Recommendations from Dagstuhl 2015

The DDI Moving Forward: Facilitating Interoperability and Collaboration with Other Metadata Standards SPRINT

The DDI Sprint was held on October 18-23, 2015. This iteration of the DDI Sprint was unique in that we extended invitations to 7 external standards experts:

- Michel Dumontier, Stanford University
- Gary Berg-Cross, Research Data Alliance
- Daniella Meeker, University of Southern California
- Eric Prud'hommeaux, World Wide Web Consortium
- Martin Forsberg, Sweden
- David Barraclough, SDMX
- Alejandra Gonzalez-Beltran, Oxford e-Research Centre

Goals

Collaboration:

- To increase awareness of DDI beyond our community
- To investigate interoperability, challenges, and ways to work together
- To ensure that DDI is on the right path. Are we working in parallel? Are there opportunities for collaboration?
- To develop relationships with external standard agencies
- To provide an opportunity for an open review by invited external agencies
- Review modeling and bindings at a high level

Development work:

- Prepare all/portions of Data Description, Data Capture, and Methodology for Modeling Team review (note that these groups are in the process of setting goals for the sprint)
- Modeling Team implementation of consistency rules and decisions from Q1 review

Introduction

A metadata workshop entitled "DDI Moving Forward: Facilitating Interoperability and Collaboration with Other Metadata Standards" took place at Schloss Dagstuhl-Leibniz Center for Informatics in October 2105 in combination with a "sprint" to progress development of the next-generation model-based DDI standard (sometimes called DDI4 or DDI Moving Forward).

The goal of the workshop was to bring together representatives from other metadata standards to provide an external review of current DDI work, with an emphasis on the model-driven approach, the production framework, and the substantive content of the standard. Progress was examined in light of the high-level goals and design principles for the model-driven specification.

The model-driven endeavor is intended to be a robust and sustainable solution for the future and a solid foundation for expansion of the standard. Therefore it is important that DDI is developed carefully and in alignment with best practices in metadata development, which makes input from other communities essential. The DDI metadata standard should not exist in isolation but should participate with other standards in a larger framework characterized by interoperability and collaboration.

The workshop provided a forum for representatives of the DDI initiative and other standards to learn from each other and to share successful practices. The invited experts gave freely of their time to help improve the DDI specification, and this document provides a summary of their recommendations. Small teams worked on a variety of topics during the week, and the summary is organized according to those topics.

Design Principles

A small group revisited the design principles that had been compiled in 2012. The goal was to reorganize the principles and make them more concise to ensure that they were easily understandable and that individuals working on the new DDI specification could quickly call them to mind. The group also sought to compare the principles with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to understand any overlap.

The principles were re-envisioned in this way:

Design

The model

- is developed in an agile, modular and iterative manner
- is responsive to community needs to produce actionable metadata
- should balance complexity with functionality and understandability
- is extensible and strives towards compatibility between different versions
- is maximally interoperable with relevant community standards
- supports a plurality of implementations

Documentation

The documentation of the model

- is clear, complete, and timely
- is concise, comprehensible, accessible, and useable by multiple communities
- provides justification for design decisions
- provides reference and functional perspectives

Capability

The model and its documentation

- support the discovery, reuse, exchange, and sharing of (meta)data
- support the capture, production, management, and analysis of (meta)data
- support audit and reproducibility across the (meta)data lifecycle

The result of the reframing was a more concise and more integrated set of principles. The principles were reduced in number and better organized for application. They were written to provide a basis for both design and evaluation of the model.

A further effort involved evaluation the principles against these criteria:

- Clarity
- Self-contained nature – that is, does not require additional lookup to understand the meaning
- Implementability
- Relevance
- Broad applicability
- Justification provided; related efforts

Next Steps

The group lacked time to thoroughly evaluate each principle against the criteria, so that step should be taken. Also, specific metrics for assessing the group's conformance to the principles need to be created. Finally, the principles should be reviewed by the Modelling Team and the Technical Committee and posted for full transparency.

Class Library and Versioning

The goal of this breakout session was to develop an approach to arranging and managing the DDI4 class library and specifically to evaluate the notion of functional views, which are subsets of the library intended to support and document a specific task, such as documenting a simple dataset. The group also sought to determine how best to version the library and elements in the library and how changes might affect implementation (tools) and interoperability (between tools).

