DDI Expert Committee Meeting

May 27, 2006
Ann Arbor, Michigan

Present:
Hans Jorgen Marker, Danish Data Archive (Chair); Phil Bardsley, University of North Carolina; Grant Blank, American University; Ernie Boyko, Statistics Canada; Bill Bradley, Health Canada; Lu Chou, University of Wisconsin; Maggie Darby Townsend, University of Wisconsin; Mark Diggory, MIT; Karl Dinkelmann, University of Michigan; Janet Eisenhauer-Smith, University of Wisconsin; J Gager, Aeon Consulting; Fred Gey, University of California, Berkeley; Dan Gillman, Bureau of Labor Statistics; Arofan Gregory, Aeon Consulting; Reto Hadorn, Swiss Data Archive; Carol Hert, University of Washington; Pascal Heus, World Bank; Chuck Humphrey, University of Alberta; Sanda Ionescu, ICPSR; Jim Jacobs, University of California, San Diego; Mari Kleemola, Finnish Data Archive; Julie Lamb, University of Surrey; Ken Miller, UK Data Archive; Ron Nakao, Stanford University; Chris Nelson, Metadata Technology; Rob O'Reilly, Emory University; Tom Piazza, University of California, Berkeley; Wendy Thomas, University of Minnesota; Ed Van Duinen, University of North Carolina; Joachim Wackerow, GESIS-ZUMA; Oliver Watteler, GESIS-ZA; Marion Wittenberg, DANS.

I. Progress Review of DDI Version 3.0.

J Gager summarized where we were in terms of testing Version 3.0. The few examples that were prepared enabled spotting and fixing various issues. The entire Expert Committee was asked to be available to answer questions by the SRG, and use and experiment with Version 3.0. The SRG committed to producing reasonable field-level documentation and some early explanatory documents to facilitate this process..

Some useful examples were mentioned, such as:.

1. Ethiopia example
2. Geography example (Wendy/UMinn Pop Center)
3. Aggregate example (Wendy/UMinn Pop Center)
4. Instrument Documentation example (Tom/ID group)
5. Comparative example (Joachim/ZUMA)
6. Aggregate example (Rob O'Reilly/Emory)
7. Versioning example (Ken Miller/UKDA)

While a few of these examples have already been created, most are still commitments and should be made available soon.

People were encouraged to focus on specific parts of the data life cycle that they understood and used. However, coverage for the entire life cycle is also needed to guarantee the quality of the resulting schemas.

It was noted that communication between the SRG and other working groups needs improvement. The idea was put forward to have a simple Web site for Expert Committee reviewers, maintained and updated directly by the SRG, with all available materials, including releases of the schema for review, the comment log, and any documents that might be useful, including presentations, explanatory documents, and mappings between 2.0 and 3.0. While useful for code maintenance, CVS (on Sourceforge) was not deemed a suitable interface for the average Expert Committee member.

The Version 3.0-related presentations at IASSIST 2006 should also be made available on this Web site, or through links to the current SRG site.

A draft resolution to be voted on later in the meeting was presented and discussed. Changes were suggested primarily in the timeline, to allow for drafting in-line documentation and the European vacation.

II. Documentation of Version 3.0

Three types of documentation may be prepared to facilitate testing and use of Version 3.0:

- Field-level documentation, included in and maintained as part of the schemas.
- Higher-level documentation, including papers about topics such as: grouping, identification, versioning, and referencing; different ways of using 3.0 (from migrated 2.0 instances vs. as a "clean slate" approach); functionality of 3.0 vs 2.0; etc.
- User community documentation, in which users support each other by communicating their experiences via a Wiki-type environment, threaded e-mail discussions, examples, etc.

It was generally agreed that, at a minimum, field-level documentation and some of the higher-level documentation was needed for reviewers to really be able to use the schemas and produce markup instances.

It was agreed that the in-line documentation will be produced by the SRG, in direct consultation with the various working groups or individuals who are more specialized in certain areas.

A few simple examples are also needed, as an internal part of the package, when Version 3.0 is released for public review. These should be shown in XML, but also through stylesheets that would produce easily-readable presentations of the same metadata.

The user community documentation could follow, and would be a work-in-progress, supported by some of the approaches already starting to occur around 3.0 (as the online help put up by Pascal, etc.).

Timeframes for documentation would be:

- Field-level and basic "higher-level" documentation: available by end of June;.
- Examples and more higher-level documentation: July/August;.
- More complete documentation to support public comment: October.
- Other documentation (user community): put in place as possible following the public comment period, to support release of standard in early 2007.

Ilona Einowski and Ann Green have volunteered to be the editors and usability testers for the documentation - everyone else was to provide them with as much content as possible.

