Final Report to National Science Foundation
"Electronic Preservation of Data Documentation: Complementary SGML and Image Capture"
SBR-9617813
[August 1 1997-July 31, 2000]

1. Participants:

What people have worked on the project?

DDI (SGML/XML) Component:

- Jon Brode, Systems Programmer III, ICPSR
- Peter Granda, Senior Research Associate, ICPSR
- Sanda Ionescu, Research Assistant II, ICPSR
- Peter Joftis, Senior Information Specialist, ICPSR
- Mary Vardigan, Executive Editor, ICPSR
- Ann Green, Director of Social Science Computing, Yale University. (Ann helped to create the online DDI Tag Library and Manual.)
- Jerome McDonough, DTD Developer, University of California, Berkeley. (Jerry made revisions to the DDI Document Type Definition [DTD].)
- Wendy Treadwell, Machine Readable Data Center, University of Minnesota. (Wendy worked on developing new tags to describe aggregate data and on creating a customized codebook entry capability for the DDI.)
- Patrick Yott, Librarian, University of Virginia. (Patrick worked on developing stylesheets using the eXtensible Stylesheet Language (XSL) to display DDI-compliant codebooks.)

What other organizations have been involved as partners?

NESSTAR (http://www.nesstar.org), the Networked Social Science Tools and Resources group, is a collaboration among three European Archives: the UK Data Archive, Norwegian Social Science Data Services, and the Danish Data Archive. NESSTAR's software enables users to locate multiple data sources, browse highly detailed metadata, conduct statistical analyses interactively, visualize data statistically and geographically, and download subsets of cases and/or variables in a number of formats. NESSTAR uses the DDI metadata specification as the underpinnings for its system, so in that sense the parallel development of the DDI and NESSTAR has been a partnership. NESSTAR has recently initiated two related projects, which also have links to the DDI: LIMBER, a multilingual thesaurus effort, and FASTER, a flexible metadata interface that will accelerate access to all types of statistical data. Two members of the NESSTAR team are DDI Committee members, and NESSTAR also participated in the beta-test.

Intentions to use DDI metadata or to promote interoperability with it have also been expressed by FERRET (Federal Electronic Research and Review Extraction Tool, developed by the Census Bureau), SDA (Survey Documentation and Analysis, developed by UC-Berkeley), and the VDC (Virtual Data Center, under development at Harvard).

Have you had other collaborators or contacts?

We worked with 13 beta-testing organizations to obtain assessments of the usability, limitations, and strengths of the DDI DTD:

- London Guildhall University, Centre for Comparative European Survey Data (CCESD)
- Danish Data Archive
- The Data Archive, UK
- Harvard-MIT Data Center
- NIWI-Steinmetz Archive, The Netherlands
- Norwegian Social Science Data Services (NSD)
- University of California-Berkeley, Survey Research Center
- University of Giessen, DFG Dokumentationsberatung LLP, Arbeits-, Berufs- und Wirtschaftspddagogik
- University of Ljubljana, Social Science Data Archive
- University of Michigan, Harlan Hatcher Library
- University of Minnesota, Machine Readable Data Center
- University of Warsaw, Institute for Social Studies
- University of Wisconsin-Madison, Data and Program Library Service

We also worked with the DDI Committee itself, meeting yearly to refine the DTD further and to set direction for future activities. DDI Committee members include:

- Merrill Shanks, Chair (University of California, Berkeley)
- Atle Alvheim (Norwegian Social Science Data Services)
- Martin Appel (Bureau of the Census)
- Grant Blank (American University)
- Ernie Boyko (Statistics Canada)
- Bill Bradley (Health Canada)
- Cavan Capps (Bureau of the Census)
- Bill Connett (University of Michigan)
- Cathryn Dippo (Bureau of Labor Statistics)
- Pat Doyle (Bureau of the Census)
- Dan Gillman (Bureau of Labor Statistics)
- Peter Granda (ICPSR)
- Ann Gerken Green (Yale University)
- Bjorn Henrichsen (Norwegian Data Archive)
- Peter Joftis (ICPSR)

- Ken Miller (ESRC Data Archive)
- Tom Piazza (University of California, Berkeley)
- Karsten Boye Rasmussen (University of Southern Denmark)
- Richard Rockwell (Roper Center)
- Jostein Ryssevik (Norwegian Data Archive)
- Peter Solenberger (University of Michigan)
- Allen Tupek (Bureau of the Census)
- Rolf Uher (Zentralarchiv fuer Empirische Sozialforschung)
- Mary Vardigan (ICPSR)

2. Activities and Findings:

What were your major research and education activities?

