Expert Committee Meeting

October 12-13, 2003
Ann Arbor, Michigan

Present:
Tom Piazza, Chair (University of California - Berkeley, Computer-Assisted Survey Methods Program); Arofan Gregory, XML Consultant (AEON Consulting); Atle Alvheim (Norwegian Social Science Data Services [NSD]); Bill Bradley (Health Canada); Pat Doyle (U.S. Census Bureau, Demographic Surveys Division); Andrew Dzhigo (Princeton University Library); Ilona Einowski (University of California - Berkeley, UCDATA Archive); Fred Gey (University of California - Berkeley, UCDATA Archive); Pascal Heus (World Bank); James Jacobs (University of California, San Diego); Peter Joftis (University of Michigan, ICPSR); Ryan Johnson (Washington State University); Julie Linden (Yale University, Social Science Libraries & Information Services); Margaret Low (California Digital Library, University of California); Marc G. Maynard (University of Connecticut, Roper Center); Meinhard Moschner (Zentralarchiv, Cologne); Ron Nakao (Stanford University); Pilar Rey del Castillo (Centro De Investigaciones Sociologicas); Jostein Ryssevik (Nesstar Ltd); Janet M. Eisenhauer Smith (University of Wisconsin - Madison); Jeanne Spicer (Pennsylvania State University, Social Science Research Institute); Kevin Schurer, for Ken Miller (UK Data Archive); Wendy L. Thomas (University of Minnesota, Minnesota Population Center); Mary Vardigan (ICPSR); Oliver Watteler (Zentralarchiv, Cologne)

Also attending as observers:
Cavan Capps (U.S. Census Bureau); Mark Diggory (Harvard-MIT Data Center); Ann Green, Steering Committee (Yale University); Bjorn Henrichsen, Steering Committee (Norwegian Social Science Data Service); Dag Kiberg (Norwegian Social Science Data Service); Richard Rockwell, Steering Committee (Roper Center); Ed Thomson (Health Canada); Myron Gutmann, Steering Committee (ICPSR); Sanda Ionescu (ICPSR); I-Lin Kuo (ICPSR); Matthew Richardson (ICPSR)

Introductions and Procedural Issues

Interim Chair Tom Piazza opened the meeting and welcomed participants to the first meeting of the DDI Alliance Expert Committee. Steering Committee members Myron Gutmann, Bjorn Henrichsen, and Richard Rockwell offered additional welcoming comments and emphasized the importance of the work of the Expert Committee. Participants introduced themselves, described their interest in and use of the DDI, and discussed their expectations for the Alliance.

Tom briefly discussed the fact that the Expert Committee is a large group and will be most effective if it can form working groups to focus on specific tasks. Also, according to the Bylaws, the group must elect (1) a chair to head the Committee and attend meetings of the Steering Committee, and (2) a second representative to the Steering Committee meetings. Tom indicated that elections would take place later during the meeting.

A discussion of how the Committee should communicate took place. Virtual office software was raised as a possibility, as were listservs and bulletin boards, like the ezboard communications software that the group has been testing.

Data Model

After a brief presentation on the history of the DDI effort and a summing up of where the effort stands, there was discussion of a data model for the DDI. It is generally agreed that the XML Document Type Definition (DTD) for the DDI has limitations: it is not as modular and easily extensible as it should be and it has not been thoroughly reviewed for internal logic. The Committee needs to develop a data model, most likely in Universal Markup Language (UML), to reflect the underlying design and structure of the specification. Once the data model is in place, the DDI can be expressed as XML Schema, RDF, a DTD, or possibly other formats.

The Health Canada/Nesstar partnership has already done some work on a data model for the DAIS/nesstar software that harmonizes the DDI with ISO 11179. In the ISO standard, the notion of "concept" is central; ISO 11179 is also strong on administration and informs a number of metadata repositories on the Web.

