How Initiative Benefits the Research Community - the Data Documentation Initiative

Ken Miller[1] and Mary Vardigan[2]

1 UK Data Archive, University of Essex, England. millk@essex.ac.uk

2 ICPSR, University of Michigan, USA. maryv@icpsr.umich.edu

Abstract. This paper describes the development, establishment and adoption of an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences. It illustrates how that standard benefits the whole social science community and how that community has influenced, and is still influencing, the development of the standard. In particular, the paper highlights the parallel development of a European social science data publishing and statistical browsing tool that has utilized the standard. The paper concludes with a brief insight into the deliberations of one of the working groups set up by the project's Steering Committee and the possible future directions for this effort.

Introduction

The Data Documentation Initiative (DDI) is a project to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences. Documentation or metadata, data about data, constitutes the information that enables the effective, efficient, and accurate use of those datasets.

In response to the need for better documentation, the DDI is an endeavor to provide a straightforward, consistent means for social and behavioural scientists to record clearly, and then to communicate to others, all the salient characteristics of the empirical data for which they are responsible.

The DDI specification, although based on the electronic "codebook", greatly increases the scope and rigor of the information traditionally associated with that kind of document. In addition to documenting measurements in a body of data, such as variable names, labels and values, the DDI specification provides for descriptions of the methodology of the study and information needed to locate specific datasets. Specifically, it covers information such as: a) the mode of data collection b) the sampling methods employed c) the universe and analytical units d) the geographical and temporal coverage of study e) assigned subject classifications and keywords f) the organizations or persons responsible for the production of the final dataset g) the particular stored version of the dataset and h) the access conditions for use of the dataset.

The language in which DDI expresses its information is the most widely used specification language in the world, XML. The DDI is currently expressed as an XML Document Type Definition, or DTD, which defines all of the elements and attributes of social science technical documentation and the relationships among them. There is also an XML Schema implementation.

Hence, the DDI metadata has one major advantage over the old codebooks, in that the information displayed can be fully understood by computer software as well as by humans. The DDI enhances collections and aids interoperability by creating metadata that share a known structure and a specification language across many bodies of data.

Background

The exchange of data between scientists is one of the major requirements of scientific progress and critical to effective exchange is the existence of documentation that enables a full understanding of the data without any consultation with its creator. Within the social sciences it was widely thought that most documentation accompanying datasets was often inadequate.

Moreover, the unstructured and incompatible text formats of the documentation often made them technically obsolete, unable to be read by modern computer software. If the documentation of the data of the social and behavioural sciences is to be shared and reused across software, organisations, and disciplines, - that is become part of the Semantic Web -, it must use a standard specification language, and the information contained must be given well-defined and structured meaning in that language.

Traditionally, it has been the data archivists who have tried to address the communication problems surrounding secondary analysis. Whereas the creators and primary users of statistics have first hand knowledge of the data, secondary users must rely on the documentation supplied with the data to fully exploit its potential. The metadata provides the essential link between the primary data source and secondary use.

The primary goals of social science archives have been preserving data resources and making them readily accessible for secondary analysis. The diversity of these archives' user communities, namely academics, researchers, journalists, planners and decision-makers not involved in any primary data collection but requiring answers from it, goes a long way to explain the high priority that the archives have given to the development of metadata.

Archives are also in the business of data transfer and know only too well that data files can become corrupted during transmission. Good accompanying documentation should therefore include check sums and complete frequencies or descriptive statistics to ensure data integrity.

The other major area in which archives have employed computer readable metadata is the field of resource discovery. Valuable information was completely unavailable for on-line search purposes when the majority of documentation accompanying datasets was only available in paper form.

Within the social science community there was a recognized need for high-quality documentation so that a) secondary investigators could understand and use the data; b) the data could be accurately preserved and transported c) the data could be found more easily d) the social sciences could truly become part of the Semantic Web and aid the cumulative building on prior knowledge.

History of DDI

During the 1970s, the Institute for Social Research at the University of Michigan developed electronic documentation in the form of the OSIRIS codebook. OSIRIS provided documentation at the variable level, such as variable names, variable labels, value labels, and missing values. OSIRIS even allowed limited information about the study itself. However, in practice, mainly due to the high cost of computing, most archives did not provide any electronic documentation at all.

