Committee Meeting Minutes

Friday, February 7, 2003
Washington, DC

Present: Grant Blank, Ernie Boyko, Bill Bradley, Cavan Capps, Pat Doyle, Diane Geraci, Dan Gillman, Ann Green, Bjorn Henrichsen (Chair), Peter Joftis, Ross Marshall, Marc Maynard, Ekkehard Mochmann, Tom Piazza, Richard Rockwell, Jostein Ryssevik, Wendy Thomas, Mary Vardigan, Ron Wilson

Ekkehard Mochmann from the Zentralarchiv in Cologne, a member of the DDI Steering Committee, was welcomed to the meeting as was Diane Geraci, sitting in for Ken Miller from the UK Data Archive. The Chair acknowledged that this was the last meeting of the original DDI Committee, and the Committee was thanked for their years of service and dedication to the effort.

Aggregate/Tabular Data Extension

Since it was formed, the Aggregate/Geography Working Group had been grappling with several important issues, including:

- What belongs in the DDI and what should be specific to an application
- How to coordinate with other standards
- How to make the specification machine-processible, not just machine-readable

A list of recommendations regarding the aggregate/tabular extension and geography in the DTD was circulated by Wendy Thomas prior to the meeting, and the Committee was asked to consider them one by one:

Recommendations

1. Add new attributes of temporal (y/N), geog (y/N), geoVocab, and catQnty to Variable 4.3 and remove 4.3.12 timeDmns. The attributes temporal and geog permit the markup author to flag a variable as a time or geography variable. In the case of a geographic variable, geoVocab indicates the vocabulary used. It was pointed out that time has distinct properties that are different from other dimensions; one can't aggregate time but instead must flag it. We can extend the way we handle time at a later stage and may even want to form a separate Working Group to concentrate

on time-related issues. At this point, the additions are processing attributes rather than content standards with controlled vocabularies. The CatQnty attribute indicates the number of categories found on the variable and is useful in determining how granular categories are and in facilitating machine processing. The addition of catQnty was one of the recommendations of a small group of DDI users (representatives of ICPSR, University of Minnesota, and California Digital Libraries) who had met earlier in 2003 in Berkeley, CA, about DDI-related issues. These recommendations for new attributes of Variable were approved by the DDI Committee.

2 Remove the recursive nested category feature in 4.3.18.5, add to Category Group 4.3.17 attributes of "levelno", "levelnm", "completeness (true/false)", and "exclusivity (true/false)", and add to Category 4.3.18 the "exclusivity" attribute, removing from Category the attributes "other" and "total", which were part of the nested scheme. There are two ways of creating an aggregation hierarchy: through recursive nesting or through category groups. There are problems with each method, but the category groups approach has proven superior. It is more flexible and permits an unordered incomplete hierarchy. The nested category feature was part of the original aggregate recommendation but should be removed from the final version of the specification. The new category group attributes will enable the description of the logic of a hierarchy. Without these additions we can't do on-the-fly aggregations. Levels are needed for nesting order. The DDI specification should allow the tagging of existing tables and the ability to use OLAP cubes. Also, we need to treat Category and Category Groups in a more parallel fashion. These recommendations were approved.
Regarding hierarchies, it was also suggested that we investigate an attribute on Variable similar to geog and temporal that would flag a variable as containing some sort of hierarchical coding list, but no final action was taken on this.

3 Add new elements from MathML called "mrow" and "mi" (under mrow) to Category 4.3.18 with an IDREF, "varRef", on mi. This new structure basically describes rules for creating a concatenated key; it enables one to string characters together and to treat that string as a single unit. The need for concatenation is particularly salient for geography but also applies to complex data files in

which one may want to create a unique record identifier through concatenation. The proposal is to borrow these two tags from MathML now and to investigate MathML and OpenMath in greater depth in the future to see what benefits might accrue from their use in describing derivations. It was suggested that the same objectives might be accomplished through the use of namespaces, but since an XML schema is needed to use namespaces, that suggestion was tabled. It was noted that we may also need a Working Group to focus on issues related to derivation. This recommendation was approved.

