Informal DDI Expert Committee Meeting

October 16-17, 2006
Zentrum fuer Umfragen, Methoden und Analysen (ZUMA)
Mannheim, Germany

Present:
Hans Jorgen Marker, Danish Data Archive (Chair); Iris Alfredsson, Swedish Data Archive; Frederic Berger, CEPS-INSTEAD, Luxembourg; Karl Dinkelmann, University of Michigan; Jan Goebel, German Socio-Economic Panel Study; Peter Granda, University of Michigan; Reto Hadorn, Swiss Data Archive; Janet Harkness, GESIS-ZUMA; Jeremy Iverson, Colectica; Ken Miller, UK Data Archive; Peter Ph. Mohler, GESIS-ZUMA; Meinhard Moschner, GESIS-ZA; Beth-Ellen Pennell, University of Michigan; Sigbjorn Revheim, Norwegian Data Archive; Pilar Rey del Castillo, Spanish Data Archive; Dan Smith, Colectica; Wendy Thomas, University of Minnesota; Mary Vardigan, ICPSR; Joachim Wackerow, GESIS-ZUMA

Monday, October 16

Introduction

Peter Ph. Mohler, Director of the Zentrum fuer Umfragen, Methoden und Analysen (ZUMA), welcomed DDI Alliance participants to the meeting, which was intended to be an informal meeting of the Alliance to provide a forum for questions and answers and a time for working groups to meet. He indicated that the larger social science community is now being exposed to the benefits of the DDI, and the Alliance should make it a point to involve not only the archives but other constituencies like data producers. Hans Jorgen Marker, Chair of the DDI Alliance, then outlined the agenda for the two days.

Timetable to DDI Version 3.0

Mary Vardigan, Director of the Alliance, presented some information about the timetable for moving the draft DDI Version 3.0 specification to the period of Public Review. She indicated that Version 3.0 is still in the period of Internal Technical Review and that this will continue through November 2006. The Alliance will then vote in December (online vote) to move to Public Review, which will take place in January-February 2007. The goal is to publish Version 3.0 in the Spring of 2007.

This is a critical time for the Alliance. The timetable is ambitious and we need to have a good set of tools and supporting materials ready for the Public Review. We all need to pitch in to make sure our rollout of Version 3.0 is successful.

For Public Review, it will be important to have good examples of DDI markup to highlight the ways in which the new DDI specification can be used. During the period of Public Comment, we must demonstrate Proof of Concept. In the context of DDI Version 3.0, there are two components to Proof of Concept: (1) the markup examples themselves, and (2) implementations. Regarding #2, important implementations will be a 2.x to 3.0 migration tool and a stylesheet/viewer that enables one to browse an instance. At the end of the period, the Alliance will vote to accept the specification and to publish it. In terms of suggestions for where to publicize the public review, the group suggested the ddi-users and IASSIST listservs as well as listservs for users of statistical software and data collection software.

Currently, we plan to present the following examples:

- Simple ICPSR metadata record level markup (Sanda Ionescu, ICPSR)
- Variables from an ICPSR media poll (Sanda, ICPSR)
- Variables from American National Election Study (Sanda, ICPSR)
- Variables from a country-level time-series data file (Rob O'Reilly, Emory University)

- Comparative example: Variables reflecting question wording changes (NSFH or other longitudinal dataset) (Janet Eisenhauer, University of Wisconsin)
- Comparative example: Variables reflecting language and geography differences (Meinhard Moschner, Zentralarchiv, and Peter Granda, ICPSR)
- Hierarchical data example (fixed format file with multiple record segments such as household, family, person records with a clear linking variable, like the Current Population Surveys) (still need volunteers)
- Instrument documentation example (Instrument Documentation Group)
- Aggregate data examples: 3 scenarios for aggregate data corresponding to the 3 different aggregate components (still need volunteers)

The group also discussed all the components that will be necessary for Version 3.0 Public Review (see the section of the minutes on the Tools Working Group for more detail on some of these items):

Basic components

- Separate page on DDI site for 3.0 review
- Schemas with internal documentation
- UML Diagrams
- Markup examples with descriptions; viewable in XML and through stylesheet
- Form to provide feedback -- written in DDI 3.0 to show its use; mechanism for others to send examples

Supporting materials

- Document on what and how to evaluate the specification with checklist of points to consider in the evaluation
- Supporting documentation describing modules
- Supporting documentation describing principles behind the structure
- Tutorials on XML Schema and UML
- Getting Started document
- Table mapping DDI 3.0 to other metadata standards -- Dublin Core, MARC, SDMX, ISO 11179, etc.
- List of mandatory elements
- Presentations

Tools

- DDI Version 3.0 Viewer/Stylesheet
- DDI Help page developed by Pascal
- Schema visualization tools by Achim (XMLDOC, XMLSpy) and Arofan (HTML, sequential for printing)
- List of recommended DDI Core (Lite) elements
- Templates (DDI core, geography, grouping)

