Expert Committee Meeting

May 22-23, 2005
Edinburgh, Scotland

Present:
Tom Piazza, Chair (University of California - Berkeley); Iris Alfredsson (Swedish Data Archive); Titto Assini (UK Data Archive, Manchester University); Atle Alvheim (Norwegian Data Archive); Ernie Boyko (formerly of Statistics Canada); Bill Bradley (formerly of Health Canada); Gina Cheung (University of Michigan); Mark Diggory (Harvard-MIT Data Center); Ilona Einowski (University of California - Berkeley); J Gager (AEON Consulting); Ann Green (Yale University); Arofan Gregory (AEON Consulting); Reto Hadorn (Swiss Data Archive); Pascal Heus (World Bank); Lon Hofman (Statistics Netherlands); Sanda Ionescu (ICPSR); Mari Kleemola (Finnish Data Archive); I-Lin Kuo (ICPSR); Hans Jorgen Marker (Danish Data Archive); Marc Maynard (Roper Center); Kate McNeill-Harman (MIT); Meinhard Moschner (Zentralarchiv, Cologne); Ron Nakao (Stanford University); Rob O'Reilly (Emory University); Ed Ross (OpenSurvey Project, Triple-S); Jostein Ryssevik (Nesstar Ltd); Jeremy Iverson (University of Wisconsin - Madison); Ken Miller (UK Data Archive); Ingo Sieber (German Socio-Eonomic Panel Study); Dan Smith (University of Minnesota); Wendy Thomas (University of Minnesota); Mary Vardigan (ICPSR); Joachim Wackerow (ZUMA); Oliver Watteler (Zentralarchiv, Cologne); Marion Wittenberg (Dutch Data Archive)

Introductions

DDI Alliance Expert Committee Chair Tom Piazza opened the meeting and welcomed participants to Edinburgh.

Data Model

Wendy Thomas and Arofan Gregory presented an overview of the Version 3.0 DDI specification, based on documents provided prior to the meeting. The goal of these documents is for the Working Groups to understand the structure of Version 3.0 in order to submit proposals that adequately cover their respective substantive areas. The presentation emphasized the following points:

- A shared terminology is very important for developers of Version 3.0.
- With the new version, there will be a conceptual model in UML and various technical implementations, with the canonical expression being an XML Schema.
- Version 3.0 will have a modular structure and will reflect the true life cycle of a dataset, rather than focusing solely on its archival manifestation.
- HTML tagging for formatting will be possible.
- In place of the Version 2.0 "generic elements," there will be reusable classes that can be applied anywhere in an instance.
- Version 3.0 will have increased support for computer processing and will move in the direction of being machine-actionable (as opposed to human- or machine-readable).
- An upper-level wrapper will be part of each instance and will reference all the modules used in the instance. The wrapper concept is taken from METS -- the Metadata Encoding and Transmission Standard.
- For complex instances, there will be a grouping mechanism with six different parameters or dimensions. This will help to describe complex longitudinal and comparative datasets. The idea is that information common across all files only needs to be entered once, and then subsequently only differences among them need to be entered. Formal relationships among the grouped units are specified in the grouping module, and inheritance is used. Because instances will grow in size with the new modules, the capability to inherit properties is important.
- Version 3.0 will use data typing extensively. This means that a date, for example, will have a required format and syntax.

- Version 3.0 distinguishes between simple and complex instances. Simple instances will be similar to Version 2.0 instances. Additional modules will be available for those who need them but will be optional.
- The SRG will provide a mapping tool to move Version 2.0 instances to Version 3.0 in as automated a way as possible.

Questions were raised about the life cycle model and where the knowledge products (e.g., reports and articles and published tables) fit into this scheme. The consensus was that knowledge products are outside of the model.

Another question was raised about separating the question formulation from the logical encoding of a variable because the question and its responses imply a logical structure for a variable, even though there is not always a one-to-one correspondence between an original question and the variable that becomes part of the final data file.

How the new life cycle model coordinates with ISO11179 was another point of discussion. Because the new model gives more prominence to concepts, it should be relatively easy to interface with ISO11179, which is based on concepts.

The difference between the Group and Comparison modules was also discussed. For longitudinal studies, one would use both modules.

We will need good tools to maximize efficiency and automate metadata entry over the life course of a data collection. SPSS is moving toward a life cycle model, and it would be great to coordinate with them. Also, the World Bank is working on tools to archive and disseminate data, which will ultimately be open source and freely available.