4 A new geographic scheme was advanced during the meeting, but the Committee recommended that the group working on geography consult relevant standards including ISO to develop the recommendation further to accord with other standards. The following is the recommendation agreed to via email after the meeting:
Add to Geographic Coverage 2.2.3.4 three new attributes: "geotype" to indicate if the geographic locations identified in the dataset are points, line strings (e.g., streets), or polygons (e.g., states, tracts, countries, etc.); "geoVocab" to indicate for discovery purposes the geographic coding schemes used; and "georef" to link to the variable carrying the base level of geographic information in the file, e.g., Summary Level in the U.S. Census. In addition, add two new elements at the same level as Geographic Coverage: Geographic Bounding Box (optional and non-repeatable), with four sub-elements -- West, East, South, and North Bounding Latitude -- and Geographic Bounding Polygon (optional and non-repeatable) with a sub-element of Polygon (mandatory and repeatable), a sub-element of Polygon called Point (mandatory and repeatable), and sub-elements of Point (mandatory and non-repeatable) called G-Ring Latitude and G-Ring Longitude.
The Bounding Box is the fundamental geometric description for any dataset that models geography and is intended for discovery purposes. It is the minimum box defined by the west and east longitudes and the north and south latitudes that includes the largest geographic extent of the dataset's geographic coverage. It is used in the first pass of a coordinate-based search undertaken using a geographic search tool. There was discussion about whether the bounding box should be included at the file or the study level and whether it could describe shape files. The question

was also raised that the box alone is not sufficient and that describing a more detailed polygon should also be possible. (The recommendation ultimately approved by email did include the detailed polygon.) The bounding box should also be capable of dealing with three dimensions.

5 While the Complex Files Group brought a tentative recommendation to the table, they decided that it was necessary to meet again and develop a new recommendation that would ideally be incorporated into the final version (2.0) of the DTD. NESSTAR has marked up complex files using variable groups, but there should be a more efficient method, which the Working Group will construct.

6 Add a new element called nCube Group (with nCube and nCubeGrp IDREFS) to describe published tables made up of multiple nCubes. This was another recommendation that resulted from the Berkeley meeting. Essentially, the addition of this new nCube Group element would create a structure parallel to Variable Group and Variables. This recommendation was approved.

7 Add the attributes "sdatrefs"and "country" to Category Label (Label is a generic element, A2). This was intended to cover instances of comparative data in which categories are specific to different geographic areas. This recommendation was approved.

8 Make Measure 4.4.14, which is part of the nCube structure, repeatable. This will enable the markup of a time-series database. This recommendation was approved.

9 A new scheme for including map references was suggested during the meeting, but the Committee recommended that Jostein and the group working on geography develop the recommendation further to accord with other standards. The following is the recommendation agreed to via email after the meeting:
Add an element under Variable 4.3 called Geographic Map with attributes of URI, mapformat, and levelno to facilitate linking to a map external to the DDI instance. This should be repeatable for levels of the geographic hierarchy. It was noted that there are substantial version issues associated with maps.

10 Accept the locMap 3.2 structure and implicitly the new aggregate/ tabular data extension (development Version 1.3) with the amendments just discussed and publish this as Version 2.0 by the

end of February 2003. This recommendation is limited to text file storage and it is understood that the model will not work for OLAP cubes or relational databases, which will be addressed at a later date. This fulfills the goal of the Roper/ICPSR grant from the National Science Foundation. This recommendation was approved.

Working Group on Standards

This group, consisting of Ann Green, Dan Gillman, Peter Joftis, Jostein Ryssevik, and Bill Bradley, has been working on an extension of the DDI to harmonize with ISO 11179. A semantic mapping and a partial data model have been developed. Bill hopes to publish a paper on this topic soon. The data model is incomplete but could serve as a starting point for the new Expert Committee as they begin their work. The whole issue needs deeper discussion, which was not possible at the meeting. Developing a data model is an activity that many on the Committee are interested in, and it seems a logical next step to formalize the model underlying the DDI, which would give us greater flexibility in expressing the model in different formats.

Alliance Organizational Documents

These documents are currently being reviewed by the University of Michigan to enable the Alliance to set up an administrative home in ICPSR. At this point we have heard from 21 potential members, each expressing certainty that their group will join.

We still hope to find external funding. It's possible that in the European Union's 6th Framework there could be a DDI-related funding.

Expert Committee

The Chair of this new group will be really important, and it was also suggested that we emphasize to the new group what we are trying to accomplish and who it is for. We need to talk about the physical versus the logical models, what is appropriately inside and outside the specification, and why it developed as it did. We might hold a panel discussion on the early life of the DDI. Merrill as the original chair should be invited.

This transitional phase from the original Committee to the Alliance is an important time. We need to ensure continuity and momentum. We should also consider the way the Expert Committee should be structured to work most efficiently. Should meeting time be spent mostly in Working

Groups with a final vote at the end? This is one model we might follow. Financing travel is another issue.

The CAI software houses should be invited to meetings, especially Blaise, which has given a commitment to DDI. MetaNet would also be a good group to engage.

The Committee was also apprised of the MetaDater effort in Europe to develop standards for the description of large-scale comparative surveys over space and time and to provide tools for metadata creation and management for such surveys. This will involve the creation of a data model for comparative surveys and will be compatible with the DDI. The project continues for three years.