Citation Minutes October 22

**Larry Hoyle, Wolfgang Zenk-Moeltgen, Sanda Ionescu, Jay Greenfield, Sam Hume, Stuart Weibel, John Kunze, Michael Witt, Jeremy Iverson, Barry Radler,**

**Wendy Thomas, Mary Vardigan**

**From Yesterday:**

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?

Original Questions:

* Which elements do we need?
* What objects should have metadata?
* How should reuse be handled? Relates to infrastructure
* What infrastructure is needed for location?
* Which need controlled vocabulary?
* What special information is needed for the citation of stream resources? (revisit what additional information at host site)

**Questions for Nature taxonomy authors:**

Clarify distinction between categories of contributors versus roles.

answer: categories (journals can define and populate more granularity, e.g., roles)

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

Have you thought about degree of contribution?

Are you open to adding new terms?

Has the planned workshop been held yet? What are future plans?

Do you see the DDI lifecycle roles as an appropriate application of your taxonomy?

Broader lifecycle, e.g., 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.

**Proposal:**

Use case 10

Data from this workshop

 Minutes (qualitative collectionn)

 quantitative dataset(s) - can’t survey ourselves - no IRB approval,

 but we could make a dataset on our observations here (weather, food?)

 some parts without the PIs’ contribution, credit for variables? metadata?

 Different measures of degree of contribution. How difficult? Potential for hard feelings?

 To be archived - KU Scholarworks (with handle) ICPSR (with doi)

 Generate DDI3.2

Make a SAS quantitative dataset from the minutes

**Notes**

does reuse imply using the data integration with other data or just e.g. another analysis

rey on discovery - robots search for references RDA working group. (2 other groups, RMap and the National Data Service)

Should DDI Alliance take on a advocacy role re citation?

Elements needed

Should we borrow from DataCite their list of elements and attributes? How does DataCite metadata relate to DDI? Does DDI lack the ability to describe relationships between

1) new dataset is subset of old dataset does citation of new dataset indicate parent dataset

1. harmonized dataset

not in the citation in a publication but in the supporting information

E.g. ICPSR fertility dataset

“isDerived from” etc

DDI has these relationships how do you serialize

Will everyone using DDI use Datacite? no

DDI4

vocabulary for relationships?

Provenance chain

Is Derived From

Is Part of

Can furnish relationships inside DDI; how can we expose them so they are serialized?

One way of linking to another standard is to say go here. Steal from DataCite?

Do XSLT of DDI and DataCite to find gaps

Recommendation:

There are hundreds of standards for serializations of citations. Publisher (/funder? / data provider?) requirements will determine the serialization of the citation metadata. Not our role to recommend one serialization style. We need to ensure that the information is there to be serialized.

We should be in the modeling business not the serialization business.

**Should we remodel and make creator and publisher a role of contributer ?**

Dublin Core could support this approach this would have the advantage of having the degree of contribution weight having the same scale

Should degree of contribution even be part of our role?

A simple citation - What is core? There could be a best practice that some should be there.

A minimal serialization is needed to test our theories.

**Question 1: Which elements do we need for citation?**

**Minimal set**

Title

ID

Contributor (OtherID) (Role attribute -- Needs controlled vocabulary)

-Author (creator)

-Publisher

-Publication date

Version

 -Number

 -Date

 -Responsibility

Resource Type (Needs controlled vocabulary) (could be derived from DDI or from external sources) (Jay)

Locator

Recommended but not required? - Pointer to metadata -- Need Wolfgang’s guidance on this

What will this imply for data creators - is this enough of a burden to make this pointer impractical. Alternative link to metadata is loaded onto the locator

NIH DDI - requirement for URI for metadata

UML model allows extension beyond minimal set

Short list of conditionals for additional elements under sets of circumstances

These instructions could be in a profile - which in turn could be used for validation.

**Additional set of recommended**

actionable link to dataset

Copyright

(Access restrictions)

**Additional set of Objects needing development**

Permanence or Stability Information (NLM vocabulary?)

 multiple dimensions - Permanence of the object, permanence of the identifier, stability of the object

Data Fingerprint (enable this at the citation level? could this apply to DDI elements (e.g. a question?) - lack of agreed standard, but stable possibilities (UNF reference Altman et al )

actionable link to other metadata (part of locator?)

