

DDI Alliance Expert Committee Meeting Minutes

Stanford University, Palo Alto, CA
Saturday, May 31, 2008

Participants

- Iris Alfredsson (Swedish Social Science Data Service -- SSD)
- Atle Alvheim (Norwegian Social Science Data Service -- NSD)
- Nikos Askitas (Institute for the Study of Labor, IZA, Germany)
- Michelle Edwards (University of Guelph)
- Alexis Furuichi (Princeton University)
- Julie Gibbs (University of Surrey)
- Arofan Gregory (Open Data Foundation)
- Pascal Heus, (Open Data Foundation)
- Sanda Ionescu (Inter-university Consortium for Political and Social Research -- ICPSR)
- Jannik Jensen (Danish Data Archive -- DDA)
- Uwe Jensen (German Social Science Infrastructure Services, Central Archive for Empirical Social Research -- GESIS-ZA)
- Mari Kleemola (Finnish Social Science Data Service -- FSD)
- Stefan Kramer (Yale University)
- Fredy Kuhn (Swiss Foundation for Research in Social Sciences -- FORS)
- Sonia Latour (University of Alberta)
- Hans Jorgen Marker (Chair, Danish Data Archive -- DDA)
- Marc Maynard (Roper Center)
- Kate McNeill (Massachusetts Institute of Technology)
- Ken Miller (United Kingdom Data Archive -- UKDA)
- Meinhard Moschner (German Social Science Infrastructure Services, Central Archive for Empirical Social Research -- GESIS-ZA)
- Ron Nakao (Vice Chair, Stanford University)
- Robert O'Reilly (Emory University)
- Tom Piazza (University of California, Berkeley)
- Janet Eisenhauer Smith (University of Wisconsin, Madison)
- Jon Stiles (University of California, Berkeley)
- Wendy Thomas (University of Minnesota)
- Mary Vardigan (Inter-university Consortium for Political and Social Research -- ICPSR)
- Joachim Wackerow (German Social Science Infrastructure Services, Centre for Survey Research and Methodology -- GESIS-ZUMA)
- Wolfgang Zenk-Moltgen (German Social Science Infrastructure Services, Central Archive for Empirical Social Research -- GESIS-ZA)

- Bill Block (observer, new member July 1) (Cornell University)
- Dan Gillman (observer) (Bureau of Labor Statistics)
- Linda Harding DeVries (observer) (Statistics Canada)
- Nancy Hoebelheinrich (presenter) (Stanford University)
- Jeremy Iverson (observer) (Colectica)
- Dan Smith (observer) (Colectica)

Welcome, Introductions, and Congratulations

DDI Alliance Chair Hans Jorgen Marker opened the meeting. After a round of introductions, Alliance Director Mary Vardigan stated that DDI 3.0 had been officially published a month earlier and that the Alliance members deserved congratulations for all of their hard work. She singled out members of the Technical Implementation Committee (TIC) who were present -- Chair Wendy Thomas, Vice Chair Achim Wackerow, Arofan Gregory, Pascal Heus, and Ken Miller -- for their dedication to creating the new specification.

Building Alliance Membership

The DDI Alliance Steering Committee supports a proposal to create a new "allied" category of DDI Alliance membership. This membership type would permit statistical agencies, national statistical offices, and other data producers to be members of the Alliance with a voice in the development of the DDI standard. This category of membership would not require a membership fee (many statistical offices cannot obtain money to join organizations like the Alliance) and would most likely not carry voting rights. A document describing this membership category will be drafted.

We need to begin to market DDI to statistical agency data producers in order to have DDI markup generated automatically out of data producing systems. Wendy Thomas will be attending the 2009 meeting of the International Association of Official Statistics, a section of the International Statistical Institute (ISI), in Durban, South Africa, and will present to this group about DDI. It is also important to be on the program of groups like METIS (UNECE-Eurostat-OECD) that discuss statistical metadata. Dan Gillman chairs this group and will bring to their attention the allied membership opportunity.

