DDI Alliance Expert Committee Meeting Minutes
Ithaca, New York – Cornell University
May 31, 2010

Participants

Iris Alfredsson (Swedish National Data Service -- SND)
Atle Alvheim (Norwegian Social Science Data Service -- NSD)
Bill Block (Cornell University)
Michelle Edwards (University of Guelph)
Jane Fry (Carleton University)
Dan Gillman (US Bureau of Labor Statistics)
Arofan Gregory (Open Data Foundation)
Alistair Hamilton (Australian Bureau of Statistics)
Sue Ellen Hansen (Survey Research Operations, University of Michigan)
Chuck Humphrey (University of Alberta) -- Chair
Sanda Ionescu (Inter-university Consortium for Political and Social Research -- ICPSR)
Jannik Jensen (Danish Data Archive -- DDA)
Anne Sofie Kjeldgaard (Danish Data Archive -- DDA)
Mari Kleemola (Finnish Data Archive – FSD) – Vice Chair
Stefan Kramer (Cornell University)
Arja Kuula (Finnish Data Archive – FSD)
Jared Lyle (Inter-university Consortium for Political and Social Research -- ICPSR)
Marc Maynard (Roper Center)
Steve McEachern (Australian Social Science Data Archive -- ASSDA)
Katherine McNeill (Massachusetts Institute of Technology – MIT)
Susan Mowers (University of Ottawa) -- observer
Ron Nakao (Stanford University)
Tom Piazza (University of California, Berkeley)
Anita Rocha (University of Washington – member as of July 1, 2010)
Laurents Sesink (Data Archive and Network Services – DANS)
John Shepherdson (United Kingdom Data Archive -- UKDA)
Samuel Spencer (Australian Bureau of Statistics)
Jon Stiles (University of California, Berkeley)
Wendy Thomas (University of Minnesota)
Mary Vardigan (Inter-university Consortium for Political and Social Research -- ICPSR)
Joachim Wackerow (GESIS)

Environmental Scan

Chair Chuck Humphrey opened the meeting with introductions of all participants and then moved on to pose the questions: What are the aspects of the current environment in which DDI operates that are different from 8-10 years ago and is the DDI specification – along with the DDI Alliance itself -- structured properly to address current conditions?

Meeting participants responded with a lengthy list of changes (this is a sample):

- E-science and cyberinfrastructure
- Transparency initiatives such as data.gov
- Increasing amounts of health and biomedical data
- Structured searching for data and LinkedData
- Fine-grained metadata and reuse of it
- Mixing of data types
- Increased collaborations
- Use of Open Archival Information System (OAIS)
- Emphasis on register data
- Multiplicity of metadata standards and need for mappings
- New data formats
- More data available from statistical agencies
- More research data centers and restricted data
- Emphasis on machine-actionability
- Data visualization
- Vendor interest in data
- Growing appetite for data in general
- Growing mentions of DDI in international contexts (e.g., METIS)
- Importance of intellectual property
- More undergraduate use of data
- Increasing data complexity
- Open access policies
- Tools to integrate and harmonize data
- Links between publications and data
- Social network data
- Qualitative data and combined qualitative/quantitative
- Self-archiving
- Influence of Google
- More data on the Web of unknown provenance
- Lack of mechanisms to evaluate data quality
- Mashups and APIs
- Replication datasets
- More comparative cross-cultural research
- Paradata available from the beginning of data life cycle
- Confidentiality challenges

Evolution of the Alliance

Given the changes and complexity described, the Alliance needs to think about whether the DDI specification is structured to meet the needs of the current environment and additionally, whether the Alliance itself has appropriate structures in place, especially with respect to intellectual property rights.

While not likely, it is possible that a member of the Alliance would assert ownership of the DDI IP or that a vendor would try to claim it. The IP section of the Bylaws is currently not fleshed out and needs to be much more specific in terms of who owns the IP and who protects it. The specification is currently distributed under the Lesser Gnu Public License (used often for open source software) to protect it and provide controls on how it is used.

