Citation Minutes October 21

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?

Question - If citing the data file should that include a reference to (identifier for) the metadata?

Use case structure:

* scenario (include actors, systems, goals)
* citable objects (from DDI model)
* elements used, where they come from
* example citation(s)
* discussion

Continuing the Use Cases:

**Scenario**

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.

**Citable Objects**

Citable object(s): [StudyUnit](http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/); DataFile;

**Elements Used**

Title, Creator, Publisher, PublicationDate, Identifier (DataCite Kernel) also ResourceType?

**Example Citation**

Scientist, Sally. Coffee Consumption at Schloss Dagstuhl. Leibniz Institute for Informatics [publisher], 2014. <http://doi.org/10.3886/11223344>

There is citation metadata at different levels. Should there be a rule for inheritance built into the model? DDI can enable inheritance by reference but tools are needed to enforce this.

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?

There may be a separate citation for the codebook (different use case).Citation located in Study Unit. Citation information a components of the study could reference the study level citation.

Need for best practices. When the user goes to a landing page what should they find there.

The group reviewed the elements and attributes that are currently used for citation in DDI Codebook and Lifecycle. Copyright should probably be moved to another location as it is not really related to the citation. Publisher has attributes of producer and distributor.

**Scenario**

Same dataset but including enough information to credit many co-workers defining the roles of those co-workers and degree of contribution. Bill does data collection, Jane does data analysis (statistical), and NIH was the funder.

**Citable object(s)**:

StudyUnit, [DataCollection](http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/), Instrument?, Statistical analysis (nothing in DDI now about analysis)

**Elements Used**

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

**Example Citation(s)**

Scientist, Sally, Bill Smith, and Jane Doe. Coffee Consumption at Schloss Dagstuhl. Leibniz Institute for Informatics [publisher], 2014. <http://doi.org/10.3886/11223344>

Bill and Jane’s respective contributions would be acknowledged in journal Note/Acknowledgments, which are increasingly structured, e.g.:

**Author Contributions**

From article in PLOS (http://www.plosone.org/article/authors/info%3Adoi%2F10.1371%2Fjournal.pone.0109687):

Conceived and designed the experiments: DR ZE. Performed the experiments: DR ZE. Analyzed the data: DR ZE. Wrote the paper: DR ZE.

**Discussion**

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.

Citation information for funder at the Study Level - in DDI3.2 this would be in a separate element)

Citation information at some object at the DataCollection level for Bill’s role

Citation information at the Methodology level? for Jane’s role

Modelers need to ensure that all roles are inheriting from the same core class. Make Creator a subset of Role so that we are not missing any roles.

1. **Scenario**

Same dataset cited by a scholar wanting to reference the work in a publication.

**Citable Objects:** StudyUnit; DataFile (PhysicalInstance)

; Codebook(DDIInstance)

(Some people do cite codebooks because they don’t know that they should cite the data.)

**Elements Used**

Title, Creator, Publisher, PublicationDate, Identifier

**Example Citation(s)**

Scientist, Sally. Coffee Consumption at Schloss Dagstuhl. Leibniz Institute for Informatics [publisher], 2014. <http://doi.org/10.3886/11223344>

*Steve and Mary. and Sam, and Stu*

**Discussion**

Need pointers to both data and metadata. Need to know which data file (physical instance was used).

1. Suzie Student in Genomics generates lots of unpublished data over the course of 4 years and needs to share citations to intermediate findings for the purpose of review.

Suzie includes all the fields she can from an official citation, and adds “[work in progress, subject to change]” or “[work in progress, fixed as of 2014-10-21]”.

**Discussion**

Permanence categories might apply

1. **Scenario**

Sam Streamer uses data currency exchange data to normalize price data over time .

**Citable object(s)**:

StudyUnit, PhysicalInstance

**Elements Used**

Title, Creator, Publisher, PublicationDate, Identifier, last revision date, queryDateTime, PID to query/selection

If a subset was extracted, add a queryID with timestamp.

**Example Citation(s)**

Streamer , Sam. Normalized Coffee Price Data at Dagstuhl October 2013, DDI Alliance [distributor], [http://doi.org/10.3886/11223354](http://doi.org/10.3886/11223344)**,** 2014-10-21T10:24:15

Sam cites:

Bucks are Us, Streaming Currency Exchange Rates, Daddy Warbucks Inc. [publisher], Last revision date of 2014-10-21T10:00:00, http://BucksAreUs.com/?myquery

**Discussion**

Permanence Level should be in the metadata.

queryID,

1. Thomson Reuters Data Citation Index harvests data citations for indexing activity and discovery. *Michael, Jay, Wolfgang, and Jenny.*
2. **scenario (include actors, systems, goals)**

Dataset deposited in a repository (e.g., ICPSR) that puts metadata into DataCite Metadata Store and then is harvested by Data Citation Index (TR, a commercial service) using OAI-PMH.

