Dagstuhl Sprint Summary Report

October 31, 2014

The fifth DDI Moving Forward sprint convened on October 20th in the Schloss Dagstuhl – Leibniz Center for Informatics. Attendees:

Kelly Chatain
Survey Research Center
Institute for Social Research
University of Michigan

Daniel Gillman
Office of Survey Methods Research
Bureau of Labor Statistics

Jay Greenfield
Booz, Allen, Hamilton

Arofan Gregory
Open Data Foundation

Denis Grofils
Eurostat

Larry Hoyle
Institute for Policy & Social Research
University of Kansas

Sam Hume
CDISC

Sanda Ionescu
Interuniversity Consortium for Political and Social Research
Institute for Social Research
University of Michigan

Jeremy Iverson
Colectica

Jon Johnson
Centre for Longitudinal Studies

John Kunze
California Digital Library
University of California

Thérèse Lalor
Australian Bureau of Statistics

Jenny Linnerud
Statistics Norway

Justin Lynch
Australian Bureau of Statistics

Steve McEachern
Australian Data Archive

Barry Radler
MIDUS-Institute on Aging
University of Wisconsin

Ørnulf Risnes
Norwegian Social Science Data Services

Chris Seymour
Statistics New Zealand

Wendy Thomas
Minnesota Population Center

Mary Vardigan
DDI Alliance & Interuniversity Consortium for Political and Social Research
Institute for Social Research
University of Michigan

Joachim Wackerow
GESIS - Leibniz Institute for the Social Sciences

Stuart Weibel
Online Computer Library Center

Knut Wenzig
German Institute for Economic Research

Michael Witt
Purdue University Libraries

Wolfgang Zenk-Möltgen
GESIS - Leibniz Institute for the Social Sciences

Overview

A significant amount of production and explorative work was accomplished during the week. The initial agenda included finalizing views for Simple Instrument, Simple Data Description, and Simple Codebook, as well as exploring Qualitative Description and Methodology, and Enhanced Data Citation. In the course of events, other areas were identified and prioritized in order to achieve the original goals and maintain progress as a whole.

The following topics were addressed during the week:

- Enhanced Data Citation
- Simple Instrument View (renamed DataCapture)
- Simple Data Description View (renamed PhysicalDataDescription & LogicalDataDescription)
- Simple Codebook View
- Qualitative Description
- Methodology Description
- Integration of Views: Conceptual/Classification/LogicalDataDescription (how to use Process)
- Controlled Vocabularies
- Identification

Enhanced Data Citation

An NSF-funded project to enhance data citation information in DDI brought together a group of ten people from different stakeholder communities. One of the issues that the project addressed is how data citation can be extended to acknowledge the contributions of different types of contributors in the development of research data, leading to the possibility of generating metrics to better understand and measure those contributions. To underlie this, structured metadata is needed.

From the proposal, the group noted the following key questions:

> What objects should have metadata?
> Which elements are needed?
> How should reuse be handled?
> What infrastructure is needed for location?
> Which need controlled vocabulary?
> What special information is needed for the citation of stream resources?

A distinction was made between citation and source information. Source information was defined as a package of information describing the source of an object. Citation was defined as the serialization of some of those objects. It provides attribution for intellectual effort and

distinguishes the object. Source information can be used for other purposes, e.g. describing an instrument and its calibration.

It was decided that all versionable objects should be able to be cited, but the decision of which objects to cite is more of a social issue and ought to be left to the responsible institution.

The following elements were identified as the minimum required for citation with additional recommendations:

Minimum
Creator (with role)
Title
Publisher
Contributor (with role)
Publication Date
Identifier/Locator
Resource Type

Additional Recommended
Version (number, date, responsibility)
Pointer to metadata
Copyright (access restrictions)
License
DataCite elements that support citation

Both the Contributor Role and Resource Type should have controlled vocabularies associated with them. The group had a conference call with Dr. Micah Altman, Director of Research from MIT, to specifically discuss the Contributor Role and the degree of contribution. The role was mapped to the CV for DDI Lifecycle events, and Dr. Altman was open to collaborating further on the taxonomy, potentially building a hierarchical classification scheme with the Harvard/Wellcome taxonomy at the top level. Using numeric metrics to establish the degree of contribution is problematic. The Harvard Wellcome group came up with a three-tiered approach using Lead, Equal, and Supporting roles. It was agreed that this would be a good solution for DDI.