The Alliance leadership has had some discussions on this topic and has investigated what it would take to restructure the Alliance as a legal non-profit entity. According to the lawyer consulted, this approach would necessitate a large administrative burden, and it is not clear that the benefits in terms of IP protection would outweigh the work involved to achieve official non-profit status.

Another possible way to protect IP is to join an existing standards organization like OASIS -- Organization for the Advancement of Structured Information Standards. OASIS is "a not-for-profit, international consortium that drives the development, convergence and adoption of open standards for the global information society." Joining OASIS would require a fee and a different way of operating. Another solution would be to fast-track the DDI specification to become an ISO (International Organization for Standardization) standard. Other suggestions were to solicit advice from ANSI (American National Standards Institute), despite its being an American entity, and to explore the United Nations Economic Commission for Europe (UNECE) statistical metadata group METIS, which "facilitates harmonization of data models and structures for statistical metadata in the context of statistical information processing and dissemination."

The Expert Committee finally recommended that work to research IP and the best way to handle it become a budget item for prioritization by the Steering Committee, which was to meet the following Wednesday. The group expressed its intention that this work be the highest possible priority before considering other changes to the Bylaws and to the governance structure and that a consultant be contracted to research the issues thoroughly and provide a report. This will need to be considered in concert with the license, a contributor agreement, and the membership agreement. It was also mentioned that the Alliance does not wish to change its core values but rather its structure so it can be most effective and responsive to the community in the current environment.

Semantic Web/RDF/LinkedData and DDI

This topic is receiving a lot of attention currently and poses both challenges and opportunities for the Alliance. The group discussed whether the Alliance should recommend the creation of an RDFa implementation of the DDI specification. RDFa (or Resource Description Framework – in – attributes) is "a W3C Recommendation that adds a set of attribute level extensions to XHTML for embedding rich metadata within Web documents." The RDFa serialization would express the ontology for microdata in the social sciences. There is another W3C standard, the Web Ontology Language (OWL), which is more expressive than RDF and might be another alternative.

While the LinkedData project has created a lot of interest in the Semantic Web, there are challenges for social science including protection of confidentiality because of the rich web of linkages and connections in the Semantic Web, which can lead to disclosure of identities. Statistical literacy is another issue as is the lack of provenance information. LinkedData is currently linking in an ad hoc way, so it is in the best interest of the Alliance to be at the table when the social sciences are discussed. If the Alliance had an RDF serialization, it could go to Google and say that this is the statistical model for social science microdata. Another advantage would be that in terms of the Open Archival Information System (OAIS) reference model, the DDI RDF could be an expression of the Dissemination Information Package (DIP).

While the committee agreed on the importance of this work for the Alliance, it still needs to determine how to resource this. The core of the model would be a subset of DDI 2 and 3 containing concepts and classifications, using the ISO standard for time, etc. The work to create the RDF expression of DDI should not be too difficult.

It was decided that the Alliance should sponsor someone to be at the table where the Semantic Web standards are being discussed. The semanticweb.org group has a vocabulary called SCOVO -- Statistical Core Vocabulary (SCOVO), for representing statistical data on the Web. This is the group we should try to contact. Also, the Open Data Foundation is coordinating a meeting about Semantic Web topics in July at the University of Tilburg in the Netherlands. The committee recommended that, subject to approval by the Steering Committee, the Alliance should support Dan Smith to attend the OdaF meeting and then send him to other relevant meetings.

DDI Alliance Matters

Finances

The DDI project is set to close out the year with a positive fund balance, somewhere in the range of $90-100K. In addition to standard expenditures for staff time and XML consultant support, the main expenditures during the year were to support training at Dagstuhl and the TIC meeting in Ann Arbor. The Alliance had 32 paid memberships this fiscal year.

StatTransfer

Having the StatTransfer package export DDI would be a boon to stakeholders in the DDI community. Accordingly, the Alliance made contact with the company and has received a commitment that they will provide an export to DDI 3 in their next release. Joachim Wackerow will send the company examples of DDI export from SAS and Stata to facilitate this work. It is not known at this point whether the export will be DDI 3 plus ASCII data, DDI 3 with data inline, or some other form.

