DDI Alliance Expert Committee Meeting

IASSIST 2007 - Montreal, Quebec
May 19, 2007

Minutes

Members Present: Hans Jorgen Marker, Danish Data Archive (Chair); Ron Nakao, Stanford University (Vice Chair); Tuomas J. Alatera, Finnish Social Science Data Archive; Iris Alfredsson, Swedish Data Archive; Mark Diggory, MIT; Jan Goebel, SOEP-Berlin; Pascal Heus, World Bank; Graham Hughes, University of Surrey; Sanda Ionescu, ICPSR; Jannick Jensen, Danish Data Archive; Mari Kleemola, Finnish Social Science Data Archive; Rutger Kramer, DANS (Netherlands); Vigdis Kvalheim, Norwegian Social Science Data Services; Ken Miller, UK Data Archive; Meinhard Moschner, GESIS-ZA; Pilar Rey de Castillo, Centro De Investigaciones Sociologicas (Spain); Wendy Thomas, University of Minnesota; Mary Vardigan, ICPSR; Joachim Wackerow, GESIS-ZUMA; Marion Wittenberg, DANS (Netherlands).

Other Participants and Observers: Ernie Boyko, IASSIST; Jack Gager, XML Consultant, Metadata Technology; Dan Gillman, ISO 11179 Consultant, Bureau of Labor Statistics; Arofan Gregory, XML Consultant, Metadata Technology; Linda Harding-DeVries, Statistics Canada; John Ladds, Statistics Canada; Chris Nelson, UML Consultant, Metadata Technology; Jostein Ryssevik, Open Data Foundation; Michel Seguin, Statistics Canada.

DDI 3.0 Status Report and Discussion of Vote

Comments Related to Public Review

Arofan Gregory, XML consultant to the DDI Alliance, presented a summary report of the status of the Version 3.0 specification after the Public Review period, which ended April 14. He indicated that he was in the process of making changes to the specification based on information and comments received, and this new draft would be considered a "Candidate Release," a concept taken from the W3C.

About 100 comments were received during the public comment phase, most of which were submitted by individuals inside the DDI community. Several comments focused on consistency issues and minor changes, and there were comments about major features that the Alliance should consider. In general, the Structural Reform Group (SRG) wanted to avoid scope creep and to keep changes to a minimum, so no changes without solid use cases to support them were accepted during this round of comments. After the fact comparison was not finalized and no new controlled vocabularies were incorporated as a result of public review.

Changes implemented after Public Review:

- The *Identification and Referencing* features were modified so that the new specification is more aligned with the model. It will now be possible to make a distinction among maintainable, versionable, and identifiable classes.
- *Grouping, comparison, and inheritance* are solid in the new specification and multiple inheritance, which can be very complicated, was avoided. Support was added for machine-actionable recode descriptions; that is, one can now provide command lines to perform recodes. While the request was made that the DDI support comparison between any two objects, the

comparison feature as implemented remains focused on core items such as concept, categories, question, variable, and universe.

- The *archive module* was generalized so that an individual or organization having stewardship or control over a dataset at any point in the data life cycle may be considered an archive, whether or not it is a formal data archive like ICPSR.
- Additional fields were introduced to support *complex files* description, and this feature of the specification is now fully implemented.
- *Category statistics* were rationalized and support for crosstabulations was added.
- More detailed *universe descriptions* have been enabled to permit both machine-actionable and human-readable versions. It is now possible to have structured universe content that can be subset.
- Gaps in *field-level documentation* were filled in and text for user guide topics was drafted. Documentation will be improved during the Candidate Release period and there will be a hyperlinked version of the field-level documentation, which is currently only available in PDF form.
- Alignment with concepts, questions, and variables in the context of *ISO 11179* has been reconsidered based on consultations with experts.
- The *conceptual model* has also changed to reflect changes to the specification.

Candidate Release Process

The goal of moving to Candidate Release status is to provide a relatively stable version of the standard around which implementers can develop software and test for bugs. This is in response to concerns that not enough internal testing occurred during the Internal and Public Review periods. Any bugs discovered will be fixed with rapid turnaround, but changes will be limited to these kinds of revisions. After the implementation period, the Alliance will have a formal online vote to publish Version 3.0. The revised specification will be posted to SourceForge during the week of May 21. Examples will be modified accordingly as soon as possible.

Testing and Use Cases

Pascal Heus pointed out that use cases are critical at this stage in the development of DDI 3.0. We need to test the functionality of different life-cycle aspects of the specification as well as new features, and thus we need a set of real-world studies to review and to provide a roadmap to the technical group. These use cases will set the target for proof of concept and provide measurable results.

Suggested use cases, of which we need around 10-15, include:

- A demonstration that 3.0 does what 2.0 does. For this we may be able to get output from the IHSN toolkit.
- Instrumentation - 2000 Census questionnaire
- Concepts - Statistics Canada IMDB
- Comparison - GESIS-ZUMA
- Geography - IPUMS
- Complex Files
- Resource package
- Question bank

In June, the SRG will brainstorm and identify the best set of use cases for development. Producers of use cases could be:

- Highly committed volunteers
- Individuals for whom the use case answers the needs of a specific agency
- Students
- Consultants or agencies developing use cases under contract

The SRG can support those creating use cases because it is no longer in schema production mode. It may be useful to point out to implementers that in doing this work they can influence the final version of 3.0, so this may provide an incentive for implementers to come forward.