DCI then processes the DDI and includes other sources to enrich (e.g., normalize to their proprietary schema and associate with citations to literature that it also indexes) and make it searchable in Web of Science/Knowledge. For example, they provide metrics of use and present data citations with papers that cite them.

1. **citable objects (from DDI model)**

DDIInstance (i.e., codebook), StudyUnit, PhysicalInstance, SummaryStatistics, QuestionGroup [there could be more…]

1. **elements used, where they come from** *We will finish this later!*
2. DDIInstance
3. StudyUnit
4. PhysicalInstance
5. SummaryStatistics
6. QuestionGroup
7. **example citation(s)** *We will finish this later!*
8. DDIInstance
9. StudyUnit
10. PhysicalInstance
11. SummaryStatistics
12. QuestionGroup
13. **discussion**

Do we supply citations of data with harvest (e.g., for DCI to act upon). Currently not in DDI?

Jay shared Aqueduct.

There is how DCI works now and then what more they could do if we gave them more/better metadata.

Jay can provide example of NIH with requirement for description of tables.

Caveat: DCI may not actually use DataCite MDS now.

Some make-believe.

Scales are licensed for purchase.

1. **Scenario**

Citation for data files received and then harmonized. Original citations and plus citation for the harmonization effort. *Wendy Thomas and Knut.*

**Citable object(s)**:

Study Unit, Data Sources for the Study

**Elements Used**

Harmonized work: Title, Creator, Contributor (role), Publisher, PublicationDate, Identifier,

Original data contributors: Contributor (role),

**Example Citation(s)**

Minnesota Population Center.*Integrated Public Use Microdata Series, International: Version 6.3*[Machine-readable database]. Minneapolis: University of Minnesota, 2014. <http://international.ipums.org> [Source Data Contributor: France, National Institute of Statistics and Economic Studies; Germany, Federal Statistical Office]

**Question:**

*What is the purpose of noting the sample sources?*

We have the strong feeling that this is not required and is basically a social/political motivation to acknowledge contributors to the overall project.

This should be handled by the example capturing the extract specification if one is citing the extract rather than the project database (IPUMS). This is the only technically satisfying solution.

*Should it be noted in the citation?*

If so, how to differentiate between the direct source (harmonized in this case)and the contributing sources (listed here as Source Data Contributor).

1. **Scenario**

Joe Vari creates a variable “AgeInMiilliseconds” to be stored in an archive for others to discover and reuse. Larry *and Jeremy.*

**Citable object(s)**:

Variable

**Elements Used**

Title, Creator, Publisher, PublicationDate, Identifier, Resource Type

**Example Citation(s)**

Vari, Joe. AMs - Age in Milliseconds, DDI Alliance Repository, 2014, <http://doi.org/10.3886/11223345>, [Variable]

ResourceType:Variable

**Discussion**

Note that the identifier will include a version. In this use case the the variable is likely to evolve over time.

For others to reuse the variable they would likely need a reference to the Represented Variable. On the other hand an InstanceVariable might be cited in the reuse of a dataset or inn a repllication study where the the original variable was comapred to a new instance variable.

Scenario:

Robert Researcher uses an extract from a larger published dataset (e.g. IPUMS xxx) done in a controlled access environment. Robert is able to construct a citation identifying the selection criteria for the subset of data that he used. *Sanda and Ornolf.*

**Citable Objects:**

DDIInstance, StudyUnit, PhysicalInstance (data file)

**Elements used:**

Author: Robert Researcher

Title: Subset of Original Title

Publisher: Robert Researcher

Publication Date: Date when new subset was published (2004)

Related Resource:

 Author: Census Bureau

 Title : IPUMS

 Publisher: Census Bureau

 Publication Date: 2000

Relation type: Is Derived From

Selection criteria: List of variables selected

 Set of KEEP criteria (case selection)

**Example citation:**

Researcher, Robert. Subset of IPUMS. Robert Researcher 2004. Is Derived from: Census Bureau, IPUMS, Census Bureau, 2000.