The limits of the OSIRIS documentation were apparent, especially with regards to study-level information, and archives made efforts during the 1980s to develop a standard that would address these limitations. However these developments were only implemented at a few archives, and a study-level standard was not generally accepted.

In 1993, staff from data archives and members of the International Association for Social Science Information Service and Technology (IASSIST), the professional association of data archivists, formed "The IASSIST Codebook Action Group" to work on problems of electronic codebooks.

Richard Rockwell, then the Executive Director of ICPSR, constituted a committee to develop a metadata specification to replace the obsolete OSIRIS standard. The first meeting was held in conjunction with the annual meeting of IASSIST in Quebec City, 1995, where an initial list of codebook elements was drafted. Further meetings resulted in a sample SGML DTD version, prepared by John Brandt of the University of Michigan Library and released in 1996. In 1997 subcommittees were formed to conduct a review of the elements of the DTD and to address the issue of handling aggregate data. In the latter half of 1997 the DDI specification was translated to XML by Jan Nielsen of the Danish Data Archive and it has stayed in this format ever since. Further revisions were carried out by Jerry McDonough, a DTD developer at the University of California-Berkeley Library.

The beta testing of the DDI DTD began in March 1999 and at its conclusion the approved changes were incorporated into Version 1.0, published March 2000.

Several enhancements to Version 1.0 of the DTD were made subsequently, with the most recent stable version of the DTD published as Version 2.0 in July 2003.

Obtaining external funding for the initiative was always challenging. In June 2002, a meeting was held to discuss a draft charter, written by Richard Rockwell, to create a DDI Alliance, with a new membership structure and funding base that would provide self-sustaining support so that the initiative could continue. The charter document provides for an Expert Committee with representation from the DDI Alliance membership, with each member of the Committee having a vote and thus a say in the future of the DDI. The Alliance Steering

Committee met for the first time in February 2003, and the Expert Committee meets once or twice each year.

The DDI metadata specification originated in the ICPSR but is now the project of an Alliance of about 27 institutions in North America and Europe. Together, the member institutions comprise many of the largest data producers and data archives in the world. Virtually every kind of data produced by social and behavioural scientists are to be found in these participating organisations.

Benefits of DDI

The empirical observations of the social and behavioural sciences derive from surveys, censuses, administrative records, experiments, direct observation, and other systematic methodologies for generating empirical measurements. They may pertain to individual persons, households, families, business establishments, transactions, countries, and many other subjects of scientific interest. The observations may consist of measures taken at a single point in time in a single setting, such as a sample of people in one country during one week, or they may consist of repeated observations in multiple settings, including longitudinal and repeated cross-sectional data from many countries, as well as time series of aggregate data. The DDI specification has been designed to fully encompass all of these kinds of data and to provide all the information a potential data analyst needs.

However, the structure supplied by the tagged file of a DDI XML metadata record is perhaps the greatest strength of the standard, in that in allows computer manipulation of the information contained within those tags. This structure allows multiple usages of the information stored within the tags: a) the ability to input the data directly into software packages b) the tailored display of the information through style sheets to satisfy unique user needs c) the ability to perform complex precision searches d) the output of traditional style codebooks.

Basically, the DDI produces a single document with multiple purposes in which changes made to the core document will be passed along to any output generated.

Another major benefit of using the standard is that interoperability within and between systems and organizations becomes that much easier. Codebooks marked up using the DDI specification can be exchanged and transported seamlessly, and applications can be written to work with these homogeneous documents.

The richer content of the DDI encourages the use of a comprehensive set of elements to describe social science datasets as completely and as thoroughly as possible, thereby providing the potential data analyst with broader knowledge about a given collection and facilitating informed use of the data.

Because the DDI markup extends down to the variable level and provides a standard uniform structure and content for variables, DDI documents are easily imported into on-line analysis and sub-setting systems, rendering datasets more readily usable for a wider audience.

Since each of the elements in a DDI-compliant codebook is tagged in a specific way, field-specific searches across documents and studies are enabled. Thus, it is simpler for researchers to discover the studies, variables, or populations relevant to them. With the ability to embed hyperlinks in the DDI XML and through the use of controlled vocabularies for retrieval software to utilize, other relevant work also becomes easier to discover. Through this enhanced discovery functionality, comparative secondary analysis becomes much more feasible.