Permanence -- is this misleading?

Streaming is a separate category from the permanence levels. There may be future work related to permanence needed for streaming resources.

Big Data: Don’t get analysis dataset until you push it through pipeline -- at what stage is data sitting?

Data Lake: Key value stores where the value can be anything. Warehousing data without worrying about the data type. But to use it you need to know something about the data type.

Library of Congress vocabulary of roles?

Citing algorithms as part of process -- this can happen

Qualitative data segments -- they should also get citations; sql queries also

Replication data should be considered like any type of data in terms of citation

Terminologies or dictionaries apply to datasets. useful to cite them - link in metadata for dataset to link to terminologies or dictionaries

Replication datasets - is there an object that binds the data and the processing information

How do you generalize citation to the medical instruments that Jay describes? Should we distinguish between citation and annotation?

We have minimal set and a DOI -- how do you get to the supporting metadata?

DataCite - persistent identifiers to objects, infrastructure to resolve object

In DataCite the goal was to give persistent IDs and there is a Metadata Store to hold the supporting metadata. DataCite provides stable infrastructure to resolve the DOI but the resource holder takes responsibility to keep the IDs up to date.

Are we duplicating metadata stored in DataCite? Responsibility of data owner to synchronize

We need a clear mapping to DDI (DataCite metadata document - mapping was complicated)

DDI already had duplication DDI Title and dc:title. If we get rid of Dublin Core still need a clear mapping to it.

We need human-readable reference information to support the machine-readable ID. Mandatory elements are stored together with the larger metadata set.

Try to align with schema that others use.

DataCite could have role as contributor type.

Which roles are more important than others - would you name all of the people involved in a survey

not all in serialized record but maybe all in the record.

(8K authors might be too few)

What is the minimum set of information objects to generate a DataCite doi reference

Should we be thinking of how to generate a DataCite record from DDI?

Have core and extensions that you would use to programmatically to generate a citation?

Each data center can decide which objects to assign citations to

Minting a DataCite DOI is one form of serializing a data citation.

Should we try to make them identical? Would this be a big win?

Five mandatory for DataCite -- we add version and resource type. Should we make these optional and push them “down” into the metadata record? +1

DataCite record like a catalog record no richer metadata - schema has related identifier with type of hasMetadata. Wolfgang created a DOI for the codebook of the World Values study. Now the DOI points to the rich metadata (XML) and avoids the landing page.

Content negotiation mechanism? but landing page has different content than XML

Could we put something on the landing page that a machine could read (like a meta tag?) to access the richer metadata? These are two different systems, though.

There is a more general problem. We are registering DOIs for landing page. What if we cite a variable? Don’t have a landing page for that.

Citation to a variable can’t yield a valid DOI?

From a modeling point of view it is easy to articulate the requirement that all objects should be citable. But citing a variable may not make sense as a DOI may not be appropriate for a variable. Practice is different from modeling. For some things you need human-readable information and for others not.

Citation with option of name value pairs?

All nodes have a type? Attributes depend on the type?

Park

Metadata repository expected to keep metadata up to date, but errors are always there. Not sustainable in the data world?

**Interview with Micah Altman**

**Presentation**

authorship in scholarly presentation average number of authors per publication increasing, commonly serialization only lists some

contributed equally?

order has no agreed upon meaning, varies among fields

improved analyics? connections among authors, …

acknowledgement practices vary

Authorship statements collected by journals - (limiting misconduct)

Who is recognized as an author varies across vendors even within a field

Objective - define source of contribution typically recognized (whether or not authorship)

Economics - alphabetical - early listing has advantage

IWCSA workshop 2012 -

contribution taxonomy

another dimension - what you made a contribution too - paper , dataset, section of work

intended to apply to natural, medical, physical, health, social science research

intended to be usable in publication environment

could ask questions - who was software developer?

Publication workflow - focus on corresponding author

surveys, analysis of acknowledgement statements

some categories controversial as authorship (e.g. funding acquisition)

pilot study with contributing authors

how much they had contributed was missing - level of contribution

(lead, equal, supporting) ←-

 Dec 2014 meeting

 CASRAI dictionary

scale issues - 3 authors vs 100 authors

contributership relates only to a particular subset of the work.