Mapping to Other Standards

The metadata standards landscape is confusing to users, so we need to think about how to communicate the relationships between and among the standards that are relevant in the social sciences and also provide guidelines for usage. Mappings and crosswalks can illuminate the differences and similarities. DDI particularly needs to clarify its relationship to the Metadata Encoding and Transmission Standard (METS) and to the Preservation Metadata Implementation Strategies (PREMIS) standard. This work is on the agenda of the Alliance as is a mapping to ISO 11179, the metadata registries standard. Because DDI 3.0 was developed to align closely with ISO 11179, it is basically an XML implementation of that specification, so the relationship between the two is evident. An SDMX-DDI mapping is also being developed and will be available in about six months.

METIS publishes a Common Metadata Framework, divided into four parts:

- Part A - Corporate Context
- Part B - Metadata Concepts, Standards, Models and Registries
- Part C - Metadata and the Statistical Cycle
- Part D - Implementation

Each part concentrates on different practical and theoretical aspects of statistical metadata systems, vital knowledge for any person working with statistical metadata. It would be useful to have DDI 3.0 be represented in Part B, where the Neuchatel Terminology Model on variables, concepts, objects, and attributes is presented.

Update on Tools and Related Projects

A summary of the DDI Foundation Tools Program (FTP) was provided. The goal of this project, to which several partners are contributing, is to construct a core library of open source objects and components upon which DDI tools may be built. A Roadmap lays out the sequence of tool development and specific guidelines for tools creation. To date, the Tools Program has produced a Web site, a set of Apache XML Beans (Java objects that represent DDI 3.0 content), a URN generator, and a basic validation tool. Some tools have been donated to the Program, specifically the DDI/DExT tools (sponsored by UKDA/ODaF) that convert from SPSS to other statistical formats and both versions of DDI, and the SPSS/SAS to DDI converter (GESIS ZUMA). Colectica has created a tool called SurveyViz that imports Blaise, CASES, and other CAI formats and permits the visualization of question flow in a survey. Recently, an SPSS to DDI 3.0 converter for R has been developed at the Instituto Regionale di Ricerca della Lombardia (IRER) in Italy.

A new project to create a suite of DDI editing tools building on the FTP is now ramping up. About ten organizations have indicated interest and a meeting was held earlier in the week to assess what each potential partner could contribute in terms of funding and/or in-kind contributions. The goal is to coordinate resources and to build a DDI editor as soon as possible. The first phase will produce a light editor, with more advanced features being added in Phase 2. It is hoped that the basic editor can be created within six to nine months. There are many ways to contribute to the project, including working on documentation, writing up use cases, testing, etc., and any Alliance members are welcome to be part of the effort. We also need a separate tool to convert existing markup to DDI 3.0. It is possible that this could develop into an API with documentation explaining how to use it.

An update was also provided on the following DDI-related projects:

- NORC Data Enclave — This project involves not only documentation but also parsing and sharing of source code
- EURASI (European Access to Statistical Information) proposal — This project would connect Research Data Centers (RDCs) in Europe and provide virtual access, networking, and anonymization features along with a DDI 3.0 based component for metadata management.
- Canadian RDCs — This is a project to connect all of the RDCs in Canada and to use DDI 3.0 for metadata management.
- DANS MIXED (Migration to Intermediate XML for Electronic Data) project — DANS has created an XML format for the preservation of microdata.
- International Household Survey Network (IHSN) and the Accelerated Data Program, sponsored by the UN OECD/PARIS21/World Bank — These efforts are providing tools to national statistical offices in developing countries to permit them to document and disseminate their data according to best practices. The tools need to scale up because over 60 countries are involved and the number could reach as many as 100. There is now a Web-based data catalog used by national ministries and line ministries. The IHSN has no short-term plans to move to DDI 3.0 as their tools are built on DDI 2.0.

It is hoped that DDI can be designated a United Nations Statistical Commission "preferred standard," as has occurred with SDMX recently.

Transitioning from DDI 2.* to 3.0

A presentation was made on creating an interim version of DDI (a "2.2" version) that would permit a two-step transition from 2 to 3. This would move us away from the DTD, incorporate some features of 3, and perhaps help large-scale users like IHSN and Nesstar to make the transition. This transitional version would be schema-based, have ID validation, add mandatory elements like agency, and facilitate the expression of a codebook view from 3. It would be the final 2.* release and would carry no new functionality.

The Alliance voiced a number of views on this topic. Several members were not convinced that a two-step migration path is necessary. Also mentioned was the fact that another 2.* version could be confusing to users and it is not clear how we would brand this version. It was pointed out that as a standards body, the Alliance cannot control what users do, but we do not want a branching version tree.