The next steps for the citation group:

- Now that there is a minimal and recommended set of elements, they can go to the modelers.
- A way of handling controlled vocabularies must be defined. A mechanism for constraints on the vocabulary (e.g., to validate use or to manage relationships among the vocabulary items) is required.

- The group is going to work to ensure alignment with DataCite.
- The group is writing at least one paper and perhaps more to be published in the near future.

For more information see the wiki page here:
https://dditools.atlassian.net/wiki/display/DDI4/Enhanced+Data+Citation+Working+Group

Simple Instrument (renamed DataCapture)

The charge given to this group was to model a simple survey instrument. They began with the assumption that limiting the model to simple surveys was too restrictive and would likely cause the model to break down when applied to other domains, of which there are already multiple use cases already applying DDI. They broadened the goal to develop a robust and parsimonious data capture model that could be easily extended to non-survey data collection situations.

Progress for the week includes the following items:

- Renaming SimpleInstrument to DataCapture
- Substantial refining of all objects in Drupal, including definitions and examples
- Fully fleshing out the concept of Capture, and the elements associated with InstrumentComponents
    - Capture, Statement, Instruction, and ExternalAid (and its relatonship to DDI object OtherMaterial)
- Reconceptualizing physical and logical instruments as Implemented and Conceptual, respectively
- Identifying 'touch-points' where DataCapture interfaces with other models
- Most significantly, they gained a great degree of simplicity and clarity by incorporating the generic Process model to handle much of the heavy lifting associated with actually implementing and controlling data capture instruments.

A number of Dagstuhl participants had opportunity to comment on DataCapture; they feel it is ready to be handed over to the Modeling group. The package graph cannot be represented in this document, but may be seen in Drupal here:
http://lion.ddialliance.org/package/datacapture

Simple Codebook Working Group

The challenge for the codebook group was to consider a view that is actually a combination of other views, such as data description, data capture, and discovery. Because these views were being developed and stabilized concurrently, the codebook group began their work by analyzing the profiles used by CESSDA, IHSN, ICPSR, and the Nesstar codebooks to find commonalities, documenting their work as a list of suggested elements in an excel spreadsheet. The list identifies the element’s origin, the link in DDI 3.2, the package and object in DDI 4 (if it exists), DDI 4 properties and the date it was last checked. This is important due to the ongoing revisions to all of the packages currently in Drupal.

A simple codebook view was created in Drupal and within the list of elements it was documented whether they were found or added, e.g. StudyUnit has a new property for “documentedVersion” (as opposed to inherited ddi version attribute). There is still a lot of work to be done in adding objects in the codebook view.

While there is still some work to do in incorporating the IHSN elements into the spreadsheet and defining the codebook view in Drupal for objects that do exist, the group must wait for other important decisions to be made in integrating the other views, their objects and relationships.

For more information see the wiki page here:
https://dditools.atlassian.net/wiki/display/DDI4/Simple+Codebook+Working+Group

Simple Data Description (renamed Logical and Physical Data Description)

The logical "cascading variable model" was on the table before the Dagstuhl workshop, and the Data Description group defined their task to:

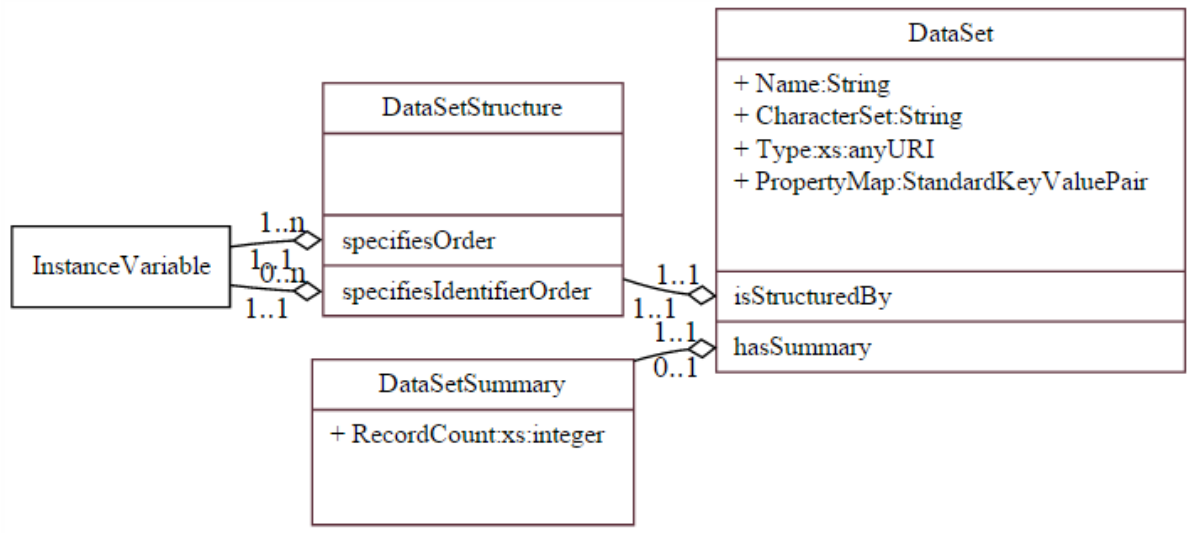
- Bring cascading variable closer to the physical data
- Align with cascading variable and GSIM as much as possible

Given this, the group started out with the assumption that they had a simple data file (i.e. a "rectangular" one, with variables on columns and units on rows). After clarifying the intentions of GSIM and cascading variable, it became clear that the Instance Variable as it emerged contained more application specific information than before; notably sentinel value domains and application specific data types.

It sounded very close to what one has when one defines a rectangular table in a relational database (via CREATE TABLE), therefore the group decided to start out with a thin model roughly based upon CREATE TABLE logic. From there they enriched it with additional information needed in the domain (but that relational databases don't care about).

The following model was produced and can be seen in Drupal:

Package Graph



Qualitative Working Group

This group assembled to explore approaches to describing new data sources that cannot currently be described in DDI, the understanding of which either is not good enough, completely agreed upon or completely unknown. Three approaches were outlined:

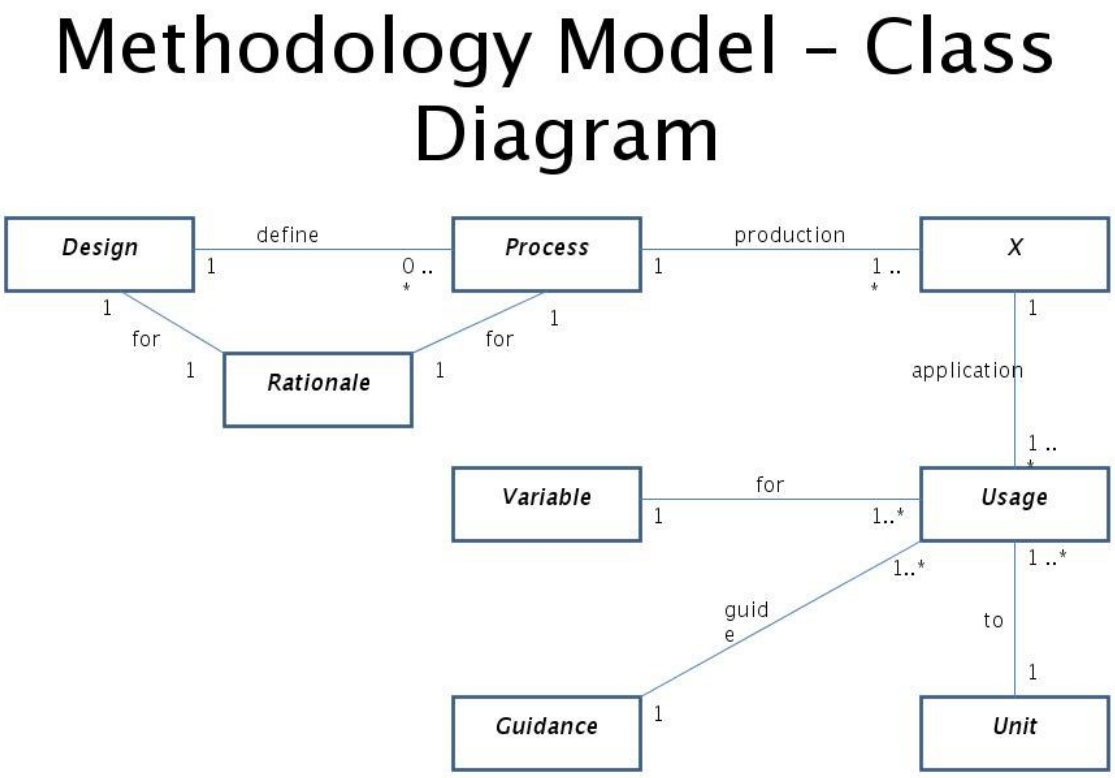
1. Simple controlled vocabulary
2. Using the DataRecord with the variables defined by DDI (Similar to SDTM in CDISC)
3. Controlled schema language for describing data records - set of objects that provide a metamodel for describing data records (similar to SDMX reference metadata – also similar to SDTM in CDISC)

