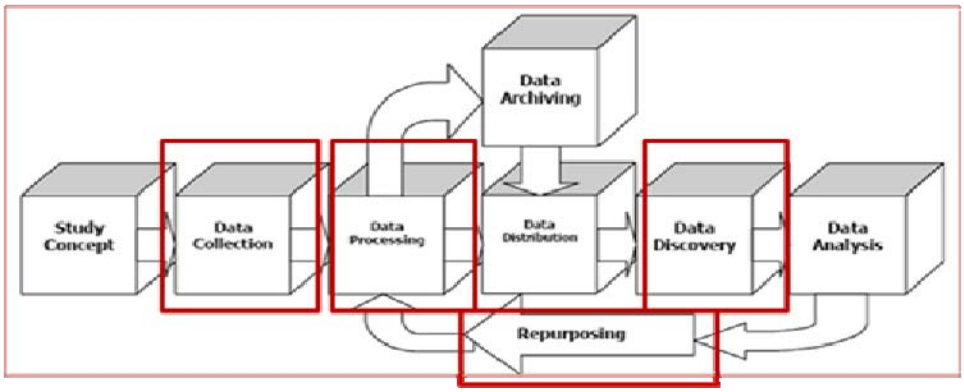
The needs of the partners span several phases of the research data life cycle. From SRO's perspective, what they wanted to derive from the project were tools to complement the MQDS (Michigan Questionnaire Documentation System), which produces XML documentation from Blaise instruments used during data collection. They also needed a tool to permit external users to add metadata for the National Survey of Family Growth during the time when data are being processed in order to produce a public use file.

On ICPSR's side, there was a need to create a robust variable-level search across ICPSR collection for resource discovery, comparison of variables, and ultimately the creation of new datasets and questionnaires. In addition, ICPSR needed a tool to perform internal searches across variables to aid in data harmonisation, which is a type of data repurposing.

Figure 2 shows the life cycle phases covered by the four metadata tools developed from the core relational database.