Others indicated that if people need features of 3.0, they will adopt it. Nesstar and the CESSDA Preparatory Phase Project (PPP) are evaluating DDI 3.0 and will determine how to proceed based on the results of the evaluation.

The point was made that we have nothing to lose by keeping the migration process internal because we all need a way to migrate. Rather than designating a version 2.2 that would go on SourceForge with other versions and lead to confusion, we should incorporate the concepts of 2.2 into a migration toolkit that the Foundation Tools project has a mandate to create. Part of this migration would be migration guidelines and best practices.

The Alliance decided that there was no support for a designated 2.2 but that we clearly need a migration toolkit, to be developed by the separate Foundation Tools project. Also necessary is a comprehensive 2 to 3 mapping.

Version Branding

Having different versions of the DDI standard has generated some confusion among the current and potential user communities. Now that DDI 3.0 is available, users want to know (1) whether the earlier versions of DDI will remain available and be maintained, or perhaps improved, and (2) whether DDI 3.0 is "superior" to the earlier versions and perhaps supersedes them as the higher number suggests.

The DDI UOG sought to clarify this situation and offered some branding suggestions, including branding DDI Version 2.* as *DDI Classic* and DDI 3.0 as simply *DDI*. Other suggestions included *DDI First-Generation (2) — Second Generation (3)*; *DDI Codebook* or *DDI-Publication (2)* and *DDI Life Cycle (3)*; and DDI 3.0 as *DDI ++*. It was decided that this type of branding may be limiting and may cause future problems; thus we should stay with version numbers. However, the UOG recommends that "DDI" be the term used to market the standard in general (as opposed to referring to specific version numbers in general publicity materials). In terms of the UN's possibly approving DDI as a preferred standard, no version distinctions are necessary in that process.

Timeline and Next Steps

With DDI 3.0 having been published on April 28, the specification is now in "stabilization mode." At this point, we need to focus our energies on education, tools, and outreach. The Usability and Outreach Group (UOG), in consultation with the TIC, will need to be very active during this period. However, this does not mean that working groups should be idle. They should continue to meet because the process of developing new elements takes a long time.

In terms of uptake of the standard, we have a small group of early adopters and will be creating editing tools in the next six to nine months. It will probably be 12 to 18 months before we have a more significant number of adopters. With the new life cycle orientation of DDI, new audiences like statistical agencies are showing interest, and there may be new constituencies that we are not yet aware of.

With respect to the revisions process, according to the Bylaws we have four basic types of changes: bug fixes, minor revisions without review, minor revisions with review (a scaled down process), and major versions requiring the full review process.

There is a provision for the Director to approve some minor changes upon recommendation by the TIC (these types of changes include bug fixes, documentation changes, controlled vocabularies, etc.), but other minor revisions require review. The TIC proposed that we redefine the revision process in terms of whether there is backwards compatibility.

Specifically, the TIC proposed (and the Alliance approved) that:

- TIC will assess the bug tracker Mantis every three months and make recommendations to the Director: No release; bug fix; minor with review; or minor without review (a major revision will always be decided by the Alliance).
- Not more than two releases will take place per year (each release requires namespace changes in applications).
- We may move to a system of fixed-date releases once the specification is stable.
- Controlled vocabularies will be produced and disseminated separately from the schemas and will use OASIS Genericcode standard for markup. They will have their own publication cycle. We expect the first release within three to six months.

It was further emphasized that we do not want a new version 3.1 coming out soon because we want people to be able to adopt the specification with confidence.

Hosting Public Registries

The DDI Alliance needs to publish a formal list of organizations using DDI so that each gets its own unique agency code. Each institution would complete a Web form that would generate a URN for its code. We also need a service for resolving references that are external to the DDI instances in order to share metadata. This needs to be available all the time.

It was suggested that we all agree to use the DANS scheme of persistent identifiers based on URNs. This is an alternative to Digital Object Identifiers (DOIs) and other such schemes. We also need to register DDI as a URN domain, which will take some time. The TIC will draft proposals for the URN service and a potential membership list that includes CESSDA, ICPSR, IHSN, Data-PASS partners, University of Tilburg, and others.