III. Vote on Next Steps for Version 3.0 and Agreed Timeline

After some brief discussion resulting in minor edits, the following resolution was unanimously passed by the present members of the Expert Committee:

Resolved

We approve the scope and lifecycle approach taken in DDI 3.0 internal XML schema drafts, including the inclusion of grouping; modularity; time, geography and aggregate data; complex files; instrument documentation; and comparative capabilities. In moving forward to a draft for public comment, we understand that further planned functionality such as extensions and use of controlled vocabularies will be included. To facilitate public comment, more complete documentation will be provided both for the schemas and as explanatory documents, including a formal conceptual model and examples of the DDI instances. By the end of June we will expect internal review to begin within the Expert Committee, for a period of three months. After approval of the draft, it will be released for public comment in the fall of 2006, with anticipated vote for adoption of the specification occurring in the first quarter of 2007, pending the results of the public comment.

During the discussion, further timeline detail was outlined as described below. This was not part of the formal resolution, but was generally agreed upon as reasonable. The internal review period was extended from two to three months due to the tendency for European participants to be on vacation during July and August. The target-date for end release was seen as anytime within the first quarter of 2007, so as to be ready for next year's IASSIST.

Timeline Detail:

(1) In-line documentation sufficient for internal review by end of June;

(2) Internal comments due by end of September;

(3) Vote to submit for public review by mid-October;

(4) Public comment begins on November 1, through the end of January;

(5) Depending on results of public comment, vote for adoption mid-February or early March.

It is important to publish Version 3.0 as soon as possible, since many actual and potential users will only find bugs and other problems after it gets in production mode. Then we can follow up with minor versions that will address those concerns.

It was also suggested that all Expert Committee members should be invited if there is a face-to-face meeting of the SRG next fall. Since Version 3.0 is becoming something that requires input and testing rather than just straight development, this proposal seemed appropriate, with the caveat that some members might not be able to make it if they cannot get funding from their institutions.

IV. Working Groups and Organization Moving Forward

It was generally agreed that the current working groups will remain active through the internal review period, providing support in the production of documentation, markup samples, and testing parts of the new specification.

Some of the current groups will continue to exist beyond the internal and public review:

- SRG will continue to focus on further developing the specification, and provide technical leadership and support.
- The Usability and Outreach group will also continue to exist, with an attempt to better define its goals. Ron will try to reorganize and refocus the group on a number of priorities, such as documentation, increasing functionality of the Web site, marketing, training and user support, etc. Sub-groups may need to be created to more specifically support each of these functions.
- The Instrument Documentation group will continue, with an eye toward adding some pieces that were not in scope for the 3.0 release.
- The Comparative group will continue for review and example-creation purposes. Beyond that, it will need a new set of goals, and may have to recruit new members.

A few new groups were proposed:

- An Early Lifecycle group that will focus on SMDS – DDI compatibility, as well as other projects related to the early stages of preparing and producing a social science study. (SMDS captures metadata at the stage of data collection.)
- A Tools group was proposed and seen as useful by everyone; Pascal, Arofan, J Gager, Mark Diggory, and Joachim are natural candidates for this group.
- A standards alignment group was suggested, although the need for this was not recognized by everyone. Also, there is a question as to whether liaison functions belong within the Expert Committee or at a higher level.
- A group was suggested to support research collections and general content standards (recommended classifications, etc.) Such a group may prove useful in planning for creating and maintaining registries, which is a future goal of DDI.
- An example group was suggested, to illustrate the production use of 3.0. Two datasets were suggested as useful:

  Oliver suggested the ISSP, which had some appealing features: it is multi-lingual; it is not too large; it's part of the GSS; it encompasses the full life cycle; it is a multi-year study; and there already are existing 2.0 instances. This would show off the new features of 3.0 to advantage: for example, comparisons can be drawn across countries, topics, and time.

  Joachim suggested the European Social Survey, which seemed to be a smaller and easier set of data with some of the same features as ISSP.

The general response was that both examples would be great, but that the role of an actual Example Working Group within DDI was unclear. Would these examples be specially supported by the DDI Expert

Committee, to serve as reference examples? No decision was made on this point, but the idea of having these examples was well-received. Collaborative work on an example might also help in exploring new tools.

Suggested working groups on qualitative data, historical data, and secondary analysis were seen as good ideas, but it was deemed that the time was not yet right for their organization, as Version 3.0 needs to be more firmly established first.

It was also suggested that an "interactive" Web site might serve as a more efficient means of forming/adding to working groups: the proposed goal(s) and timeline(s) would be posted on such a site, and interested parties could sign up, or be solicited.