There are clear differences between the DDI and ISO 11179, but they can be harmonized and can enrich one another. As developers of the DDI specification, we need to determine how we can learn from the ISO 11179 approach and how we can create interoperability with the ISO standard so that data that are ISO 11179-compliant can migrate into DDI.

Currently, the ISO 11179 standard is just a UML model, and there is no other representation. It basically lifts the data element, or variable, out of the study context that archives are familiar with. There is strength in this approach: it may help us to solve the problems inherent in documenting series and longitudinal data. We may decide that the new DDI data model should also focus on the variable and should in effect break the link between the variable and the survey. We need to recognize, though, that there are serious research issues in comparing variables across studies; one needs to know with certainty that variables are comparable in terms of sampling frame and other methodological issues.

In thinking about the aggregate extension recently added to the DDI specification in Version 2.0, it is clear that the logic of the "nCubes" is working but that the extension is positioned wrongly in the larger DDI structure. These sorts of issues can be remedied through the construction of an accurate and well thought out data model. The future of the DDI may involve a UML model and Schemas for separate modules. The Committee was encouraged to describe the process that we want to document first and then to construct a model based on the process.

We cannot abandon the DTD since a lot of markup has been done according to Versions 1 and 2. We probably need to proceed on parallel tracks, moving the DTD along from Version 2.0 to 2.x at the same time that we begin to create a modular Version 3 based on the new data model.

SDMX (Statistical Data and Metadata Exchange)

Arofan Gregory, a consultant with expertise in XML, apprised the group of another initiative he is involved in called SDMX. This is a project to develop an interchange format for time series data and metadata. In the SDMX model the data transfer format is separate from application-specific information, like OLAP cubes. The SDMX model is mainly about data; the metadata it carries is mostly about how the cubes were constructed. Thus, there seems a natural complementarity between the two efforts, and we should look for the points of intersection. Others on the Expert Committee also view this as a natural partnership.

XML Schema Version

We need to be thinking of a master plan with goals and a strict timetable to inform our work. The MetaDater and MADIERA projects in Europe are designed to be compatible with the DDI model, so we need to interface with those efforts. We need to think about moving the DTD to a Schema to take advantage of the modularity in Schemas, the capability for local extension, and the flexibility of namespaces. If we stay with the rigid DTD, any extension will break the standard, which is not the case in Schemas. With Schemas it is possible to control the amount of extensibility permitted. The METS Schema also has potential for the DDI project. METS is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, and some members of the Committee have used the METS Schema and inserted a DDI namespace. Namespaces lets the user control ownership and versioning.

ICPSR and Harvard-MIT Data Center have been working on a Schema version of the DDI that incorporates all of the documentation found in the Tag Library as well as the DTD comments.

Working Groups

It was decided to set up two major working groups, with subgroups:

- Structural Reform Working Group

- Substantive Content Working Group, broken out into Group 1: Aggregate Data, Geography & Time; Group 2: Comparative Data/Families of Datasets; Group 3: Complex Files; Group 4: Instrument Documentation
- Usability and Outreach Working Group

The Structural Reform Working Group will take on the task of "schematizing" Version 2.0 of the DTD. Substantive proposals can be fed to the Structural Reform group, but we first need to set up some style guidelines for the architecture of the proposals.

We also need to bring the DDI into XML.org and OASIS. This can be a task for the Usability and Outreach group.

There is some funding from the Alliance for meetings of the working groups, either face-to-face or telephone conferences. It should be noted that the XML specification was developed without any face-to-face meetings. ICPSR will explore the cost of telephone conferencing. Another useful tool for communications is the WIKI, which ICPSR also has experience with. MhonArc, a threaded email list, is another tool that the DDI committee used previously.

Elections

The Expert Committee Chair and a second representative to the Steering Committee will have voting privileges in the Steering Committee, which has budgetary and oversight responsibilities for the Alliance. It was decided that the two elected representatives should have staggered terms, with the Chair elected for a two-year term and the second representative, who could also function as a Vice-Chair, for a one-year term, with the following term lasting two years. Nothing precludes nomination of the same person after the first term is completed.