With the DDI becoming a widely adopted standard, the potential for sharing software developed to manipulate and utilize it increases, creating a software community similar to that of open-source. This, along with the XML specification being widely supported, can also help to reduce the cost of producing good quality documentation.

Influence on the DDI

Through use of the DDI standard, social science documentation becomes available in a computer friendly form. It therefore becomes possible for statistical software to read both the data and associated documentation, thus making possible on-line Web tools that can perform simple statistical analysis. It is the parallel development of such a Web tool, utilizing the DDI standard, which has also influenced the direction that the standard has taken and helped to encourage its use.

Although XML has the ability to be both human-readable and machine-readable, the language itself does not guarantee that documents marked-up using an XML DTD will be machine-actionable. This is only achieved if it is specifically considered in the development process. It is in the development of the machine-actionable aspects of the DDI that the NESSTAR project (Networked European Social Science Tools and Resources) and the resulting Nesstar Limited Company have most influenced the standard.

A metadata standard was vital for the NESSTAR project, since it was aiming at integration across the distributed data holdings of the archives belonging to the Council of European Social Science Data Archives (CESSDA). Before organizations are willing to convert existing metadata to a new standard, they need proof of productivity gains or improvements in the quality of products or services to demonstrate that the investments are worthwhile. The fact that the Nesstar project was so successful meant that software tools were readily available to demonstrate the usefulness and efficiency of the DDI, and thus the standard gained a wider and more general acceptance.

Users of statistical information require a technological environment that enhances their productivity as well as the quality of their work. Such an environment should include at least the following: a) quality empirical data available on-line b) a resource discovery gateway to these data c) multimedia metadata integrated with the data d) on-line browsing, analysis and visualization of data e) data conversion and download into numerous formats and f) hyperlinks from the data to other resources such as publications, other researchers and organisations.

The aim of Nesstar has been to realise as many of these items as possible. The project itself was initiated by the Web-revolution and a drive to develop a common Internet-based gateway to the various data holdings of CESSDA. The company has taken those aims into the social science community as a whole. The DDI XML documentation in the Nesstar system is used for: a) producing the internal files used for on-line analysis b) an export option for the metadata c) the source to generate the indexes used by the resource discovery system and d) navigating the data and metadata via the DDI XML tree.

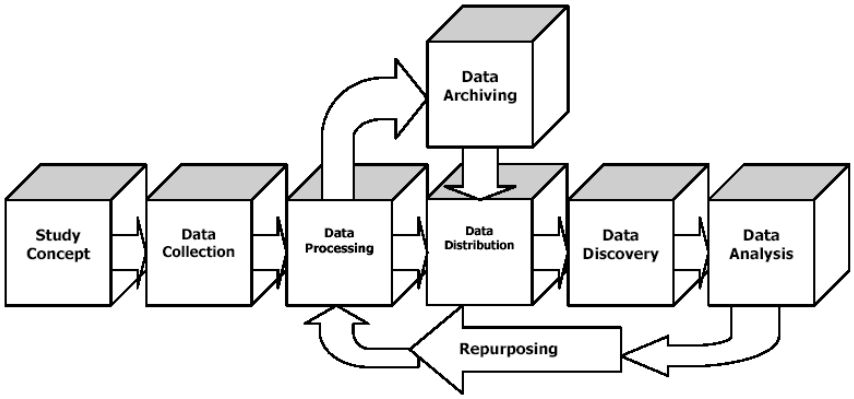
The Nesstar system can be regarded as a software implementation of the DDI standard and has certainly influenced its development. However, given the aims of the system, it is difficult to see how it could have been built without a widely accepted, highly structured metadata standard like the DDI.

Future DDI

The DDI Alliance has set up several working groups to investigate areas of expansion for the standard and outreach activities. These include structural reform, handling complex file types and spatial data, linking to related software resources, improved online support, linkages to other standards and controlled vocabularies.

Of these working groups the structural reform group is at present the most active. It is working towards a more modular and extensible version of the standard. Unlike preceding versions, the new version of the DDI standard will consist of two parts: a) the conceptual model, and b) the XML schemas and DTDs which are derived from it.

This version will also reflect a major change in the scope of the standard. Up until now the DDI has focused on data archiving needs, but with Version 3.0 it is being extended to cover all aspects of the data life-cycle, from conception to reuse. The DDI metadata collected at each stage of this cycle will be passed on to the next so that important information is recorded throughout the life course.



