# Committee Meeting Minutes

January 14 and 15, 2002
Washington, DC

Committee Members Present: M. Shanks, Chair, M. Altman, G. Blank, E. Boyko, B. Bradley, C. Capps, C. Dippo, P. Doyle, D. Gillman, P. Granda, A. Green, Mike Haarman, Bjorn Henrichsen, P. Joftis, T. Piazza, R. Rockwell, J. Ryssevik, W. Thomas, M. Vardigan

Other Meeting Participants: S. Ferg, BLS; C. Freas, Georgetown University; D. Geraci, UKDA; C. Oster, Health Canada; R. Wilson, ICPSR

Monday, January 14, 2002

Brief Funding Update

Richard Rockwell thanked Bill Bradley and Health Canada for providing bridging funding for the DDI project at a critical time in the DDI's evolution. He then related recent developments in the funding situation. A joint proposal between ICPSR and the Roper Center was submitted to the National Science Foundation in August 2001. The proposal was reviewed as worthwhile and fundable but, as a result of NSF budget constraints, was funded at only a minimal level. However, this funding is seen as an endorsement of the DDI project and will keep the DDI alive for another year. We need to think ahead to submitting a larger proposal, possibly to the Computer and Information Science and Engineering (CISE) arm of NSF.

Hierarchical Data

The current specification is lacking the proper structure to document hierarchical and relational files adequately. Users need to be able to identify keys to use for merging and to pick up the right weights for different record types. Merge order is critical. It should be possible to make a few additions or changes to handle hierarchies properly. At the very least we need:

1. An indicator of record type, e.g., household, family, person, etc.
2. For each variable, the record group it belongs to
3. The key or CASEID for the highest level record

The physical and logical structures of files are not documented properly in the DDI. Currently, Section 3 of the DDI specification documents the

physical structure, Section 4 the logical. We need to be able to capture one-to-many relationships. The logical model is too primitive to describe hierarchies well, and a semantic model should be developed.

One approach would be to start at the highest level with an entity relationship model. This would allow us to create an ideal implementation, which we could then map to an actual physical implementation. This is a major activity envisioned for DDI Version 2.

A suggestion was made to add additional attributes to the record element to indicate primary and foreign keys. We need to be able to equate two types of information on different record types.

Aggregate/Tabular Data

Wendy Thomas will continue to work on documenting the aggregate data extension in the Tag Library.

The current specification is lacking a controlled and precise way to indicate specific types of derivations. We may need a new attribute for this. It is not clear, however, whether this can be achieved in such a way as to make it machine-understandable. Some calculations are so complex that human intervention will always be required. This may be a good application for MathML. Can we define classes that derivations might fall into?

We also need a presentation layer or a default view of a cube - a model to describe how a data cube should be displayed to be understood by humans. We need to be able to describe how cubes can be combined.

Another issue is that there are currently two ways of specifying a hierarchy of categories. We either need to provide only one way of doing this or else we need to strongly encourage only one way of doing this. This is necessary for telescoping categories like occupation and disease codes. Category hierarchies can be accomplished by nesting categories (catgry is now a recursive element) or by using category groups. Neither mechanism is perfect as it is currently structured. There is no level name or number, for example.

The category hierarchies should not be different for microdata and for aggregate data. Perhaps we should keep both options and just specify the nested categories option as the preferred method in a best practice document. This is a microdata problem that becomes apparent in a data cube. We need to be able to drill down and roll up.

Other suggestions relating to the timeDmns, measure, and aggrMeth elements were made, along with suggestions for linking to maps.

It is clear that we are not yet ready to freeze the data cube model; more testing needs to occur first.

Weights in the DTD

We need the capability to specify that a variable can have more than one weight. We also want to be consistent with weight-related elements for catStat and sumStat. This issue is not clear and needs to be pinned down via email.

Tuesday, January 15, 2002

Funding

The next meeting of the DDI group will be in Storrs, CT, at the time of the IASSIST meeting in June. Members will be traveling on their own funding.

We need to begin to generate some sustaining funding for DDI activities. The NSF grant will permit us to get by, but we will have to prioritize the proposed tasks since we cannot accomplish them all with the limited funding awarded.

Health Canada can contribute funding for working groups if the meetings occur before March 31, 2002.

Given the funding situation, this may be a time to explore an alternative arrangement to generate core funding. We could form a consortium with each institution paying a fee to send a representative to the meetings; each paying member would have one vote. There would be a small core budget supplemented by external grants focused on specific projects. The institutional fees would not cover travel. We would probably invite members from the private sector, especially survey research firms, statistical software companies, markup software companies, etc. We need to have rules about how many units from a large institutions can join. It was suggested that ICPSR be the host institution initially.

This proposal will be on the table for June. Before then, we need to look at other consortia and their bylaws, particularly organizations that are producing a product.

Operating as a consortium would enable the DDI Committee to function more like other standards bodies. We need a membership structure that

permits decisions to be made efficiently, more like the W3C model. W3C has a specified route that all versions of their standards take: implementation stage, requests for comments, tentative recommendation, etc.

Such a membership model might raise the visibility of the DDI effort. If it were formalized and advertised well, we might get corporate support from agencies like the Census Bureau.

The consortium would have a budget approval process. Dues could be used for publicity, brochures, the Web site, training for Committee members, representing the organization, administrative staff, etc.