Longer-term tools and projects

- DDI to SPSS/SAS and SPSS/SAS to DDI conversion tools
- DDI Tools Forum for sharing and discussing tools
- DDI-aware editor with graphical user interface for people with non-technical background
- Tool with GUI for creation and organizing groups
- Sophisticated report and search engines
- DDI Registry
- Ultimately, integration with statistical software packages and Blaise

The next meeting of the Alliance will be held on May 19, 2007, in Montreal after the IASSIST meeting. At that time we will review the feedback received on the specification, look at progress towards Version 3.1, and undertake a publicity and awareness campaign.

General Discussion

The remainder of the meeting was devoted to questions fielded by Wendy Thomas, Chair of the Structural Reform Group (SRG), the Alliance's technical committee. Wendy and Achim also presented some materials she had prepared earlier for other meetings.

Variable structure
Wendy emphasized that the structure of the variable has changed significantly across DDI Versions 1, 2 and 3. With the new lifecycle model, it is now possible to begin with a concept, operationalize the concept into a question, set up response categories, create a variable, and build a physical instance. While there is more complexity in Version 3.0, there is also more potential to express relationships in the data never before possible. In terms of where to begin with markup, it is now the case that one builds a new variable from the bottom up, starting with categories.

Grouping
Comparison across variables is by definition and not by label since the same label may express different definitions. Grouping provides needed new functionality. Object-oriented programming provided the model behind the grouping model. The specification is designed to capture comparison by intent in the metadata; for comparison after the fact, we need to capture the researcher's methodology. Right now, grouping is defined as using two or more study units; a single study unit implies a simple instance rather than a complex one.

Maintainability
In DDI 3 there is a new emphasis on maintainability. Concepts, questions, and variables can be maintainable objects which can exist outside of a DDI instance and be drawn in full or in part into other instances. This means that such objects must be identified and controlled strongly in order to exist on their own.

Data typing and controlled vocabularies
In terms of data typing, one declares a type of data (date, for example), and then software will "expect" content in that format. We need controlled vocabularies for many of these typed elements and will solicit volunteers to work on these vocabularies.

Migration from 2 to 3
The user doesn't have to make this conversion, and in fact there are situations in which one may want to convert from 3.0 back to 2.x if one has built a system based on 2.x. To convert, it will be relatively easy to do a basic migration of the straightforward elements, but there are some that will require human decisions.

Geography
Geography has more detail in Version 3.0 and we have expanded our ability to reference external maps and shapes. Again, from DDI Version 1.0 to 3.0, there has been a lot of change and innovation. In Version 1.0 we basically had only the nation, geographic coverage, and geographic unit elements. Now we have bounding boxes and polygons and have made geography available to every module. It will be easier for the end user to understand what is being described. Our metadata will inform GIS software what to search for.

Versioning
Versioning is handled differently in DDI 3.0 and a comment was raised that we need to be clear that we can version different iterations of the datasets described by the DDI instances, not just the XML instances themselves. Each time a physical data file changes, the physical instance changes -- there is a one to one relationship between physical instance and the data file. We need to be clear that we can document the history of version changes. We may need to document versions of the instrument as well.

Tuesday, October 17

Tuesday's meeting opened with a demonstration by Sigbjorn Revheim (Norwegian Data Archive) of the newest version of the Nesstar Publisher, which has been developed for the International Household Survey

Network Microdata Toolkit but which will be rolled out as a general upgrade soon. This new version has several new features:

- Documentation of more than one study with a single study description
- Many possibilities for templates with metadata about them, which can be customized
- Inclusion of Dublin Core
- Integration of the hierarchy builder into the main program (structure similar to a relational database with checks that merging files is possible)
- New validation functions
- Export of documentation into PDF
- Translation of elements into other languages
- User interface in other languages

Still to come is the ability to document additional cube functionality.

Three working groups then met to discuss issues specific to their groups.

Tools Working Group

(Jan Goebel, Hans Jorgen Marker, Sigbjorn Revheim, Pilar Rey del Castillo, Dan Smith, Mary Vardigan, Joachim Wackerow)

It was suggested that we create a new tab on the DDI Web site to be called DDI Version 3.0 Public Review. This would be the site where all the information about evaluating Version 3.0 would go. It was thought that giving out the URL ddialliance.org would be advisable, so ICPSR will try to move quickly on this. The group discussed the audience for the Public Review. It will be technical people but also others who may not have the same technical expertise. It is really easy to overwhelm people, so we want to make sure that the Version 3.0 Web page is organized well.

We want to encourage a broad review of Version 3.0. To that end, we will provide a checklist of points to consider. We want to encourage people with new perspectives to think about this in a larger way, but if we are too open we may not get the responses we want. It was suggested that we create an HTML form to capture feedback and that the form itself be marked up in DDI 3 to demonstrate a use of the standard.