On March 24, 2000, the Data Documentation Initiative (DDI) Committee published Version 1 of the DDI Document Type Definition, or DTD. This accomplishment was made possible in large part by this NSF grant, which began in 1997.

This project actually began in 1995 when Richard Rockwell, Executive Director of the Inter-university Consortium for Political and Social Research (ICPSR), appointed a committee to develop a structured format for codebooks documenting social science datasets. The original codebook specification was written using the Standard Generalized Markup Language (SGML), a nonproprietary language that permits the tagging of fields, or elements, for both content and display. Later, after the development of XML, or eXtensible Markup Language (an instance of SGML better suited for the Web), the DTD was "translated" into XML. The specification is represented by a Document Type Definition, or DTD, that presents all of the elements of a codebook in a hierarchical structure.

Following are summaries of the objectives and deliverables of the DDI component of this NSF grant.

Goal 1: Markup of a dozen codebooks, representing a variety of data types, according to the DDI standard. Once Version 1 of the DTD was published, this work could move forward. It has now been completed, and the resulting marked up codebooks can be viewed using Internet Explorer 5 at:

http://www.icpsr.umich.edu/DDI/codebook/sample.html

Goal 2: Evaluation of markup software. In terms of *commercial* products, ICPSR evaluated the following packages:

FrameMaker+SGML 5.5 (Adobe). We were unsuccessful in loading the DTD into the software. Adobe is promising that the next version of the software will be more user-friendly with more of an emphasis on XML. It is fairly expensive, $348.

Adept Editor 8.0 (ArborText). This software required numerous consultations with the technical support staff at ArborText in order to load the DTD because modifications were required for XML. The package is fairly robust but has some drawbacks: it is possible to create an invalid document instance if required elements and attributes are not turned on, and the export capability produces lowercase tags, ignoring the case-sensitive nature of XML tags. This product is very expensive: the editor costs $1500 and another piece of software called Document Architect ($7000) is also required for initial DTD loading. Fortunately, we were able to persuade ArborText to lend us a copy of Document Architect.

XMetaL (SoftQuad). This is the program we ultimately used for markup. We were able to load the DTD with a minimum of effort; the product appears to be adequate but not as fully functional as we would like. The price of XmetaL is $395. ICPSR has also looked at Author/Editor (previously SoftQuad, now Interleaf), the precursor to XMetaL.

In terms of *shareware*, ICPSR has evaluated the UNIX EMACS add-on for SGML, which is the authoring software used by the DTD developer Jerry McDonough. This program requires a lot of specialized technical expertise. Several beta-testers assessed other shareware, such as SPY XML (Icon Information Systems), Clip! (Techno 2000, Alpha), and XML <PRO> (Vervet Logic). Most of these products have significant limitations.

We have not yet had a chance to evaluate *automated markup* tools like Omnimark and Balise. These tools require a fair amount of technical skill to use, and it is not clear how helpful they would be given the heterogeneous nature of our codebook holdings.

Goal 3: Development of a procedure for automatically producing SAS and SPSS data definition statements from the DDI metadata source. This task is not yet complete, but we are working with Patrick Yott at the University of Virginia to develop a stylesheet using the eXtensible Stylesheet Language (XSL) Transformations, or XSLT. See Findings, Goal 3, for more information.

Goal 4: Translation of OSIRIS codebooks into the DDI standard. We have written a program for this conversion and have applied the program to ICPSR's existing OSIRIS codebooks. Because there are differences in format among the codebooks (despite use of the OSIRIS "standard"), not all codebooks were able to be converted and there is editing work remaining. A list of codebooks converted from OSIRIS to DDI can be found at http://www.icpsr.umich.edu/DDI/codebook/sample.html.

Goal 5: Preparation of a manual and a tag library, providing instruction in and support for use of the DTD. This project, headed by Ann Green, is complete. We now have an extensive tag library and guidelines for using the DTD on the DDI website at http://www.icpsr.umich.edu/DDI/codebook/codedtd.html. In the future we plan to convert the HTML version of the tag library to XML, with a stylesheet for display, to provide a further example of the versatility of XML and the stylesheet approach.

Goal 6: Beta-test of the DTD at six sites. The beta-test took place during March-August of 1999 with 13 test sites. All 13 submitted final reports and several have made their XML documents and programs available as well. The DDI Committee reviewed the final reports and at its meeting in October 1999 made a list of changes to the DTD based on the testers' recommendations, which were implemented.

