October 20, 2014 Shared minutes

# Day 1 Dagstuhl

## Background

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 plans to address 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, we need structured metadata.

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

* Which elements do we need?
* What objects should have metadata?
* 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?

In terms of which items relating to data should be cited, we can think about:

* Data files
* Segments
* Qualitative data segments
* Extracts from dynamic data
* Replication data

We may also want to cite various aspects of instruments and the data collection process:

* Questions
* Categorizations
* Equipment
* Software
* Algorithms

Procedures and Conceptual components may also need citing. In addition, we want to think about data reuse and the infrastructure for forward and backward searching.

Other questions for the group:

* What contributor roles should be captured?
* How can we come up with a flexible controlled vocabulary for contributor role?
* Should we look at the Generic Longitudinal Business Process Model (GSBPM)? This may provide some ideas for roles as we look across the lifecycle of producing longitudinal data.
* How might we measure degree of contribution? Some ideas: Order of listing; Percent of total project; FTE; Importance to project

The group will look at the current citations elements in DDI. Numeric fingerprints/qualitative fingerprints may be possible elements of data citations that we could add.

We also need to think about tools to make it easy to capture information in the research stream.

Having the capacity to track the provenance of questions would be very useful. Also, what happens when you want to cite multiple combinations of objects?

## Question 1: Which DDI objects should have citation metadata associated with them?

All identifiable objects. Every identifiable object should have a specific set of citation metadata as an optional property of the object. It is implicit that there is a relationship to the larger entity of which the object is a part. There is not a need for a specific information object indicating that relationship.

Rationale for decision: Choices of what to cite are are social choices - chasing this by choosing some and not others is a never-ending task. It is not the standard’s job to define these social questions as different organizations will have different needs. The standard should enable citation as practiced by different organizations. The structure should be there, but populating the citation metadata is optional.

Related Rule: You can cite something only when IsPublished is True. In development you can reference something, but you can’t cite it formally.

The modelers may want to focus on only those objects that are administered/versionable.

### Discussion

Each object in DDI 4 will have an identifier, but not all of them have additional metadata and are administered or versionable. Which objects should be citable? This is a social issue.

We need to separate the concepts of reference and citation. In general, we need to be clearer about the distinctions among:

* A citation in the sense of a pointer to an object (e.g., Title, Creator, Publisher, PublicationDate, Identifier) - as in citing a dataset in a journal article
* The collection of citation-related objects (a record) related to an object (in DDI Codebook and Lifecycle this is called a Citation element)
* Pointing to an object via an identifier (a reference)

By “referenceable,” we mean “addressable” -- anything that can have a URL. Note that all DDI4 objects have unique identifiers. When an object is reused, we reference it within its context.

This could be an opportunity to make things more consistent beyond name, label, and description.

Upper level of containers should come with full citation.

Creation of original object but then it may be reused.

Likert example - The citation metadata for a widely used object should have pointer to the seminal paper (where one exists) - alternatively where there is no paper directly describing the object there should be a pointer to its first use, the instantiation

heavily managed objects - e.g. medical classifications

We already have a socially acceptable way to cite through writing a paper and citing something in the references section, but we want to give credit to others who have contributed to a dataset.

DataCite was set up to make mainstream the convention of citing data without scholarly papers. There are 4 million DOIs to things that don’t have papers.

A core idea of DDI is that a variable as an object could be put into a repository and reused; is there a way to attach the credit to that variable?

The Altman and King paper on citation has a section on “deep citation” and provides an example of using three variables from a dataset for a table in a publication. Variable names are added to the citation of the data file.

When creating an analysis file, we should reference the original data file and describe selection criteria for the analysis file through syntax and codes.

Proposal: Give every DDI object the “Citation object” metadata (but call it something else). There is the potential for a “creator” at the data file level and at the variable level, for example. There is a relationship between the variable and the data file, so the best practice would be to treat it by inheritance and follow the relationships to the dataset.

Administered objects need citation metadata, but do the others?

Do we need to specify when in a process this information needs to be populated? In DDI, this happens when the isPublished flag is set to True.

As some terms are not clear, we should provide a glossary defining what publishing means and what citation means. Publishing in this community means making the object available and this is static. The definition for “isPublished” says: “Indicates that the maintainable will not be changed without versioning and is a stable target for referencing.” This definition is clear but instead of isPublished, we might think about using another term -- for example, “IsRegistered”, “IsFinal”, “IsCommitted”, or “IsVersioned”.

In clinical trials research, something might become citable when it becomes part of the audit trail.

## Question 2: Which information elements are needed for comprehensive data citation?

### Discussion

To answer this fully, we first need to generate a document with a list of all citation information elements from 2.5 and 3.2. [Sanda to do this.]

OthID is the contributor element in DDI Codebook and maps to Dublin Core contributor. There is a Contributor element in DDI Lifecycle.

You have an addressable object and you can associate the citation metadata with it. The idea was raised that we could add an actionable citation pointer to any object and the user could then decide what elements to use. Users would structure the information in the referenced object.

A hierarchy of pointers (typed) could degrade gracefully from DDI to something else, to a lesser something else; could be: #1 structured citation metadata in DDI; #2 link to structured citation located elsewhere but is in a recognized/sanctioned format; #3 link to citation in an unknown format.

Could we borrow DataCite as the core metadata for citation? This is focused on data but what about other kinds of objects?

The consequence of having a large list of attributes is that the large list may put a burden on implementers.

The citee has all the information about the object and can provide instructions on how to cite. The citer can use these instructions.

Allows for corner cases

To make this machine-actionable, the citee would provide the bare minimum requirements. The actual citation might be unstructured. However, we don’t want to break backwards compatibility with machine-actionable objects.

Citee describes how to be cited.

Edge case: citee provides citation to different object outside of the DDI, e.g., they want the paper cited as a proxy for the dataset.

A use case might be to find all roles a specific person has played with respect to published research data. This requires structure that the citee can provide.

JK: proposed...

Citation involves objects, cite-ers, and cite-ees.

Cited objects have, at minimum, a URL (with embedded URN).

Cite-er determines citation metadata appropriate to stage

(embryonic, adolescent, mature, version 1, version 2)

...

A set of use cases was developed:

1. Sally Scientist, working alone, wants to create a data file with a codebook including enough information about how to cite the work in an article she’s writing.

The citable object is a [DataSerialisation](http://lion.ddialliance.org/ddiobjects/dataserialisation). What is in DDI now may be adequate, but is currently scattered through the structure. Right now her role in the components of the project is only determined by implication from the study level citation.

She gets a DOI and generates a UNF. Should the Fingerprint go into the citation? Yes, as an optional element.

For a file-level citation to be possible, what elements should she use?

Title, Creator, Publisher, PublicationDate, Identifier (DataCite Kernel) also ResourceType? There may be a separate citation for the codebook (different use case).

1. Same dataset but including enough information to credit many co-workers defining the roles of those co-workers and degree of contribution.

Elements needed: Title, Creator, Contributor(s) with roles, Publisher, PublicationDate, Identifier

The codebook would have the complete set of metadata about the contributors, including degree of contribution.

Degree of contribution is difficult to address as there is no right or wrong way. Suggestions made included author order, primary-secondary, 0-1 weight with key value pair defining the semantic.

1. Same dataset cited by a scholar wanting to reference the work in a publication.
2. Colleagues in a given field need to share references to the same embryonic data.
3. Data citations harvested for indexing activity.
4. Citation for data files received and then harmonized. Original citations and plus citation for the harmonization effort.
5. Citation via OAI-ORE resource map.
6. Joe creates a variable to be stored in an archive for others to be discovered and reused.