Committee members need to take this idea back to their home institutions and discuss it over the next few weeks.

Harmonization with ISO 11179

An extension to ISO 11179 for statistical data was developed at the Census Bureau by Dan Gillman. This extension is called the Census Metadata Repository (CMR). If we make the DDI map semantically to the CMR, then the DDI can be viewed as compliant with ISO 11179.

A key difference in the standards is that the ISO 11179 has nothing at the study or dataset level. The CMR adds some description at this level. Basically, ISO 11179 has a concept level above the variable level, and variables are representations of the concept. The DDI is weak in administrative elements and control of metadata collections. ISO 11179 is weak in providing access. We could add an administrative table to the DDI to associate data objects with this administrative piece. It is not clear whether this should become part of the DDI or whether we can just provide a linkage. Administrative information would have to be updated, making this cumbersome if it were incorporated into the DTD. Resolving this may require modularization that would come with DDI 2.

Fast-tracking the DDI for Accreditation

The issue of fast-tracking the DDI in the ISO context was raised. We could pursue the path of making it a Publicly Accessible Standard. What would be involved would be putting the DDI specification into the ISO template and writing an accompanying document that would "pass muster." The benefit of doing this would be getting formal acceptance; there is a "halo effect" from compliance with any recognized standards. The standard should be fairly stable before we embark on this path. There is a review of standards every five years, but incremental changes

can take place in the interim.

These issues will be raised again in June.

DTD Issues

Mike Haarman suggested that it makes sense to register the standard with the OASIS XML registry and to apply for a Formal Public Identifier (FPI). He also proposed that the CALS table model and the TEI declarations be formulated with regard to definitive, network-accessible versions for these resources so that local copies are not required. These are simple syntactical changes posing no backward compatibility problems.

Local formatting can be solved in a couple of different ways. We can turn formatting on/off on an element-by-element basis for any element that is parseable text. We could use the TEI content model or perhaps XHTML. With the latter option, performance would be impacted because the parser needs to read the XHTML DTD for every instance.

This discussion needs to continue via email.

XML Schema/RDF

Mike will prepare a stub schema definition for the DTD, and we can experiment with this now. The advantage of XML Schema is that it enforces data typing, meaning that, for example, the content of a date element can be disciplined so that only a date in ISO format will be accepted. It was noted that data typing poses problems for archivists who receive documentation in a certain form and then have to document what they have.

Schemas are the direct descendants of DTDs and will fully supplant DTDs. All parsing tools now support schemas, and there are markup authoring tools that support schemas as well. The mechanism for local formatting in schemas is different, though.

RDF deals with the semantic side of things, whereas schemas/DTDs deal with structure, so the two formats are complementary. RDF has been patented, however, and it is not clear what the implications of this are for our activity.

CESSDA DDI Report

The CESSDA working Group met in September at the UK Data Archive. The CESSDA Integrated Data Catalog is being moved from WAIS to

NESSTAR and will be DDI-compliant. The Working Group defined a minimum DDI element set to put into English to make cross-archives substantive searches possible. In March there will be a meeting in Amsterdam about controlled vocabularies.

The nation element should probably be replaced with a country element, but this will not be possible in the current version.

Should all Dublin Core elements be mandatory? It is possible we could have conformance levels: Level 1 might have only the title, and Level 2 might have a larger set of elements.

At the next DDI meeting, we should make sure that the CESSDA and DDI Working Groups are in agreement. The DDI was created purposefully to be loose so groups like CESSDA can do what they are doing.

Bjorn Henrichsen described three major proposals in which the DDI plays an important role, all of which were just submitted to the EU.

Geography

Ron Wilson reported that the National Institute of Justice has announced that any datasets funded by them must comply with FGDC standards. More and more data producers are attempting to provide precise geographic information without compromising confidentiality.

It should be possible to reference external entities describing geographic schemes to get the precision in geography that we need. References to points like coordinates and addresses can be handled with the geographic unit element. We may need a URI attribute to map to.

Web-based Version of Health Canada DAIS Software

Bill Bradley and Jostein Ryssevik demonstrated a prototype for the new version of the DAIS software that drives the Health Canada data analysis system. The XML editor takes various SAS and SPSS files as input and exports DDI-compliant files. Users can click on other elements and add them easily to the marked-up files. Several Committee members expressed interested in getting copies of the software.

Next Meetings

The next meeting of the DDI Committee is tentatively scheduled for Saturday, June 15, in Storrs, CT, right after the IASSIST meeting.

Several working groups were constituted to reflect members' priorities for next steps in developing Version 1.1 of the DTD. They are as follows:

- ISO issues: Bill Bradley (leader), harmonization; Dan Gillman (leader), accreditation
- Aggregate data issues: Wendy Thomas (leader), Cavan Capps, Jostein Ryssevik, Ann Green, Bjorn Henrichsen
- Hierarchical data issues: Tom Piazza (leader), Grant Blank, Jostein Ryssevik, Pat Doyle
- Geography issues: Ron Wilson (leader), Bjorn Henrichsen, Cavan Capps, Peter Granda, Wendy Thomas, Mike Haarman

Group leaders will decide when to meet in person, if necessary. There is financial support for meetings from Health Canada if the groups meet before March 31. The European members suggested that work be accomplished via email to reduce the number of trips they have to make. It is possible that the aggregate and geography groups could meet together.