Citation of versions??

with each article a packet of metadata

**Questions**

*Clarify distinction between categories of contributors versus roles.*

The way they formulate this is that role is essentially incomplete subcategories of the taxonomy. What each person who had a relation to the work did in relation to the higher level.

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

contributor statement qualitative - would like to see it structured at least in terms of taxonomic categories

Ideally every contributor has an ORCID

*Publish the taxonomy in a way we can reference it?*

Yes working with CASRAI and to a lesser extent NISO both standards-based organizations. Large dictionary that is increasingly comprehensive. Integrate taxonomy into official CASRAI standard dictionary. Will be schematized.

*If we have ideas for new terms or definitions, would be open to them?* Yes

Send comments to Amy Brand and Micah Altman

*Broader lifecycle, e.g., what about reviewer (for funding proposal or article)?*

Doesn’t call out specifically peer review

Look to funders to give more acknowledgment to reviewers

*Have you thought about degree of contribution?*

yes see above - lead, equal, supporting

Was population primarily publishing data or article?

Almost certainly traditional publication. It would be interesting to see taxonomy applied to that

*Are you open to adding new terms?*

yes sooner is better

Has the planned workshop been held yet? What are future plans?

Do you see the DDI lifecycle roles as an appropriate application of your taxonomy?

*Any data publishers in the sample*?

already published authors - very tiny data publications

Micah:

Main problem: ways that authorship has changed in scholarly communications. Number of authors increased.

The order of authorship has no generally agreed upon meaning

Improve analytics

New measures become feasible

Measure new things

Reduce errors in standard analytics like impact factors

Acknowledgments are not authorships.

Citations ambiguous.

Journals collect qualitative author statements not available for analytics.

For limiting misconduct.

Who is recognized as an author differs from journal to journal

Objective is to define sorts of contributions typically recognized not define authorship

Author statements not typically in structured form

Research shows name order effects are important

Has an effect on careers if your name comes earlier in the alphabet

Workshop 2012

IWCSA Workshop 2012

Goals for Prototype Taxonomy

Describes types of contributions

Cover broad domain of research: natural, medical, physical, health, social sciences

Theoretically justified

Aligned with researcher behavior

Usable in publication environment

Not goal of describing all ways to make a contribution

Hope that taxonomy can be applied in this way

Number of steps from abstract taxonomy to something you can use in a real publication environment

How to elicit information about people and integrating with publisher workflows difficult

Most of workflow centers on corresponding author which is somewhat limiting in looking at contributorship

Keep this in mind when thinking about what interventions are possible in this space.

To develop prototype there were existing specifications especially in medical field

Empirical analysis of practice

Looked at what terms researchers used to describe their contributions qualitatively and which terms mapped to existing terms in frameworks

Checking this against acknowledgement statements

Ran a pilot test with select publishers

Generated taxonomy

Some categories are controversial as authorship in particular fields

Not claiming that these contributions should be authors but are evalutable and can be usefully tracked nad some stakeholders are interested in them

Designed survey to test whtether authors contriubtions can be assigned into a series of specified roles

Reflected on the process. Got a few hundred responses from people who found it reasonably easy to do this. Focused on articles wiht 7 or few authors.

Had generally positive comments about formalizing the process.

It was clear that there was at least one thing missing: people wanted to state how much different contributors had done. There is another dimension of level of contribution in each category. Something on this is needed for people to consider this adequately expressive.

Working with CASRAI and NISO to formalize this. There have been minor changes in wording with the exception of adding level of contribution as another taxonomic dimension.

Community meeting at CNI. WIll be incoroporated into CASRAI dictionary. PLOS will be one of the early adopters. Accredited standard later maybe.

Test with larger community

Develop guidance for use of taxonomy -- examples of questions you can use to elicit information about contributorship; how implement online

Socialize the taxonomy in the community

**Afternoon Session**

Discussion with Barry Radler re: MIDUS

The question of: “where did specific measures and questions come from”? can arise. A formal approach to developing an audit trail for these objects could be based on a citation element attached to each object as they are versioned.

Another question is “Who brought this question into the survey?” this could be tracked with a custom attribute in a citation element.

Jay noted that it is important to agree on a place for this information. It is also important to balance the collection of the information against being burdensome to researchers.

Copyright information, and perhaps access controls are important for questions in some contexts. Some questions have restrictions barring the publication of their wording or even the instructions around them.