CESSDA Preparatory Phase Project (PPP) Report

The Council of European Social Science Data Archives (CESSDA) has received a major award to develop the CESSDA research infrastructure and to focus on tackling and resolving a number of strategic, financial and legal issues in order to ensure that European social science and humanities researchers have access to, and gain support for, the data resources they require to conduct research.

The project consists of several interlinked yet individually focused work packages, including work on developing the data portal to allow seamless access to data holdings across Europe, developing common authentication and access middleware tools, developing metadata standards, creating thesauri management tools, extending the coverage of and strengthening the CESSDA research infrastructure, investigating the potential of grid technologies, and improving data harmonization tools. The project will include an evaluation of the capacities of DDI 3.0, so the Alliance will be in communication with the project about this and there is the possibility of a meeting.

Introduction to METS

Nancy Hoebelheinrich, Metadata Coordinator at Stanford University and Co-Chair of the METS Editorial Board, met with the DDI Expert Committee to discuss ways in which DDI and METS, the Metadata Exchange and Transmission Standard, could collaborate. She explained that METS, which got its start in the archival environment, is a standard for digital objects and metadata that are encoded together. METS provides a way to aggregate, package, encode, render, share, and archive information. In Open Archival Information System (OAIS) terms, METS creates syntax for the Submission Information Package, the Archival Information Package, and the Dissemination Information Package.

One possibility for collaboration is to link from a METS record to a DDI instance. METS also has a mechanism called the Endorsed External Schemas, which is another possibility for DDI. DDI is currently on the METS controlled vocabularies list, along with MARC, MODS, EAD, PREMIS, and others, so it is a known metadata type. Another possibility is a Profile Schema that allows the declaration on an instance of best practice in human-readable form. The process is that one submits to the board, which reviews the application for the profile, and then the profile is published in 30 days. A logical use case would be to package a dataset and DDI metadata using a data subset. We could create METS profiles for Version 2 and for Version 3.

Future Meetings of the Alliance

The CESSDA Expert Seminar will be held in Odense, Denmark, in September 2008 and will focus on DDI. If the Alliance has significant agenda items by that time, we could perhaps have a DDI Expert Committee meeting there. There are other meetings to consider — for example, there will be a CESSDA PPP Community Coordination Workshop late in October or early in November focusing on specific work packages. We will determine the date of this workshop and solicit suggestions for agenda items for a fall meeting.

DDI Fall Training

DDI 3.0 training, sponsored by GESIS-ZUMA, will again be offered at Schloss Dagstuhl Leibniz Center for Informatics, November 3-7, 2008. This workshop is geared toward the staff of archives and data producing agencies. The multi-day structure of the workshop provides participants with an opportunity for in-depth assistance on the specialized features of DDI that are important to their organizations' activities.

The following week (November 10-14), there are plans for an Expert Workshop at Dagstuhl, sponsored by the DDI Alliance, GESIS, Minnesota Population Center, and the Open Data Foundation. The premise of this workshop is that with the publication of DDI 3.0, there is now a question of how best to implement it and thus there is a natural growing interest in documenting best practices to support interoperability and sharing of metadata.

The focused expert workshop would be designed to produce two papers addressing this need: one on technical best-practices and the other dealing with organizational approaches, governance, and content. These will form the basis of ongoing work in this area within the DDI community. The intention is to publish these papers on the DDI Alliance Web site and in other places as appropriate.

Suggested topics for the expert workshop include:

1. **Controlled vocabularies:** How these should be selected, used within an organization, and communicated to users of DDI.
2. **Governance:** Best practices for agreeing on a set of metadata elements within a community of use, how to make decisions over time, how to communicate within a community.
3. **Use of DDI 3.0 Schemes:** How best to organize metadata such as variables, questions, concepts, categories and codes so they are optimized for reuse.
4. **Management of DDI 3.0 Identifiers:** How organizations should assign and manage the identifiers for which they are responsible across their entire set of metadata.
5. **URNs and entity resolution:** How to provide these services at the level of an organization, and at the level of the community.
6. **Workflows:** Recommended workflows for archival use, for data production, for data dissemination, for provision of services such as question banks, variable banks, etc.
7. **Metadata storage:** How to maintain the metadata set independent of its expression in XML as it flows into and out of applications.
8. **Versioning and publication:** What constitutes a published version of a metadata set?
9. **High level architectural model:** Recommended multi-level architecture for a suite of applications that use DDI 3.0.