Vote to Approve DDI 3.0 Candidate Release

The group spent some time clarifying the substance of the vote. The vote was basically seen as a one-time process change to the Bylaws since the Bylaws do not include the Candidate Release mechanism. Getting consensus of the group to move forward was seen as crucial.

It was pointed out that during the Candidate Release period we need to be working toward proof of concept for the major features of 3.0. When all the use cases are supported, the Candidate Release will be ready for approval. A question was raised about what should happen if we do not reach the goal of having the use cases supported, and the response was that if there is no proof of concept yet, we cannot take a vote. Other questions about how to decide when it is time to vote and who will make the decision were raised as well. The Director will work in consultation with the SRG to make the determination.

Another question focused on how much is enough for a use case since there are no tools to do the markup easily. This is a chicken and egg situation at this point. We are asking for a vote of confidence, based on the small set of examples already produced, that larger, more comprehensive and representative examples will be possible. We may want to organize our use cases to align with the data life cycle. At this point we need to stop development and make sure what we have is good and usable.

In terms of a timetable, the committee wanted a time limit to be part of the resolution voted upon, and an outer bound of nine months was established. There may be EU funding to evaluate 3.0 starting in 2008, which would mesh well with this timetable. The wish list of use cases will be developed in June, and at the SRG meeting in October in Germany the group will look at the use cases developed by then to determine what else is needed and whether we are close to being ready for a final vote.

The following resolution passed unanimously:

Resolved: That DDI 3.0 be moved to Candidate Release status. A set of production use cases will be identified, involving each of the major features of DDI 3.0 throughout the lifecycle. These cases will result in documentation of each case and accompanying DDI 3.0 XML instances. In combination with tools implementation of the standard, these cases will be used to identify bugs and issues for fixing, and will also provide the basis for evaluation of the standard. No new features will be added during this period. A vote as called for by the existing bylaws will be taken at the end of the Candidate Release period, which will last no longer than 9 months.

Strategic Plan

The Expert Committee spent some time discussing its draft Strategic Plan, which lays out a vision for where the DDI Alliance is headed and discusses some tactics and plans for how to reach the desired goals. Specific objectives cover areas such as tools, outreach (to data producers, vendors of statistical analysis software, adopters, etc.), membership, training, funding, and ISO status. The plan also calls for the Alliance to use its monies strategically to create the sorts of products that it needs to be successful -- for example, documentation, a user manual, use cases, and promotional materials. A suggestion was made that a new section be added to the Plan to focus on continuing the further development and refinement of the standard for use with complex and longitudinal files.

An important undertaking for the Alliance is clarifying how the various metadata standards relate to one another. SDMX and DDI, for example, complement each other. It would be good to have mappings among all the major metadata standards and a document that guides people toward the best standards for their specific needs. It was suggested that users may want to wrap up DDI metadata in METS, an approach that the UK Data Archive and MIT are taking.

Promoting DDI

We need to begin to market DDI more aggressively by having a presence at major meetings. Meetings and conferences that are on the agenda of the Alliance include:

- A workshop in September 2007 in conjunction with the Eurostat meetings that will involve SDMX and DDI. Twenty-seven national statistical institutes will be in attendance in London for this, so it would be good to have some promotional materials ready at that time.
- In terms of promotion, the 25 IHSN countries are also using DDI and we should try to harness that group as promoters. There will be regional conferences to develop communities of practice.
- Wendy is on the program at the 2009 ISI conference in South Africa.
- Sponsored by GESIS-ZUMA, training for mostly European participants to learn about using DDI 3.0 will be held at the Dagstuhl conference center in Germany the last week in October 2007. An SRG meeting will also take place following the workshop.
- DDI will be on the agenda of the Comparative Survey Design and Implementation (CSDI) meeting in Berlin in June 2008. This may include a workshop and papers. Wendy and Achim have submitted a proposal for an invited paper.
- Coalition of Networked Information (CNI) meeting in Washington, DC, in December 2007 and the Digital Curation Conference around the same time. Pascal may be able to attend.
- E-Social Science, Ann Arbor, MI, October 7-9, 2007. We may be able to submit a short paper and poster sessions.
- OpenForum meeting on metadata registries in July 2007. Wendy Thomas is attending.

Working Groups

Tools Developers

The goal for tools is not so much a formalized DDI Tools Working Group but rather a community of developers. We need a Web page devoted to the tools effort that indicates which organization is working on tools, what the tools are, and who the contact is. We can link out to other sites as well. We need a mechanism for reporting.

The issue of branding of tools and their conformance to DDI was raised. Because most of these will be open source, they will be freely available for use and misuse. We cannot control how the tools are developed and used. Licensing of the schemas themselves to protect authorship and release from liability is another issue. We need guidelines on using the brand carefully. SDMX is a good example. It is freely available to all, but if one obtains it from ISO, there is a charge. We need IP language. The DDI Alliance is not a legal entity itself, so there is no one to own the IP. Because the business of the Alliance is run through ICPSR, ICPSR could be the target of legal action, however remote.