Figure 2

Data life cycle phases covered by SRO-ICPSR project

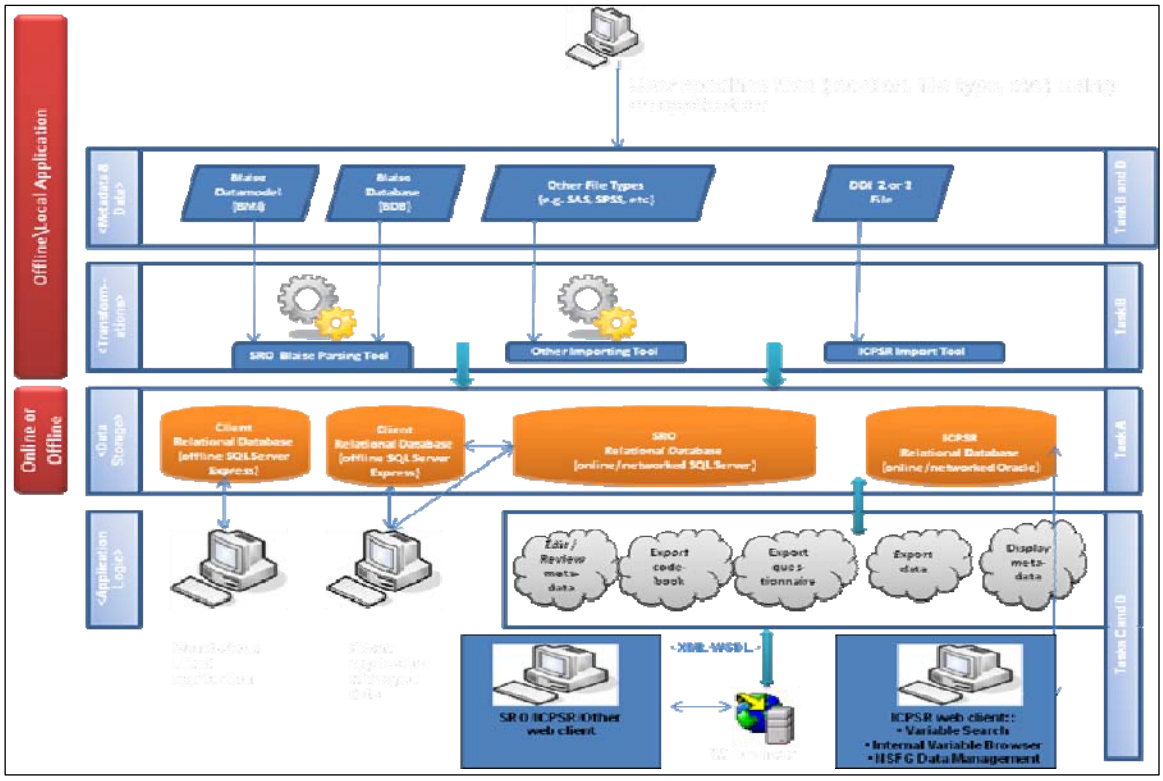


Project design

Teams from the two collaborators met for several months during 2007 and 2008 in order to finalise the data model. Because SRO and ICPSR work in different technical environments, the database had to be implemented on two separate platforms. Figure 3 shows a diagram created early on in the project to guide activities. A flexible model was necessary and thus the partners agreed that while the two implementations would share a common core of metadata elements, each group could create local extensions. This is possible in DDI and does not “break” the standard.

Figure 3

High-level diagram of project functionality



Data collection phase – (MQDS)

SRO uses Blaise for computer-aided telephone interviewing (CATI)/computer-aided personal interviewing (CAPI) surveys, and in recent years they developed a system called MQDS to facilitate automated documentation and harmonisation of Blaise survey instruments and datasets and to extract survey question metadata in a standardised format. The survey metadata MQDS provides include question universe, variable name and label, question text, question variable text (fills), data type, code values and code text, and skip instructions.

MQDS version 1 extracted metadata from the Blaise data model as XML tagged data and provided a user interface for selection of Blaise files, instrument questions and sections, types of metadata to extract, languages to display, and a style sheet for generation of instrument documentation or codebook. However, the first version of MQDS had limitations in that the XML was not DDI-compliant because DDI version 2 did not have XML tags for all metadata provided by Blaise and did not provide easy means of adding XML tags without becoming non-compliant. Another problem was that XML files for complex surveys can be very large; entire files had to be processed in computer memory and there was limited ability to fully automate documentation.

With DDI 3.0 it became possible to document the instrument and it was decided to move from processing XML metadata in memory to streaming metadata to a relational database. The resulting database solution includes DDI-compliant standardised tables and flexibility for SRO and ICPSR to add extensions that meet their specific organisational needs. It also allows automated documentation of any Blaise survey instrument, importing and documenting data produced by other software, and results in lower costs for development of other tools that facilitate editing and disseminating data.

In documenting both the instrument and the data, MQDS offers unique functionality to complement Blaise and other CAI systems.

Data processing phase – editing tool

The relational database also enables the development of new tools to deal with the practical problems involved in transforming data and documentation derived from Blaise instruments into public use products. One such product is an editing tool to load MQDS output into database tables with a Web interface to permit quick viewing. This is an application that permits both internal and external clients to access and edit variable-level information and also provides the ability to include disposition codes to designate which variables to include in public use files. It permits the maintenance of a permanent record of decisions made throughout the editing process.

Data discovery phase – social science variables database (SSVD)

The SSVD enables ICPSR users to search variables across datasets. Furthermore, it assists in data discovery, comparison, harvesting, and analysis and is useful in question mining for designing new research.

The concept was first tested in a pilot project funded by the National Science Foundation and completed in 2005. This product had good functionality and demonstrated the benefits of using DDI markup with easy import, complex, granular searches, and a user-friendly display. However, it included a limited number of datasets (69 ICPSR studies included) at the time.

ICPSR was able to automate the production of DDI-compliant metadata from SPSS source files, making input to SSVD much more efficient. This software produces the full suite of archival distribution products, including data files in SPSS, SAS, and Stata formats, as well as raw ASCII text plus setup files and a variables description in DDI.