To answer these questions, the group looked at the architecture and versioning mechanisms discussed so far in the Moving Forward process. Furthermore, they looked at UN/CEFACT (United Nations Centre for Trade Facilitation and Electronic Business) and UBL (Universal Business Language), which are both model-driven standards with an XML syntax, and at CaBIG (and related projects.

It was clear to the group that there were some options when it comes to versioning, the management of namespaces, and the planning of releases. One set of possible solutions is suggested by UBL while the experience of CEFACT and CaBIG (Cancer Biomedical Informatics Grid) provide some examples of how not to approach these issues.

Following are the recommendations of the group, which are often posed as a set of options with tradeoffs:

Versioning

Object-level granularity is desirable but needs to be balanced against resource demands. Also, effective quality assurance must be emphasized.

Release Strategy

A key issue here is the impact of new releases on implementers. Thus, the strategy requires a balance between size and frequency. One possibility discussed was large, infrequent releases containing several functional views; however, it was pointed out that we should learn from the CEFACT experience that binding version to namespaces makes this impractical for implementers.

More frequent view-level releases would be possible if we do not impact implementations too severely (that is, bind only major versions to the namespace and supplement with a version element, etc.).

Namespaces and Version

Again, options were presented. Using the approach of indicating a major version in the namespace URN supplemented with an XML element giving full version information will require processing of XML in order to know which version of the XML schema to use. An alternative is to rely only on a version element without any version-binding in the namespace. Another option is to use a totally version-bound namespace as is the case with DDI Lifecycle.

Backward Compatibility

The group agreed that if both of the following types of backward compatibility are not met, the release should be classified as a major release.

- *Data value level* -- Data can be transferred from one version to another without risk of losing or changing the meaning of data (i.e., DDI-Lifecycle compatibility)
- *Tool/Instance format level* -- All instances of a previous version are valid according to the new version of the XML schema (i.e., DDI-Codebook backward compatibility)

To guarantee backward compatibility, testing procedures may be necessary.

Next Steps

The modelers need to review the options and make decisions for the way forward. The development team should also give thought to how users might influence/support the change process.

Production Workflow

The goal of this group was to evaluate the production workflow and tools being used to build new versions of the model-driven specification. To that end, the group sought input from experts about their experiences with automated production workflows to get from a Platform Independent Model (PIM) to the actual implementations of a standard (in preparation for the upcoming sprint in Copenhagen).

Recommendations included the following:

- The configuration should be separated from the transformation scripts
- Simplicity is a key rule
- The Master should be a git repository that contains ONLY plain text files (XMI, RST, configuration files, XSLT, etc.). By implication, Drupal/Lion (the current mechanism for generating the model) cannot be the single source of truth. The canonical version is the XMI (PIM) and the associated documentation.

Next Steps

- Discussion should continue on specific points such as:
    - Flattening XMI
    - Translation of XMI to XSD or OWL, etc.
    - Transformation configuration (RuleML, etc.)
    - Use of Schematron for views / profiles
    - Updated visualization of the production flow that combines both the generation of the schemas and the documentation

Patterns

The objectives were to review the process and collection model patterns, with the idea that they could be more modular, and ultimately to rationalize and formalize the model and its relationships. Recommendations were made in three areas:

Process

Several issues were identified regarding the process pattern. These pertained to the way in which roles were related to process steps, and how time was related to the process model. It was suggested that we look at the "path" (semantic trajectory) pattern from VOCAMPS to add more rigor to our modelling. PROV-O might need to be revisited. Changes to the Process model as a result of the feedback from the Data Capture group are also required.

It was also suggested that the group explore other existing ontologies/ODPs might exist that might be leveraged.

Bindings

Bindings were discussed, with an explanation of the collections pattern within the process model. Again, the VOCAMPS path pattern might be usefully applied.

Collections

The group also looked at collections, but did not identify any major issues. They did confirm the value of XKOS, the DDI RDF Vocabulary built to extend SKOS.

Next Steps

The group noted that some open items remain to be addressed:

- Granularity of the model
- Whether Participation/Role should be added to the process model
- Need for creating the binding relationships for Inputs and Outputs to refer to Types and Instances of data/metadata objects
- Do we build a single pattern for both prescriptive and historical process instances? How do we manage time in relation to process?

