Supporting Extended Citations in DDI4

North American DDI Users Conference
University of Wisconsin, Madison
April 2015

Larry Hoyle, University of Kansas, Institute for Policy and Social Research

Mary Vardigan, Inter-university Consortium for Political and Social Research (ICPSR)
Dear Colleague Letter - Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution

Date: April 11, 2014

National Science Foundation
Directorate for Social, Behavioral & Economic Sciences (SBE)
Division of Social and Economic Sciences (SES)
Directorate for Computer & Information Science & Engineering (CISE)
Division of Advanced Cyberinfrastructure (ACI)
Role

Citation and attribution:

- Novel mechanisms for citation of software and datasets as distinct products of scholarship, promoting standards of academic credit and rigor for these cyberinfrastructure components.
- Novel citation methods for new forms of publication and scientific expression so that researchers are able to ensure their work is citable, and others are able to discover and access it.
- Citation patterns that include a role for citations (e.g. to value activities such as “data provider/curator” and/or “software tool provider” alongside “data analyzer” or “computational modeler”), which can help create a credit market for data and software sharing.
NSF Grant 1448107

• Brought a group of data citation experts into a workshop at Schloss Dagstuhl, event 14432 (DDI4 Sprint)

• Goal was more nuanced citation of data and related objects in DDI

• Side benefits: Workshop familiarized the DDI community with data citation and introduced citations experts to DDI
Citation Workshop Group

- Larry Hoyle (PI), University of Kansas
- Mary Vardigan (Co-PI), University of Michigan
- Jay Greenfield, Booz Allen Hamilton
- Sam Hume, CDISC (clinical research data standards)
- Sanda Ionescu, University of Michigan
- Jeremy Iverson, Colectica
- John Kunze, California Digital Library
- Barry Radler, University of Wisconsin
- Wendy Thomas, University of Minnesota
- Stuart Weibel, Dublin Core
- Michael Witt, Purdue University
Citation vs the Information Supporting Citation

- We found ourselves hanging up on the word “citation”.
  1. The act of citing something
  2. Supplying the information needed to perform that act
  3. Supplying additional information once one identifies the resource

- DDI3.2 has a “Citation” object – a mix of the above
Content model elements (11):

AlternateTitle, Contributor, Copyright, Creator, InternationalIdentifier, Language, PublicationDate, Publisher, SubTitle, Title, dc:any

Included in content model of elements (20):

AuthorizedSource, BudgetDocument, Collection, DDIInstance, ExternalAid, ExternalInformation, ExternalInterviewerInstruction, Group, Item, LocalGroupContent, LocalResourcePackageContent, LocalStudyUnitContent, Origin, OtherMaterial, PhysicalInstance, ResourcePackage, StandardUsed, StimulusMaterial, StudyUnit, SubGroup

May contain elements by substitutions (48):

Structured Annotations and Description Types

• Confusion about what objects merit “citation”
• Structured annotations may be needed for different purposes, including attribution, administrative information, characterization information
• System of description types proposed
  – Citation type
  – Sourcing type
    • Example: (U.S.) OMB required information for vetting questions in federally administered questionnaires
  – Instrument type
    • Example: Instrument properties, settings
  – Dataset type
High Level Structure - W5HSP

Dataset type properties support traditional citation content and help to facilitate data reuse:

• Who
• What
• When
• Where
• Whether
• How
• Structure
• Provenance
DDI4
Annotations on Almost Everything

Most objects inherit from AnnotatedIdentifiable. Examples:

- DataStore
- Concept
- Question
- ConceptualVariable
- Annotation
- AgentAssociation
- PairedCodeValueType
- CodeValueType

The Annotation object will also have an additional property capable of containing administrative, characterizing, and other information structured by an external vocabulary.

- title :InternationalString [0..1]
- subTitle :InternationalString [0..n]
- alternateTitle :InternationalString [0..n]
- creator :AgentAssociation [0..n]
- publisher :AgentAssociation [0..n]
- contributor :AgentAssociation [0..n]
- date :AnnotationDate [0..n]
- identifier :InternationalIdentifier [0..n]
- copyright :InternationalString [0..n]
- language :CodeValueType [0..n]
- typeOfResource :CodeValueType [0..n]
- informationSource :InternationalString [0..n]
- versionIdentification :xs:string [0..1]
- versionResponsibility :AgentAssociation [0..n]
- abstract :InternationalString [0..1]
- relatedResource :ResourceIdentifier [0..n]
- provenance :InternationalString [0..n]
- rights :InternationalString [0..n]

Example:
- extent: codeValue = Lead
- role: PairedCodeValueType [0..n]
  - role: Conceptualization

Example:
- agent: BibliographicName [0..1]
  - agent: ContactName

Example:
- Individual
- Organization
- Machine
DDI4
Annotations on Almost Everything

Most objects inherit from AnnotatedIdentifiable
DDI4 Inheritance

AnnotatedIdentifiable
AnnotatedIdentifiable
Creators and Contributors Are AgentAssociations