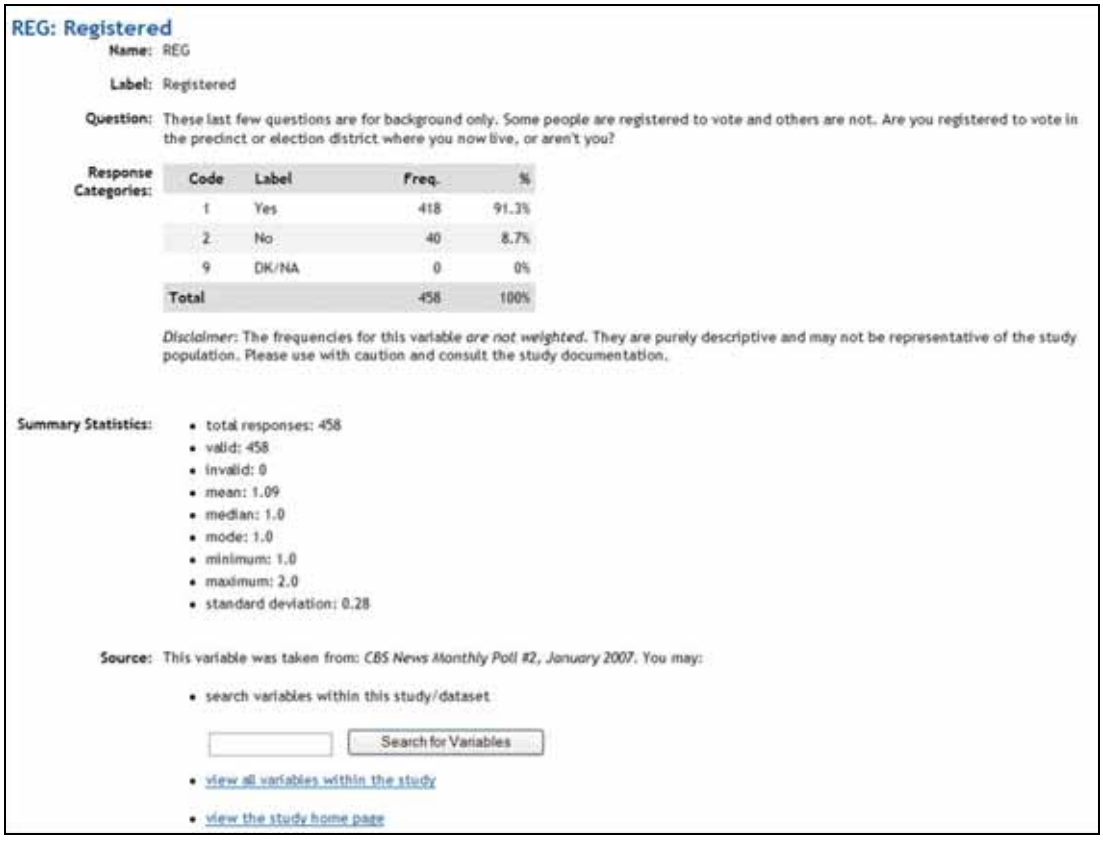
To maximise the effectiveness of the public search, it was necessary to perform additional work to enhance the quality of the machine-generated DDI documentation. ICPSR needed to add question text, whenever available, and needed to increase the readability of variable/value labels, especially if question text was not present.

The new SSVD, which was finalised in autumn 2008, was built to match the DDI 3.0 data model and to be both DDI 2.x and DDI 3.0 compliant. It was designed to accept both DDI 2.x and 3.0 as input and to produce output in both versions. ICPSR version currently uploads DDI 2.1 and generates DDI 3.0 individual variable descriptions.

The variable-level description files in SSVD number more than 5,100 files representing about 2,000 studies and 40 percent of the ICPSR holdings with setup files. Over 1.5 million variables can now be searched, and ICPSR continues to add content. The database can be searched here – http://www.icpsr.umich.edu/ICPSR/ssvd/index.html – and Figure 4 shows a sample variable.

Figure 4

Sample variable from SSVD



REG: Registered

Name: REG

Label: Registered

Question: These last few questions are for background only. Some people are registered to vote and others are not. Are you registered to vote in the precinct or election district where you now live, or aren't you?

Response Categories:

Code	Label	Freq.	%
1	Yes	418	91.3%
2	No	40	8.7%
9	DK/NA	0	0%
Total		458	100%

Disclaimer: The frequencies for this variable *are not weighted*. They are purely descriptive and may not be representative of the study population. Please use with caution and consult the study documentation.

Summary Statistics:

- total responses: 458
- valid: 458
- invalid: 0
- mean: 1.09
- median: 1.0
- mode: 1.0
- minimum: 1.0
- maximum: 2.0
- standard deviation: 0.28

Source: This variable was taken from: *CBS News Monthly Poll #2, January 2007*. You may:

- search variables within this study/dataset

Search for Variables

- view all variables within the study
- view the study home page

Repurposing phase – internal search for data harmonisation

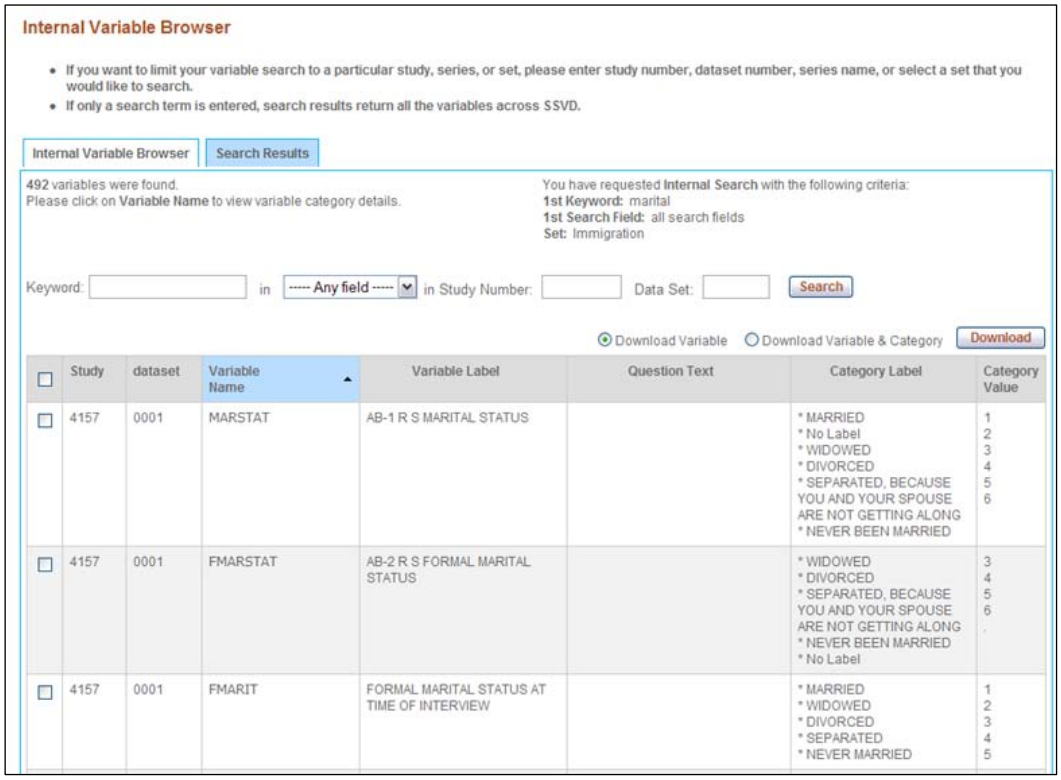
The fourth tool based on the shared data model was designed to aid in post hoc data harmonisation. ICPSR received a five-year grant from the National Institute of Child Health

and Human Development to harmonise data from 10 large surveys of marriage, fertility, and child-bearing in the United States. These surveys, running from 1955 through 2002, comprised the data series Growth of American Families, National Fertility Surveys, and National Surveys of Family Growth.In order to make decisions about harmonisation across all files, ICPSR needed access to question text for all variables along with value labels and categories. Staff needed to be able to find and export metadata from all 10 files at the variable level and to have the capability to document and recode each variable, as well as variable choice. Also important was the ability to do nested searches that were documented, search all fields individually and in sequence, and download results and document what search terms were used.

To that end, all 10 datasets were loaded into ICPSR's version of the shared database, which was designed to capture all of the relevant fields that were marked up in DDI. Figure 5 shows the variable marital status and the different categories to be harmonised.

Figure 5

Internal variable browser results



Downloaded search fields serve to identify variables to be harmonised and provide metadata for translation tables which are used to harmonise files.

Conclusion

The relational database project has shown the benefits of cooperation across research data life cycle stages and the possibility of creating multiple metadata products from a core database designed to be compliant with DDI 3. In addition to providing a means for producing structured documentation for archiving and distribution, the project also shows the

potential for going beyond the traditional codebook in terms of providing instrument documentation, including universe statements and a better view of what interviewers saw when administering surveys. There are also advantages in terms of repurposing data items, as the internal variable search for the harmonisation effort shows.

Statistical agencies that use Blaise will soon have the capability to produce DDI-compliant documentation from Blaise when the MQDS module becomes part of the Blaise package. At this point, export to DDI 3, either directly as XML or potentially into a database, will be possible. Agencies may want to consider storing, presenting, and distributing documentation in this format. As the CPES example (Figure 1) shows, XML can be rendered in any number of ways on the Web and is also easily convertible to PDF for dissemination copies. It lends itself to fielded searching, which is a boon for users. And compliance with the DDI standard ensures that the content can be preserved and will remain usable and reusable over time.