V. DDI Web site

Several attendees pointed to a need for a more dynamic, interactive site, where documents may be posted without delay, the contents changed promptly, and various updates and exchanges made possible.

In additon to the public site, where content needs to be controlled and more stable, it makes sense to also maintain a "working site" that would be more dynamic and flexible, answering the Experts' and various groups' need to communicate and share information or other materials. Such a site would become the responsibility of a larger group of people, which would allow new entries and updates to be handled more easily.

It was decided that Ron would investigate the possibilities of creating and maintaining such a site.

VI. Tools

A report was given of the discussions during the DDI 3.0 Tools meeting that had taken place two days earlier (May 24) in conjunction with IASSIST. It was made very clear that a few simple tools would be needed without delay to enable less technically sophisticated users to test, and work with, Version 3.0. Three types of tools were discussed:

- A start-up toolkit
- An Eclipse/EMF-based development environment
- Registry tools

The Start-Up Toolkit will need to include:

- A 2.0-to-3.0 converter (XSLT)
- A 3.0-to-2.0 converter (XSLT)
- A tool for grouping several stand-alone 3.0 instances (XSLT)
- A tool for breaking out simple 3.0 instances from a group (XSLT)
- A Forms-based tool for viewing and editing small instances (Pascal's X-Forms tool)
- An XSLT stylesheet for doing simple human-readable displays of 3.0 instances
- Conversion tools from statistical formats to DDI 3.0

Joachim is already working on an SPSS system and a SAS system converter to Version 3.0. Arofan and J, working with Mark Diggory, can probably provide most of the stylesheets described above.

Achim was also in favor of adopting a general reporting tool that exports to docBook format. This, in turn, is easily convertible to multiple formats like PDF, HTML, Eclipse, etc.

Nesstar support for Version 3.0 was discussed. Nesstar is looking at Version 3.0 and will become Version 3.0-compatible, although the timeframe is not clear at the moment.

It was also mentioned that the output of software like Blaise or CASES needs to eventually become translateable into 3.0. Persuading the vendors who produce statistical packages to support DDI was also discussed.

Other topics included an Eclipse-based DDI development environment using EMF and other tools, and the creation of a statistical registry tool.

Chuck Humphrey described a funded open-source project in Canada that would be working on some of the lifecycle aspects, and he used the phrase "open data environment" to describe the easy availability of data and metadata across organizations and applications, which was the target. The University of Alberta is home to this project, which involves documenting approximately 100 studies in DDI 3.0 and creating tools for the markup.

The existence of the Open Data Foundation was also mentioned as a space in which open-source projects could be conducted, and as a vehicle for assuming liability.

It was agreed that the Tools group (Pascal, Arofan, J, Mark Diggory, Joachim, some others) would have responsibility at least for:

- Creating a start-up toolkit

- Supporting vendors who wanted to integrate with the DDI.

A listserv needs to be created specifically for tools, involving the tools group, other tools creators, and users.

The creation of more sophisticated tools was also discussed, but may not be a DDI Expert Committee activity as such, aside from being something that would obviously be worth supporting from the standards side. It sounded as though Chuck Humphrey's project might provide one good venue for such development to take place.

VII. DDI -- The Big Picture

Becoming an Official Standard

Dan Gillman outlined the steps involved in becoming an ISO standard. The procedure is quite complex, and considerable effort is needed to format and rewrite a standard to make it compatible with the ISO requirements (an effort of 2+ years). The process for getting the status of a full ISO standard is very heavy and quite political.

Arofan mentioned other options (W3C, OASIS), but these are not the most appropriate for DDI. W3C is domain specific, and OASIS' rules regarding an open process conflict with our own regulations about membership.

It was generally agreed that, in the long term, ISO is the best option for DDI. ISO has a deserved reputation for excellence, its rules are well tested and work, and its standards are recognized world-wide.

Moving to ISO is well worth the effort, and will remain our ultimate goal. However, we are not ready to embark upon this process right now. First we need a stable specification, and a commitment, coupled with the necessary resources, to ensure a five-year maintenance cycle.

Timeline for Moving Forward, Outreach, Next Meetings

The timeline for the next year or so was agreed upon, as presented above (see point III. Vote on Next Steps for Version 3.0 and Agreed Timeline). The overarching goal is to "go public" before IASSIST 2007.

As shown above, Ron will try to coordinate the outreach efforts – gaining more visibility, training and education, sharing of tools, marketing, communication.

The next Expert Committee meeting will take place at IASSIST 2007, but the possibility of a face-to-face in the fall will also be considered (as previously agreed).

The next SRG meeting should occur sometime in the fall, although the actual date is still unclear due to conflicting schedules. The timing for this meeting will be further discussed within the SRG, and the Expert Committee will be notified as soon as some consensus is reached.