S e to record the relationships between a dataset and those studies that have re-used the data. Thus groups of studies need to be described, either as a series of related studies or as more informal groupings. The ability to express comparability between complete studies and individual variables is also required.

The metadata describing the data collection phase of the life-cycle model will be improved as well. At present most archives capture the paper version of the collection instrument as an image only. A much richer description of the instrument is required to support systems that allow for the re-use of questions. It will also mean that the question text can be utilised by retrieval software, thus enhancing identification of comparable variables.

Because these proposed changes to the DDI are ambitious in scope, one of the major design goals has been to ensure that migration from previous versions is as smooth as possible with minimum resource implications. The simple use of DDI for archival purposes is not radically different between versions, and mappings of all currently used fields will be provided, as well as simple migration tools.

Advances in XML technology has meant that the use of W3C XML Schema (XSD) has become mainstream, and hence this will be used to express the conceptual model instead of an XML DTD. It will also mean that the XML instances will express more validation parameters than are possible with a DTD. The use of XML namespaces will be introduced, thus allowing the expanded vocabulary to be modularized, making it easier to manage and maintain.

Another objective of this new DDI version is to increase the degree to which the metadata it contains is sufficient to support computer processing and thus move toward the goal of the specification being fully “machine-actionable”.

Conclusion

The DDI can serve as the foundation for content, distribution, use and preservation of data collections in the social and behavioural sciences, across institutions, countries, and disciplines. That foundation will be stronger if the specification is independent of any particular software or computing platform. Expressing the specification as a generalized conceptual data model will further enhance this independence. The data model is extensible and modular, supporting the specification of even the most complex data systems in a way that is simultaneously flexible and rigorous.

In further pursuit of the goal of wide adoption, the project is seeking cooperation from

both data producers and statistical software manufacturers. It is hoped that DDI metadata can soon be produced by standard computer-assisted interviewing software and be accessed directly by many statistical software packages for purposes of data definition. When these two aims are achieved, the DDI specification can readily become the basis for the entire research process, from generation of a data collection instrument to production of research articles.

A further challenge will be the extension of the DDI standard to support more complex data. The current specification is excellent for documenting independent survey files, but further work is required to build on mechanisms already included to support aggregate data and hierarchical files.

The DDI serves the social science community well with a specification that produces quality metadata with multiple purposes. It fully documents the details of datasets, it is user friendly and accessible, it integrates into the infrastructure of the Web and it supports automatic generation of statistical software system files.

The widespread adoption of the DDI will vastly improve access to a range of varied datasets. Expanded use will greatly enhance comparative research; the ability to harmonize datasets over time and geography will lead to significant improvement in our understanding of societies. Increasing the availability of high-quality data is a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and that is what the Data Documentation Initiative delivers.

References

Blank, G. and Rasmussen, K.B. "The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." Social Science Computer Review 22, no. 3 (August 2004): 307-318.

Norwegian Social Science Data Services. "Providing Global Access to Distributed Data Through Metadata Standardisation: The Parallel Stories of NESSTAR and the DDI." Working Paper No. 10, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, September 22-24, 1999. (full text)

Ryssevik, J. "Bazaar Style Metadata in the Age of the Web - An 'Open Source' Approach To Metadata Development." Working Paper No. 4, UN/ECE Work Session on Statistical Metadata, Washington, DC, November 28-30, 2000. (PDF 54K)

Leighton, V. "Developing a new Data Archive in a Time of Maturing Standards." IASSIST Quarterly 26, no. 1 (spring 2002): 5-9.

Ryssevik, J. and Musgrave, S. "The Social Science Dream Machine." Social Science Computer Review 19, no. 2 (summer 2001): 163-174.

Green, A., Dionne, J. and Dennis, M. (1999). Preserving the whole: A two-track approach to rescuing social science data and metadata (Technical Report 83). Washington, DC: Council on Library and Information Resources.

Bethlehem J, Kent J., Willeboordse A. and Ypma W. (1999) "On the use of metadata in Statistical data processing", Working Paper No. 23, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999.

Nielsen, J. (1997) "From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation". Unpublished thesis.

The Data Documentation Initiative Web site. http://www.icpsr.umich.edu/DDI/

DDI Structural Reform Group. (2005) "DDI Version 3.0 Conceptual Model". Working draft by Gregory, A. and Thomas, W L.