IASSIST Quarterly Edition on DDI

There will be a double issue of the IQ with a special focus on DDI published this summer. The issue will contain six papers:

- "Building a Modular DDI 3 Editor" -- By Jannik Jensen and Dan Kristiansen
- "Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability" -- By Taina Jääskeläinen, Meinhard Moschner and Joachim Wackerow
- "Implementing DDI 3: The German Microcensus Case Study" -- By Andias Wira-Alam and Oliver Hopt
- "Metadata Creation, Transformation and Discovery for Social Science Data Management: The DAMES Project Infrastructure" -- By Jesse M. Blum, Guy C. Warner, Simon B. Jones, Paul S. Lambert, Alison S. F. Dawson, Koon Leai Larry Tan, and Kenneth J. Turner
- "Metadata-Driven Survey Design" -- By Jeremy Iverson
- "Questasy: Online Survey Data Dissemination Using DDI 3" -- By Marika de Bruijne

DDI Use Case Papers

Six DDI use case papers have been published on the DDI Web site as part of the DDI Working Paper Series. These papers on DDI 3 use cases are the outcomes of a workshop held at Schloss Dagstuhl - Leibniz Center for Informatics in Wadern, Germany, November 2-6, 2009. Workshop participants representing nine countries in the Americas and Europe came together to write the papers, which focus on applying DDI 3 to specific projects and systems.

- Questasy: Documenting and Disseminating Longitudinal Data Online Using DDI 3
- Building a Modular DDI 3 Editor
- Using DDI 3 for Comparison
- Extracting Metadata From the Data Analysis Workflow
- Questionnaire Management and DDI: The QDDS Case
- Grouping of Survey Series Using DDI 3

All DDI Working Papers have Digital Object Identifiers (DOIs) and the series as a whole has an ISSN number.

TIC Updates

TIC Meeting in Ann Arbor

At the Technical Implementation Committee (TIC) meeting in May in Ann Arbor, over 100 bugs in DDI 3.1 were reviewed and some major structural issues were described and noted. The current plan is to release DDI 3.2, a non-backwards-compatible version, by the end of the year with a beta release coming out first. DDI 4.0 is envisioned as a longer-term project, incorporating new features such as the qualitative data and survey design and implementation modules. DDI 4.0 would also address structural issues involving remodularization to create a data model more in line with the data life cycle. The UML model for DDI 3 is being finalized now.

It was pointed out that with a new version, implementers would find it helpful to have examples implementing the changes. The Alliance may want to develop a policy outlining the products that must be provided to support version migration, including examples. More accessible and task-based documentation is extremely important. The Alliance needs to look at creating a documentation group and providing resources for this important effort.

Report on DDI 2 Changes

The International Household Survey Network (IHSN) has proposed a series of new elements and attributes for DDI 2.1, to be incorporated in a new version 2.5. The TIC will take this opportunity to add XHTML to the DDI 2.* development line and to make changes that will facilitate transformation into DDI 3.*. In addition, the canonical version of DDI 2.5 will be based on XML schema of the same form as the DDI 3.* branch, not a DTD as in the past. A spreadsheet of the changes for the Expert Committee to review will be available shortly. The changes will be backwards-compatible and will involve no namespaces.

SDMX to DDI Mapping

Work to create a mapping between DDI 3 and SDMX version 2.0 began at Dagstuhl in 2009, and includes a Use Case paper and the detailed mappings at the field level for places where the two standards overlap. The use cases have been revisited based on projects at the Australian Bureau of Statistics (ABS), to reflect the actual production use of this mapping within that institution. The output of this work will be a draft for initial implementation and review. The ongoing work has included some implementation and mapping exercises, to validate that the approaches taken are valid, and will result in an implementable result. Release of the initial draft is expected this summer – probably July.

URN Resolution

Joachim Wackerow has made a proposal for an agency/URN resolution service that would permit DDI URNs to be resolved using the Internet Domain Name System (DNS) mechanism. This involves a very lightweight one-time registration of an agency. Agencies are resolved to a list of DDI services and all complexity below the agency level is delegated to DDI services locally. The DDI namespace must be formalized in a document, which Joachim Wackerow is drafting.