Tom Piazza was elected as the Chair of the Committee and Hans Jorgen Marker of the Danish Data Archive (not present) was elected as the second representative to the Steering Committee.

Working Groups Structure

The following Working Groups were established:

Structural Reform Working Group:

- Cavan Capps - Census Bureau
- Mark Diggory - Harvard-MIT Data Center
- Andrew Dzhigo - Princeton Library
- Arofan Gregory - AEON Consulting
- Pascal Heus - World Bank
- I-Lin Kuo - ICPSR
- Pilar Rey del Castillo - CIS, Spain
- Jostein Ryssevik - Nesstar Ltd (Chair)
- Wendy Thomas - Minnesota Population Center (Vice Chair)

Substantive Content Working Group:

- Atle Alvheim - Norwegian Social Science Data Service
- Pat Doyle - U.S. Census Bureau
- Ilona Einowski - University of California, Berkeley - UCDATA (Chair)
- Janet Eisenhauer - University of Wisconsin
- Fred Gey - University of California, Berkeley - UCDATA
- Peter Granda - ICPSR
- Ryan Johnson - Washington State University
- Julie Linden - Yale University
- Margaret Low - California Digital Library
- Mark Maynard - Roper Center
- Meinhard Moschner - Zentralarchiv

- Tom Piazza - University of California, Berkeley - CSM
- Wendy Thomas - University of Minnesota
- Ed Thomson - Health Canada
- Oliver Watteler - Zentralarchiv

Subgroup 1: Aggregate Data, Geography & Time

- Atle Alvheim
- Ilona Einowski
- Fred Gey
- Peter Granda
- Julie Linden
- Margaret Low (Chair)
- Wendy Thomas
- Ed Thomson

Subgroup 2: Comparative Data/Families of Datasets

- Atle Alvheim
- Ryan Johnson
- Mark Maynard
- Meinhard Moschner
- Wendy Thomas
- Oliver Watteler (Chair)

Subgroup 3: Complex Files

- Pat Doyle (Chair)
- Janet Eisenhauer - University of Wisconsin
- Peter Joftis
- Tom Piazza

Subgroup 4: Instrument Documentation

- Pat Doyle (Chair)
- Tom Piazza

Usability and Outreach Working Group

- Bill Bradley - Health Canada
- Pat Doyle - Census Bureau
- Janet Eisenhauer - University of Wisconsin
- Pascal Heus - World Bank
- Sanda Ionescu - ICPSR
- Jim Jacobs - University of California, San Diego
- Ryan Johnson - Washington State University
- Matthew Richardson - ICPSR (Chair)
- Jeanne Spicer - Pennsylvania State University
- Mary Vardigan - ICPSR

Each group needs to determine how to communicate and meet and the frequency of their contacts.

Deliverables

It was decided that there would be two short-term deliverables:

- Version 2.0 Schema released and fully documented by the end of November
- Six months later, another version: an interim alpha version that would incorporate information from the substantive content groups and would be leaning toward the modular Version 3

In addition, the Structural Reform group will recommend formats for substantive proposals by the end of November.

The issue of whether the DDI should deal with access conditions was raised. Four or five years ago, it was decided not to include much detail on access conditions in the specification because the assumption was that this was an application issue. We should probably revisit this issue again. XML security and access systems have been developed that we might be able to link to or adopt.

Next Meeting

We have tentatively reserved Saturday, May 29, 2004 -- the day after the IASSIST meeting in Madison Wisconsin -- for the next Expert Committee meeting. This is Memorial Day weekend, however, and it is not certain how many can attend. If Working Groups want to get together at IASSIST, they should feel free to do so and should contact ICPSR to get meetings set up. More information will be forthcoming on the next meeting of the full Committee.