Discussion:

Note: a link to the selection criteria is important

MPC provides the the selection criteria and keeps the selection criteria (time stamped and versioned)

Restful interfaces might be used

Discussion of permanence, re: [NLM permanence levels](http://www.nlm.nih.gov/psd/pcm/devpermanence.html)

-

[Citation.Contributor reference in 3.2](http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/schemas/reusable_xsd/complexTypes/CitationType.html#a5)

Brainstorming ideas for populating/refining contributor roles:

1. Use taxonomy from Allen et al. Nature 508 312-3 or follow-up workshop.
2. Are their other taxonomies or lists that we might look at? Perform a scan to identify (some already in [Google Drive folder](https://drive.google.com/folderview?id=0B5aS3-mIMlfkcWVCbUR6QWxDbEE&usp=sharing_eid&invite=COmIpYYH&tid=0B4BXffowUbWzSE1HUHM2MmVrZXc))
3. Research examples from current practice by looking at notes, etc. in published data papers and datasets.
4. Scan citation style guides and author guidelines and note roles.
5. Free brainstorming in the group.
6. Look at focus groups, case studies, data curation profiles, etc. from different research communities.
7. Identify some exemplars: in current practice, what it the most granular, complex citation or reference to data that we can find?

How important will a shared community list be? Do we need some sort of hybrid of basic list and additional controlled vocabulary?

Lifecycle events - agents in process model in DDI4 will be a source of roles

Will a controlled vocabulary for roles fit into some activity object in the lifecycle model?

GLBPM

Other notes

**We need to discuss inheritance of citation information.** Rules for this need to be codified. Could these be explicitly part of the model? Must enforcement be built into the tools?

We have to be careful about implications of inheritance - did the data collector design the collection or was just responsible for carrying it out?

DDI4 should cover capturing statistical analysis activity

National of medicine permanence ratings. almost everything subject to change. classification of dynamic datasources - growth by accumulation - older data don’t change, Characterize citation as replicable or not <http://www.nlm.nih.gov/psd/pcm/devpermanence.html>

“Default values or drop-down menus are provided wherever possible. This minimal set includes:

Title

Heading

Date Published

Date Last Modified

Next Review Date

Contact email address

Publisher

Rights

Permanence Level

Permanence Guarantor

Language

”

“**Permanent: Unchanging Content**

The National Library of Medicine has made a commitment to keep this resource permanently available. Its identifier will always provide access to the resource. Its content will not change. Example: Minutes of the NLM Board of Regents meetings

**Permanent: Stable Content**

The National Library of Medicine has made a commitment to keep this resource permanently available. Its identifier will always provide access to the resource. Its content is subject only to minor corrections or additions. Example: NLM Annual Report

**Permanent: Dynamic Content**

The National Library of Medicine has made a commitment to keep this resource permanently available. Its identifier will always provide access to the resource. Its content could be revised or replaced. Example: NLM Home Page

**Permanence Not Guaranteed**

The National Library of Medicine has made no commitment to keep this resource available. It could become unavailable at any time. Its identifier could be changed. Example: Frequently Asked Questions

”

Consider a Published paper - Digital library journal? IASSISTQ

# Roles

The Nature article roles are categories. Could we have a hierarchy within these?

We can approach this from the standpoint of the data lifecycle process to capture all the steps in the data lifecycle.

Allen et al. will be working with NISO as the next step to create a taxonomy with a standards body.

The group compared the DDI lifecycle events with the Allen et al. taxonomy of categories. Sanda created a list of roles.

A proposal is part of study conception.

Funding is in both.

The DDI list is more detailed. Methodology is broken down into several steps of the lifecycle.

The DDI list is about metadata-driven design and machine-actionability while the taxonomy is about attribution.

We could share the DDI lifecyle with the Nature authors.

We could look at all the Chicago Manual of Style to see what is already being acknowledged.

Distinction between categories and roles; start with categories (map between DDI and Nature and identify matches, mismatches, and gaps), talk to Nature guys, populate roles.

Questions for Nature taxonomy authors:

Is there an effort to make this information machine-actionable?

What about reviewer (for funding proposal or article)?

Mapped DDI Lifecycle DVG to Nature:

<https://docs.google.com/spreadsheets/d/1cHTb2tth1G9fjP4vXSmeBgBDSxpYq6Ykxikn4LQbslw/edit#gid=0>

Also should flip around mapping: use Nature to potential improve DDI CVG lifecycle events and roles.

Concern about the difficulty about parcelling out credit accurately. “I have to get listed in this role as opposed to that one”. Who will do these allocations of role, how will will it be done.

Should we keep the list of roles small? Should we consider researcher burden?

What about those who now are not listed first on the author list? Details give them an opportunity for credit.