Future Releases, Requirements, and Priorities

Based on information from the TIC, the group set priorities as follows:

1. Version 2.5 with updates for IHSN and DDI 2 to 3 migration features
2. URN resolution in parallel with #1 above
3. Controlled vocabularies roll-out

4. DDI 3.2 release with examples
5. DDI 4.0 with new features including survey design and implementation and qualitative data coverage, remodularization, and documentation overhaul
6. DDI 4 may possibly also address register data and data quality, both important areas.

European Events and Information

CESSDA Project Update

There will be a new formalized legal entity called CESSDA-ERIC (European Research Infrastructure Consortium) that will involve national statistical organizations of the participating countries. Not all of the members of the former CESSDA group are likely to join the new consortium and thus the original CESSDA will continue in some form. It was pointed out that CESSDA-ERIC is an agreement between countries while the original CESSDA is an agreement between archives. Most of the CESSDA members are using DDI 2, but there is a strong recommendation coming out of the Preparatory Phase Project (PPP) to move to DDI 3.

DDI Training

Training on DDI 3 will again be held at Schloss Dagstuhl in Wadern, Germany, during the last week of October 2010. The Expert Workshop will focus on DDI and longitudinal data and will take place October 18-22.

European Users Group meeting 2010 - EDDI

The European DDI Users Group (EDDI) will meet December 8-9 this year in the Hague. A call for papers has been made available.

DDI 2 and 3 Branding Suggestions

A proposal for branding the two development branches of DDI was viewed and discussed. With this new model, DDI 2.* would become DDI-C (codebook), and DDI 3.* would become DDI-L (lifecycle). The consensus appeared to be that this would make the difference between the two branches clearer for new users. There should be a decision tree to guide users regarding which version of the standard makes the most sense for the tasks they need to perform. The Expert Committee recommended that the Director solicit the opinion of someone with marketing expertise and then do some usability testing. The TIC agreed to investigate how these new designations might be incorporated internally into the standard itself.

Working Group Reports

Usability and Outreach Group. The Usability and Outreach group submitted a proposal for the disbanding of the group and facilitated a discussion about the possibilities for new groups that could be formed. The Expert Committee then agreed on the formation of two new groups, a Tools Catalog group and a Web Site group. Stefan Kramer volunteered to lead the Tools Catalog group. In order for the Web Site group to move forward, it needs a chair, and both groups need members. A related third group focused on documentation might also form; this recommendation will be considered by the TIC subgroup on documentation review. With respect to the tools support, CISER has volunteered to host an application that will provide profiles of all the available DDI tools and provide a way to search and browse them with a Web front-end.

Implementers Group. There is interest in the formation of a group for implementers that would bring together developers interested in issues of interoperability and open source implementations. The idea is that the Alliance would provide some support for this group to meet, perhaps just before the EDDI meeting in December.

Qualitative Data Group. The Qualitative Group has reconvened with several new members and there is a lot of interest and momentum. Up to this point the group has contributed use cases to define the scope of the effort.

Controlled Vocabularies Group. This group has held videoconferences for over two years and has made a lot of progress. They have finalized CVs for 15 elements and attributes and five more are in process. The vocabularies will be made available in Genericode, separate from the DDI schemas. The timeline is that the CVs have been revised after a CESSDA review and have been sent to the Expert Committee for review by the end of June. The goal is to publish them by late summer 2010. They will first be published in American English with support for additional languages.

Survey Design and Implementation Group. This group formed two years ago and has been working along two dimensions – questionnaire design and sampling schemes. Tests of the sampling scheme against real-world examples of surveys have indicated that it works well. The goal is for the data producer to have a specification that can document what is done from an internal perspective to drive the system and also provide what the end user needs – this is the holy grail. The sampling group has two proposed CVs that it would like to review with the CVG and a combined meeting is planned. The questionnaire group is deciding what should be in the standard with input from Wendy Thomas.

TIC subgroups. The TIC is putting together small teams to address various topics, including complex questions, administrative register data, machine-actionability, and documentation review. On the latter point, what is needed is a review of the field-level documentation for consistency, a review of Parts I and II, and a determination of what the official documentation for the schemas should be.