For more information see the wiki page here:
https://dditools.atlassian.net/wiki/display/DDI4/Qualitative+Working+Group

Methodology Working Group

Methodology was previously discussed in Toronto and described as a series of design specifications to be documented. In Dagstuhl, this group redefined methodology as the combination of the design, the process, and the result. It is important to note that this definition includes both the intent and the actual implementation along with the achieved result.

The following diagram was produced:



For more information see the wiki page here:
https://dditools.atlassian.net/wiki/display/DDI4/Methodology+Working+Group

Controlled Vocabularies

A subgroup met to discuss how DDI currently handles controlled vocabularies and what the requirements are moving forward. They concluded that two different approaches were required: a simple CV structure similar to the approach for the current DDI Controlled Vocabularies and a more complex approach for additional requirements like validation and defining requirements. Ideally the complex approach could make use of the simple approach.

For more information see the wiki page here:
https://dditools.atlassian.net/wiki/display/DDI4/Controlled+Vocabularies

Integration of Views: Conceptual/Classification/LogicalDataDescription

Objects in the conceptual, classification, and LogicalDataDescription were integrated and rationalized over the course of a few days with more work to be done. It is requested that the

Classification team should take particular note of this work, as the integration that occurred resulted in significant remodeling of their proposed model.

See the notes on the wiki here:
https://dditools.atlassian.net/wiki/pages/viewpage.action?pageId=3833976

The Parking Lot:

- Extensibility of codebooks beyond the DDI4 view (e.g. anonymization of a variable)
- How to represent controlled vocabularies?
- DataCapture: Relationship to conceptual objects
- DataCapture: Question grid and question block related to capture
- DataCapture: Open EHR – state, confounding factors associated with measurement. How to handle? Out of scope?
- DataCapture: Do we have only statement? Or do we need to add in external objects?
- Consistency between Discovery, Codebook and Coverage

Action Items

1. Issue tracker for non-technical changes (distinct from buglog (Jon requested)
2. Wiki open source application (Mary)
3. JIRA set up (Wendy)
4. Fix broken wiki links and normalize the sprint pages (Kelly)
5. ID notes uploaded?
6. CDISC notes uploaded?

Model Specific (in no particular order)

1. Drupal Library to be organized.
2. Review Process model and communications document
3. Review Discovery objects
4. Review changes to the definitions that Denis made.
5. Review changes to Agent
6. Review changes to Logical
7. How to describe hierarchies?
8. Discovery: Get rid of foreign elements – note problem of changes in foreign namespaces. May need to version the foreign namespaces to maintain the references (put them in Drupal?)
9. Correspondence table comparison
10. Remove “type” from extended primitives, except for those that are actual types
11. Arofan/Dan issue to be decided by the Conceptual Group. Steve M. has the notes on the definition of Sign.
12. Classification and ISCED – Communication with Hilde (Classification Team)
13. Jeremy to create spreadsheet on where the gaps between 3.2 and 4 are, what is supported.

14. Review relationships to Node, in the context of the difference in structure between DDI4 and GSIM.
15. The integration of the proposed Correspondence pattern needs to be finalised (particularly CorrespondingItem and VersionableType) but depends on what is decided about the treatment of Hierarchies and Nodes.
16. Determine whether CategoryItem/CodeItem/XItem is required within DDI4 (to match as it exists within GSIM). This should form part the review of the Node object.