General Binding Issues

- Greater use of languages like SHACL and Shex which are more expressive than OWL and can validate.
- Conversion between XML and RDF instance data is still a major issue. A simple dump does not take advantage of the features of RDF. Need to know the expected structure of the transformation and capture the rule in RML or similar formal language.
- All objects should be identifiable.
- Need to further explore a reliable round-trip of identifiers across instances. Do we have a commitment to provide resolution of URN's over the long term?

RDF Binding

There are two approaches:

1. The data-driven approach would have a continuum from the data-driven approach (give me a data set, generate a metadata publication with annotation/automation tools). (Example Australian Data Archive)
2. The model-driven approach would create views that would be used for generating the entire data collection methods, processing/cleaning, and publication apparatus (so the necessary information for publication would not require any human intervention [in theory]. Australian Bureau of Statistics has such an initiative. Alternative, British Cohort Study.

There is new a "gap" between #1 and #2 that does not allow the process to be fully invertible. To bridge the gap in knowledge data processing between the knowledge about raw data and the published data.

RDF representation of DDI (meta)data should be as simple and maximally useful for query answering and linked data. This means:

1. **Identifiers** All data should have their own identifier. this supports the linked data vision
2. **URIs vs URNs** Use URNs for all data items. Assign a de-referenceable URL that uses the identifier as a service for fulfilling the linked data vision.
3. **Graphs** Use a graph to store the triples about an individual; explicitly relate the individual to the graph. annotate the graph with record metadata
4. **Record vs individual knowledge** Assign a versioned identifier for the record and a version-independent identifier for the individual. Couple the right metadata to either the record or individual.

5. We will need a formal language (e.g. rules) to capture reversible structural transformations from one syntax to another (e.g. XML to RDF; xml to FOAF/DC/DCAT/etc.; XML to SIO). A generic transformation would not take advantage of the beauty of RDF.
6. The UML diagrams/XSI should be transformed into an RDF constraint language, rather than an OWL ontology, which cannot be used for validation.
7. There should be an attempt to consolidate the concept model into a tighter reusable core; an ontology-based model can help here.

Use Cases

Starting with the set of 3 uses cases (Longitudinal, Blood Collection, Data Linkage) prepared at Dagstuhl create a test suite of use cases that content and modeling groups can use to test out their models. This would also provide a basis for evaluation of content models by the Modelling Team.

Biomedical and Social Sciences

The group agreed that a collaborative project focused on making standards' vocabularies more widely accessible and understandable would be a good way to continue the work in this area. Michel volunteered to articulate a Specific Aims page to shop around that summarizes the problem and the approach. Additionally, we can propose a simple pilot to help build machine-actionable descriptions of cohorts.

Data Capture

The discussion made clear a few things about Data Capture, how it is modelled and how it will interact with other DDI models/views:

- Data Capture does not only have to document data acquisition at a macro, study, or instrument level, it has to (and can) document the acquisition of individual datums.
    - In turn, this supposes an ongoing interaction with the Data Description (and Process) model throughout any data capture event.
- Question is a type of Measurement, and there are many other "types" of measurement that do not necessarily need individual objects in Data Capture. Measurement can be extended.
    - One promising alternative is to develop a controlled vocabulary for Measurement that describes different measures or capture types. This list could be extended indefinitely. This would allow the Data Capture model to be generic enough to be broadly applicable, but also allow an easy entry point by different domains to describe their specific data capture types.
- Certain data capture types (e.g. data linkage) need to document not only the data being acquired but also its metadata.
- The Data Capture model seems to work for the use cases we applied to it (simple survey and protocol (blood assay)). There was no objection to or criticism of the elements in Data Capture.

Data Description

The discussion highlighted the different ways in which different domains understand and describe the basic concepts underlying the description of data.

- **Different communities use different terminologies.** efforts need to be made to provide self-contained explanations for terminology, and mappings to other community concepts e.g. semantic web
- **Different terminologies may not have equivalent concepts.** some terminology may not be relevant, mapping to other terminologies will need to highlight those elements which are interoperable

Progress on the development of the model, in particular **Viewpoint** which is an assignment of the Roles Identifier, Attribute, and Measure to DataPoints in a Record. In a ViewPoint each DataPoint has a single Role. This seems to correspond to a Data Structure Definition in SDMX and hold promise to assist in mapping to other terminologies such as RDF.