This would most likely be an invited workshop, with key people represented and a good mix from different types of institutions. There was a lot of interest expressed in these topics, especially from the perspective of the CESSDA PPP.

DDI Regional User Groups

Nikos Askitas from the Institute for the Study of Labor (IZA) in Bonn, Germany, and Joachim Wackerow (GESIS-ZUMA) made a proposal for a regional European DDI users meeting, modeled on the successful Stata users meetings. The idea is that the DDI Alliance would begin to sponsor regional DDI user meetings and groups, with dedicated Web pages on these meetings. The meetings would have the goal of providing a forum for exchange among users and communication of specific needs back to DDI Alliance.

IZA has proposed to host the first such meeting in 2009 in Bonn, Germany, in September/October (fall seems best since it should not be too close to the IASSIST conference and should not be in the summer). In addition to Europe, North America could host regional user meetings and ideally we can extend the geographic regions over time.

There could be a call for papers and a program committee with representatives from different institutions. Also proposed is an opening plenary event on a topic such as "How DDI can facilitate the research process" with a presentation by a distinguished researcher or a panel session with several persons. Members from Technical Implementation Committee should be available in a session on "Ask the Experts." Optional would be a workshop for beginners (1/2 day or one full day) preceding the meeting. The UOG co-chairs proposed that this meeting, as one of the user outreach activities, be coordinated with the UOG. The next steps are that IZA and GESIS will start planning the workshop.

Intranet for the DDI Alliance

A proposal was made to explore using Drupal, an open source content management system, for the DDI Alliance Web site and also for an internal Intranet for the Alliance Expert Committee members. This would create a more dynamic Web site and would also permit maintenance and creation of content to be shared among members of the Alliance. We could set permissions to only show internal information to Alliance members while the externally facing site would be publicly accessible and would also allow comments from users and more interaction.

Google Groups are another communication mechanism as are email lists and the Marratech video conferencing system that the CESSDA PPP is using.

Working Groups

Usability and Outreach Group

Kate McNeill and Stefan Kramer, co-chairs of the group, reported on the group's mission and its goal to create two-way communications between the Alliance and users of the standard. The group has been using a Google Group to communicate (this mechanism provides uploading of files, wikis, comments features, email, etc.) and has discussed topics such as the various audiences for DDI, marketing efforts including the creation of a brochure, and requirements for a DDI editor.

The group hopes to become very active this year with members making contributions in various areas. Goals include revamping the DDI site, building a list of conferences and attendees, creating training materials, and so forth. Press releases will be written to disseminate information on every report-worthy event. The UOG hopes to have some conference calls with the TIC as the year proceeds.

Qualitative Data Group

The UK Data Archive is spearheading a project to encourage qualitative data software vendors to use a standard schema to describe qualitative data including audio and video components and text analysis. This may be incorporated under the DDI umbrella or become its own product.

Controlled Vocabularies Group

This working group has been meeting via telephone conference calls to develop controlled vocabularies for elements in DDI 3.0 that require such lists. They have made significant progress and hope to publish the first version of the DDI controlled vocabularies within the next three to six months. The list will be published separately from the DDI 3.0 XML schemas.

Other Working Groups

The Alliance discussed the fact that there have been suggestions for several new groups, but we need to be strategic in terms of forming groups so that people do not get overwhelmed. In addition to groups on Usability and Outreach, Qualitative Data, and Controlled Vocabularies, other groups being considered by the Alliance are:

- *Survey Design and Implementation*. This group will become active again and perhaps gain some new members. Because this is an important piece of the life cycle, it should have priority.
- *Digital Preservation*. This group will be chaired by Nancy McGovern, ICPSR, and will explore comprehensive digital preservation solutions for referencing and DDI applications and will also complete a DDI to PREMIS mapping and recommend new elements if needed.
- *Data Reuse*. This group will cover what happens to data after they are disseminated and reused with metadata feeding back into the DDI instance.

It was decided to add the new groups to the Web page with their missions and statements of their areas of focus as well as membership in the groups and contact information so that others may join if they are interested. Other groups that should be described include Longitudinal Data, Data Processing (including disclosure risk review and risk limitation through anonymization), National Registers Data, and Data Processing. Some of the groups may discover that what they need is already contained in DDI and that what they should do is to write up best practices rather than recommend new elements.