***What are the major research findings resulting from these activities*?**

Goal 1: Markup of a dozen codebooks, representing a variety of data types, according to the DDI standard. A sense emerged early on from the DDI Committee that the first phase of the DDI should be geared toward documenting survey data only and that aggregate data and more complex file types would be dealt with later. Consequently, marking up codebooks for different types of files came to be seen as less of a priority. Nevertheless, our markup choices did represent a diverse set of studies, ranging from the March 1999 Current Population Study to Polity III, an international relations dataset, to the World Values Survey.

Goal 2: Evaluation of markup software. The basic conclusion about existing XML markup tools is that most of them are quite expensive and all are limited in some way. Three betatest sites -- University of Ljubljana, University of Minnesota, and University of Giessen -- worked on customized DDI codebook entry programs that would overcome the limitations inherent in existing off-the-shelf products and simplify the codebook production process. ICPSR also assessed an early prototype of a customized codebook entry package prepared by Blue Angels Inc. At this point in the development of XML authoring tools, applications specifically tailored to codebook preparation are sensible alternatives because they permit codebook authors to focus specifically on the task at hand and to bypass the cumbersome steps involved in loading the DTD and learning the intricacies of XML.

Goal 3: Development of a procedure for automatically producing SAS and SPSS data definition statements from the DDI metadata source. This task is not yet complete, but we are working with Patrick Yott at the University of Virginia to define a procedure using XSL Transformations (XSLT). We have discovered that the technology for working with XSLT has not advanced as quickly as we (and many others) had hoped. The processing of XSLT is not yet native to all browsers. We are working on a solution that involves employing a Java servlet, which can be implemented from the server side without requiring the browser to do the XSLT processing. While we could have elected to use perl scripts for the conversion, we have instead chosen to stay within the framework of the W3C standards and specifications as that seems the path with the most potential for this project.

Goal 4: Translation of OSIRIS codebooks into the DDI standard. As mentioned above, we discovered that OSIRIS codebooks in the ICPSR archive are not always uniform. Some sets of studies were handled differently -- e.g., for the Eurobarometers crossnational tables with frequencies were merged into the codebooks. Similarly, we

discovered that OSIRIS has different variants across other archives. Thus, this task, which we considered at the time the proposal was written to be a simple one, became more complicated as we looked into it more deeply.

Goal 5: Preparation of a manual and a tag library providing instruction in and support for use of the DTD. We determined that an online manual and tag library makes the most sense for the DDI. The expectation is that the tag library will be dynamic and will undergo changes as new uses arise and as new examples are added. Updating a print manual this frequently would be expensive and time-consuming.

Goal 6: Beta-test of the DTD at six sites. We discovered that there was a great deal of interest in the DDI from the social science research and archiving communities. The fact that so many different organizations from around the world applied to beta-test the DTD was an affirmation that the standards approach to metadata is a sound one and that there are many applications that could build upon the DDI. We decided that expanding the beta-test from the 6 sites originally proposed to 13 sites afforded considerable benefits in terms of "seeding" the DDI around the world, gaining wider support for the emerging specification, and gathering useful data to employ in refining the DTD.

Opportunities for training and development provided by the project

Training workshops in the use of the DDI DTD were offered at the May 1999 (Toronto) and June 2000 (Evanston) meetings of the International Association for Social Science Information Services and Technology (IASSIST). The June 2000 training involved instruction in the use of a tool for codebook entry developed at the University of Minnesota and in the use of the XSLT to present and display marked-up technical documentation for social science datasets.

Outreach activities undertaken

Realizing the importance of promoting awareness of the DDI in order to gain widespread acceptance and adoption of the specification, ICPSR was very active in pursuing outreach activities. We gave presentations at and participated in the following meetings and conferences:

- International Association of Social Science Information Services and Technology (IASSIST):
  Yale University, New Haven, CT -- May 1998
  Ryerson University, Toronto, Canada -- May 1999
  Northwestern University, Evanston, IL -- June 2000
- OMG (Object Management Group) -- San Jose, CA, August 1999. OMG is a non-profit organization that sets interface standards for commercial software applications, including standards for survey questionnaires. Tom Piazza (Berkeley) attended.
- Council of European Social Science Data Archives (CESSDA) Expert Seminars:
  Koeln, Germany -- September 1999

Tampere, Finland (upcoming) -- September 2000

- ICPSR Meeting of Official Representatives -- Ann Arbor, MI, October 1999
- Open Forum SC32 (Metadata Registries and Standards, including ISO 11179) -- Santa Fe, NM, January 2000
- Interface (Metadata Standards) -- New Orleans, LA, March 2000
- Metadata Experts Workshop -- Voorburg, Holland, April 2000
- TADEQ Meeting -- Lisbon, Portugal, January 2000. TADEQ, Tool for the Analysis and Documentation of Electronic Questionnaires, is being developed by a group based in the Central Bureau of Statistics in the Netherlands to document Blaise CAI instruments. Tom Piazza attended.
- Metadata Standards for End-User Access to Data -- Vancouver, Canada, June 2000

3. Products:

What publications have resulted from this project?