The best-case scenario to encourage wide use and support for the DDI is to have XML produced at source by the large CAI firms like CASES, Blaise, and CSPro (used by the Census Bureau in developing countries). We need a facility to add elements like "concept" to the programming of the original instruments and then encourage researchers to fill in this information. We need to end up with not just a dump of the instrument but with additional substantive metadata.

We also need to think about the future and where the infrastructure for social science research is heading. This may involve self-documenting digital objects that embed metadata in the resource itself. The future probably won't involve data files as we know them now.

We need to test our specification and tools for usability. Version 2.0 is not easy to use -- it has abbreviated tag names and other limitations -- and we need to ensure that Version 3.0 is a marked improvement.

It was noted that the draft of the data model moves pieces of variable-level information to different modules -- e.g., logical and physical structures. Is this necessary? We will need to deliver tools that permit the user to not have to deal with these issues.

Working Group Status Reports and Goals

Before the breakout sessions for Working Groups, questions were raised about what each group be should be focusing on. Should the groups be looking at the minimum we can accomplish now or should we take the big leap and make dramatic changes in Version 3.0? Most were in agreement that taking this big step is important at this juncture. The Working Groups have much to accomplish in a short time: the SRG needs to know soon what the important substantive structures and components are for each group.

Aggregate Data, Geography, and Time Working Group

In its group meeting, this group identified some modifications to geographic coverage that are needed, particularly to assist in data searches. The ability to reference external classification schemes for geography is important. With respect to time, the group identified three different ways in which time may be used: administrative datestamps, methodology, and time as a variable. There is also a time dimension to some external resources like maps. The group also needs to provide the ability to handle multiple time ranges adequately in Version 3.0. In terms of reviewing the aggregate data specification and n-cubes, the group suggests aligning with SDMX, which by implication means embedding data values in the instance.

The group will develop multiple use cases and plans to have weekly telephone conference calls starting on Tuesday, June 7, 2005.

Instrument Documentation Working Group

This group can take advantage of existing tools documenting the instruments of specific vendors like Blaise and CASES in order to identify elements that we need to incorporate into Version 3.0. An important issue that more recently came up for this group is the recoding issue -- that is, variables may be recoded prior to creation of the ultimate data file, and documenting this is critical. The group is also looking at how we should represent the logic of a recode -- whether with SAS code, MathML, etc.

Comparative Data/Families of Datasets Working Group

This group wants to develop the capacity to describe datasets that are comparable over time and across nations. These include data collections like the Comparative Study of Electoral Systems, the European Social Survey, and the Eurobarometer Trend File. There are two orientations in the group: the relational database approach and the XML approach. The group will take real-life examples and make suggestions for extending the comparative and grouping modules.

Usability and Outreach Working Group

While this group did not meet formally in Edinburgh, it will soon be working on a revamped DDI Web site that has a new focus on soliciting information from DDI users and prospective users.

Timetable

The goal is for Working Groups to submit proposals by September 1, 2005. Then the SRG will begin work on the 3.0 draft UML model and XML Schema. The ultimate goal will be to publish Version 3.0 sometime in 2006 after all of the required reviews and votes take place.

A meeting will be scheduled in October 2005 for the SRG and representatives from each of the Working Groups, either in conjunction with the ICPSR OR meeting or around the time of the CESSDA Expert Seminar in Madrid. We will do estimates to determine which meeting could be held most cheaply.

Additional Meeting Held

On Tuesday, May 24, members of the DDI Working Group on Comparative Data/Families of Datasets met with representatives of two other groups -- the Comparative Survey Design and Implementation (CSDI) group and the MetaDater group. The CSDI project involves participants from Survey Research Operations at the University of Michigan's Institute for Survey Research and the Center for Survey Research and Methodology (Zentrum für Umfragen, Methoden und Analysen) in Mannheim. The CSDI group has designed a tool called Survey Metadata Documentation System (SMDS) that documents a survey from the initial design stage through archiving of the resulting dataset. MetaDater is a European Union project involving a consortium of European social science archives that is working to describe large-scale comparative surveys over space and time. Since it was clear that these groups had overlapping missions and interests, a meeting was arranged to share perspectives and approaches. Each group gave a summary of progress, and there was a discussion of key ideas and themes and how the groups can cooperate and work together.