Structural Reform Group

This group needs to come up with a new name, since its current name is not really descriptive of its work going forward. The group will discuss a new name during their next conference call. The SRG is very open to new members.

Usability and Outreach Group

Now that Version 3.0 is beginning to be realized, the Usability and Outreach group must become much more active with regularly scheduled telephone calls. We need to create a campaign of awareness and adoption through a toolkit of outreach materials like PowerPoints, a professional brochure, and tutorials and training materials. We need to polish and package our products.

Instrument Documentation Group

As stated earlier, this group will become active again in August or September and is interested in taking on new members. It was reported that, as part of the work of the Instrument Working Group, Lon Hoffman from Blaise has done an evaluation of DDI 3.0 and has found that the handling of looping needs improvement. This will be followed up on when the group reconvenes.

Survey Design and Implementation Group

This group met the previous evening and will be reviewing the content of the Survey Design and Methodology System modules developed by GESIS-ZUMA and Survey Research Operations at the University of Michigan to document and monitor cross-national comparative studies. The group will begin to have regular phone calls.

Controlled Vocabulary Group

In developing the schema, the SRG created a rough list of areas in which controlled vocabularies are needed. This working group will cut across all substantive areas, so it will need to consult widely with a variety of individuals with the needed expertise. Ken Miller volunteered to chair the committee. There is the possibility of EU funding for this work in 2008. We may never achieve a complete set of values for controlled vocabularies, so we need a technical solution for extension that will not break the standard.

Working Group Practices

There has been a wide range of practice in terms of organization of the working groups, their outputs, and their communication with the SRG. Each group should have a clear agenda and scope statement as well as a feedback mechanism and a timetable. The SRG needs to understand how each working group fits into the model. The working groups should have members who have the time and knowledge to contribute to the groups. The SRG will be revisiting the submission process and will communicate its findings to each of the groups.

Tools Projects

Pascal highlighted several tools development projects centering on DDI 3.0 that are in the works:

- The International Household Survey Network (IHSN) Toolkit continues to develop its suite of products. The Toolkit includes an enhanced version of the Nesstar Publisher. Also being created are a bridge between CSPro, the data collection software used by the International Census projects, and DDI and a Stata plug-in for data disclosure/anonymization work. The Toolkit has just added a feature that makes it easy to implement a Web site for cataloging and searching studies. Quality assurance tools are also being built. The project is all open source except for Nesstar.

- The UKDA and the Open Data Foundation are creating data preservation and conversion tools through the Data Exchange Tools, or DeXt, project. The project is focused on producing a neutral storage format consisting of ASCII and DDI XML from which software-specific files can be generated. Work will be done between June and November 2007.
- The Canada Research Data Centres are working with the Open Data Foundation to create DDI 3.0-based tools for their system of secure data enclaves. There are 13 enclaves and the goal of the project is to strengthen the capacity of each. The project would first mark up metadata in DDI 2 and then convert to 3. Also envisioned are quality control tools, tools to manage concepts, and metadata mining tools. This is a four-year project starting in September. The project will provide back to the DDI Alliance what is missing in DDI 3.0.
- There is a potential three-year NSF-funded project with NORC, OdaF, and the San Diego Supercomputer to create enclave tools. The NORC Data Enclave is opening in June and will be using the IHSN Toolkit. If the project is funded, NORC will do a survey of users and SDSC will focus on security. OdaF will provide metadata management tools and usage analysis.
- The Expert Committee was also told about a joint project for a DDI 3.0 Toolkit, which involves some Alliance members and was approved by the DDI Alliance Steering Committee. This project has been established to jumpstart DDI 3.0 tools development and will create infrastructure and build a core library and technical documentation that will make it possible to easily create open source tools on the desired platform. The project will also create converters from 1 and 2 to 3 and back; a 3.0 URN resolution tool; 3.0 stylesheets with display and editing layers; a grouping tool; a validation tool; and a concept management tool. Also envisioned are question banks and registry applications. This will all be done in an open environment with contributions from several sources, including the Alliance, GESIS-ZUMA, UKDA, DDA, Canada RDCs, and OdaF. So far there is a total allocation of around $18,000 along with in-kind contributions, but the full project will require much more support, so Pascal is beginning some fundraising efforts.

We need to organize and define processes and reporting for the project; project coordination is very important for an undertaking of this nature with so many different contributors and stakeholders. It will be necessary to have MOUs from the major players. Building the core library will constitute 30 to 40 percent of the development, and it may be that with the existing monies we can only create the core library and some of the smaller tools. Interest seems to have coalesced around the Eclipse environment. We also need a UML implementation model.

The Alliance is interested in this project and would like to help it along with a contribution of $10,000 seed money. We encourage others to support as well and to become partners.

Next Meeting

The next meeting of the DDI Alliance Expert Committee will be held in Palo Alto, California, at Stanford University in conjunction with the 2008 IASSIST conference on Saturday, May 30, 2008.