- DDI Tag Library and Manual -- http://www.icpsr.umich.edu/DDI/codebook/codedtd.html

- Ryssevik, Jostein. "Providing Global Access to Distributed Data Through Metadata Standardisation: The Parallel Stories of NESSTAR and the DDI." Statistical Commission and Working paper No. 10. Economic Commission for Europe, Conference of European Statisticians. Geneva, Switzerland, September 22-24, 1999. -- http://www.nesstar.org/papers/GlobalAccess.html

- Nielsen, Jan (1997). "From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation." Unpublished thesis. This thesis focuses on the conversion of the DTD from SGML to XML, which was performed by Jan Nielsen at the Danish Data Archive.

- Musgrave, S., and Ryssevik, J. "The Social Science Dream Machine: Resource Discovery, Analysis, and Delivery on the Web." Paper presented at IASSIST Conference "Building bridges, breaking barriers: The Future of data in the global network." Toronto, Canada, May 1999. -- http://www.nesstar.org/papers/iassist_0599.html

- Kim Tully (Editor), Lu Chou (Special Librarian), and Cindy Severt (Senior Special Librarian). "DPLS Takes Part in Data Documentation Initiative." http://dpls.dacc.wisc.edu/pubs/Newsletters/may99news.html#story1

What Web site reflects this project?

The DDI has an extensive Web site at: http://www.icpsr.umich.edu/DDI

The DDI Web site is referenced by several other sites, including:

- Cover, Robin. “The XML Cover Pages: Data Documentation Initiative: A Project of the Social Science Community” -- http://www.oasis-open.org/cover/ddi.html

- Diffuse Standards and Specifications List -- http://www.diffuse.org/archives.html#DDI

What other specific products have you developed?

We have developed a conversion routine to translate OSIRIS codebooks into DDI marked-up documentation. Conversion routines to move from DDI to SAS and SPSS are in development and should be completed by fall 2000.

Earlier this year Wendy Treadwell (University of Minnesota), at our request, headed up an effort to develop software specifically tailored to codebook preparation for Version 1 of the DTD; this application, demonstrated at the IASSIST conference in June 2000, permits codebook authors to enter codebook text in a logical, step-by-step process and to export parsed XML documents.

Patrick Yott (University of Virginia) is developing stylesheets for the display of DDI marked-up codebooks and is also assisting with the routine for the conversion to SAS and SPSS data definition statements.

4. Contributions

In developing a specification for the content, transport, presentation, and preservation of social science metadata, the DDI project has contributed to the conduct of social science research and archiving in the following ways:

Improved search precision. Because elements of a codebook can be tagged separately for content using this XML application, detailed, field-specific searches are possible. This enables a user to search, for example, all question text for specific terms or to search the sampling field across studies.

Improved finding aids for resource discovery. Because the DDI provides “header” information describing the marked-up document itself, locating DDI-compliant documents is expedited.

Improved interoperability. We provide a mapping from DDI elements to Dublin Core elements to aid further in the metadata standardization effort. Our intention is to add a crosswalk to other standards such as GILS and MARC.

Ability to produce multiple products from a single source. Marked up codebooks can be used to create bibliographic citations, full study descriptions, setups for statistical packages, etc.

Ability to display information in several ways from a single source. Using stylesheets, an XML codebook can be transformed to display in different presentation formats.

Improved consistency and completeness of documentation. Markup provides a way to standardize metadata and to ensure that codebooks meet accepted guidelines for content. The DDI effort should help to encourage data producers to provide in their documentation all of the information necessary for users to determine whether a given study meets their research needs and to perform accurate analyses. While the DDI Committee decided not to specify many elements as required, it did develop a list of elements that are highly recommended for inclusion in codebooks.

Facilitation of the job of intelligent agents. The DDI metadata provides a structured schema that can be employed by systems for data extraction and analysis, such as FERRET (Federal Electronic Research and Review Extraction Tool, developed by the Census Bureau), SDA (Survey Documentation and Analysis, developed by UC-Berkeley), NESSTAR, VDC (Virtual Data Center, under development at Harvard), etc. It also provides the potential to pull together information from separate systems.