We also need a good document on how to get started because this is always difficult for people. We want our set of examples to be viewable as raw XML and through a stylesheet. For visualization of the schema itself we currently have these options:

- Pascal's DDI Help page
- Documentation that Achim will create using XSDDoc and XMLSpy
- Arofan's tool for printing HTML sequentially

There is a documentation group (Ann Green, Ilona Einowski, Sanda Ionescu, Mary Vardigan) working to review the internal documentation and external supporting materials. We are looking for a standard reference work on which to base our definitions of core social science concepts like sampling, universe, etc. Suggestions were the ISI site, the NSDStat help documentation to which several experts contributed, and the glossary that ICPSR points to (compiled by Jim Jacobs, University of California, San Diego). We will put out a call to the group for other suggestions.

We also want to have a 2-3 migration tool. This will most likely be basic, covering only the easily mapped fields, and then listing the fields about which users will have to make decisions. Given the timetable, it isn't likely that we will have time to develop a truly interactive tool. An SPSS/SAS converter is likely to come later, not by Public Review.

Templates will be very important. These will be raw XML with dummy content that can be replaced by the user to create valid instances. We will prepare templates for DDI Core fields (formerly DDI Lite), geography, grouping, etc. It was thought that "DDI Core" terminology is preferable to the "Lite" terminology.

We also want links to tutorials on UML and XML Schemas. Also suggested was a mapping tool to cover mappings from DDI Version 3.0 to DDI 2.0, MARC, Dublin Core, ISO 11179, Triple S, SDMX, etc.

A list of mandatory elements will also be good to have. Sigbjorn offered to create a demo of exporting the Nesstar Publisher into 3.0. We also want to have Wendy's useful PowerPoint presentations and presentations like Achim's grouping presentation from IASSIST.

The Tools group spent some time talking about the content and management of a DDI Registry that we might create at some point in the future. This could have DDI renderings of standards such as occupation codes, geography codes, employment codes, education codes, etc. There could be national equivalents of these code schemes. We might also have pointers to different thesauri. Components of the registry would be externally independent and would have to be maintained by a trustworthy organization so as to ensure their longterm availability. Managing the registry itself can be complicated.

We may also want a Web forum in addition to the DDI Alliance mailing lists. One such forum is PHPBB.com, which is free. We need a way for others to send us examples.

Survey Documentation and Implementation (SDI) Working Group

(Karl Dinkelmann, Janet Harkness, Peter Granda, Jeremy Iverson, Peter Mohler, Beth-Ellen Pennell, Wendy Thomas)

This new working group held its initial meeting, during which they outlined the scope and substance of their topic, their timetable, and the expertise and support they need going forward. Beth-Ellen Pennell, Director of Survey Research Operations (SRO) in the Institute for Social Research at the University of Michigan, reported that several members of the group had met in 2005 in Edinburgh with members of the DDI Comparative Group and the SRG. The SDI work actually overlaps with the tools group, since SRO has created a tool, and with the comparative group, because the impetus for their work was comparative surveys. The tool that SRO has created is software for the ongoing monitoring of survey processing from study design to data delivery. This is in the form of a Web survey with a variety of modules to be filled out by experts involved in the data collection process. Modules include:

- General project information
- Ethics review at the national level
- Sampling
- Questionnaire development
- Translation
- System development
- Interviewer recruitment
- Pretesting
- Data collection
- Quality control
- Dataset preparation

They have used the tool successfully for a World Mental Health study by the University of Michigan, WHO, and Harvard, administered in 28 countries. They want to modify the tool to make it compatible with DDI. Process data, or paradata, is becoming more important itself: data collectors want to look at call records, time of day, etc., and are emphasizing responsive design to ensure flexibility.

The SDI group will prepare a proposal along the lines suggested by the SRG in spreadsheet form. They plan three face-to-face meetings in January, February, and March, with the hope that they will have a recommendation by April. This will be part of Version 3.1. They will put out a call to the DDI group for more volunteers for their group. Wendy will be the SRG contact.

Comparative Data Working Group

(Frederic Berger, Ken Miller, Meinhard Moschner, Reto Hadorn, Iris Alfredsson)

The group agreed that Meinhard would take over management of the group at this point. They first want to evaluate the comparative module of Version 3.0 and set up a specific framework for selecting test materials. For example, they want to look at Version 3.0 in the context of single datasets and cumulative

integrated datasets. They want to look at a simple variable like age, a more complicated variable like vote intention with country-specific political parties, and harmonized variables. They want to review the specification across types of languages and cultures.

For the comparative example that Meinhard and Peter are building, they will likely use the ISSP. Achim will be the contact person in the SRG. They plan to create the example in December.