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- title : InternationalString [0..1]</td>
</tr>
<tr>
<td>- subTitle : InternationalString [0..n]</td>
</tr>
<tr>
<td>- alternateTitle : InternationalString [0..n]</td>
</tr>
<tr>
<td>- creator : AgentAssociation [0..n]</td>
</tr>
<tr>
<td>- publisher : AgentAssociation [0..n]</td>
</tr>
<tr>
<td>- contributor : AgentAssociation [0..n]</td>
</tr>
<tr>
<td>- date : AnnotationDate [0..n]</td>
</tr>
<tr>
<td>- identifier : InternationalIdentifier [0..n]</td>
</tr>
<tr>
<td>- copyright : InternationalString [0..n]</td>
</tr>
</tbody>
</table>
A Link to an Agent Object

Agent

- agent :BibliographicName [0..1]
- role :PairedCodeValueType [0..n]

AgentAssociation

e.g. role:
codeValue=Conceptualization

Agent

Individual

Organization

Machine
AgentAssociation has Role(s) with Extent

Multiple roles

Each has an extent (degree of contribution)
The CRediT Taxonomy

http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033

Article in Nature (4/16/2014) proposed a taxonomy of contributor roles for authors
The CRediT Taxonomy

http://credit.casrai.org/proposed-taxonomy/

- conceptualization
- methodology
- software
- validation
- analysis
- investigation
- resources

- curation
- writing
- review and editing
- visualization
- supervision
- administration
- funding acquisition
Each with a Degree of Contribution

- Lead
- Equal
- Supporting
# DDI Lifecycle Controlled Vocabulary

## Code List

<table>
<thead>
<tr>
<th>Value of the Code</th>
<th>Descriptive Term of the Code</th>
<th>Definition of the Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>StudyProposal</td>
<td>Study proposal</td>
<td>Defining outlines for a new study/data collection, including needs for information and study scope and methodology, usually to be presented for approval to funders, partners</td>
</tr>
<tr>
<td>Funding</td>
<td>Funding</td>
<td>Decisions to extend financial support for the study/data collection.</td>
</tr>
<tr>
<td>StudyDesign</td>
<td>Study design</td>
<td>Detailed planning for carrying out the study/data collection: refining concepts and ider population, time dimension, sampling frame and sample selection, data collection metl</td>
</tr>
<tr>
<td>InstrumentDesign</td>
<td>Instrument design</td>
<td>Building the data collection instrument, for example, the questionnaire or interview qu observational design, standardized record review, independent and dependent variable</td>
</tr>
<tr>
<td>QuestionnaireTranslation</td>
<td>Questionnaire translation</td>
<td>Translating the source questionnaire into other languages, for example, in cross-nation multilingual countries.</td>
</tr>
<tr>
<td>QuestionnaireAdaptation</td>
<td>Questionnaire adaptation</td>
<td>Changing the wording of questions to reflect cultural or institutional differences if sam different regions or countries.</td>
</tr>
<tr>
<td>InterviewerTraining</td>
<td>Interviewer training</td>
<td>Training the interviewers that administer questionnaires in survey-type studies.</td>
</tr>
<tr>
<td>EthicsReview</td>
<td>Ethics review</td>
<td>Review of the study/data collection to ensure that it complies with statutory ethics req informed consent statement, performed by a qualified body, like a Research Ethics Co</td>
</tr>
<tr>
<td>LegalReview</td>
<td>Legal review</td>
<td>Review of the study/data collection in terms of compliance with the law (legislation cc etc.).</td>
</tr>
<tr>
<td>Sampling</td>
<td>Sampling</td>
<td>Selecting the sample for the study/data collection.</td>
</tr>
<tr>
<td>InstrumentPreTesting</td>
<td>Instrument pre-testing</td>
<td>Small-scale application of the data collection instrument designed to identify potential</td>
</tr>
<tr>
<td>PilotStudy</td>
<td>Pilot study</td>
<td>Dress rehearsal of the full project, for example, by administering the questionnaire to results, etc.</td>
</tr>
<tr>
<td>DataCollection</td>
<td>Data collection</td>
<td>Setting up, refining and finalizing the data collection process, including follow-up and</td>
</tr>
</tbody>
</table>
Comparison of Taxonomies

• Mapping was close
• DDI CV was more detailed but mapped well to major categories
• CReDiT taxonomy developed for authors
• Decided to adopt CReDiT taxonomy as a way to interoperate with others
Other Outcomes

• Recommended list of DDI4 properties to support citation
• Recommendations for a CDISC ODM-XML extension to support citation
• A proposed DDI4 “metamodel” object to support descriptive information structured by an external vocabulary
Drinking Our Own Champagne
(or Eating Our Own Dogfood?)

• Citation with roles and degree for our paper
• Citation information for a dataset created from the meeting minutes
• Instrument documentation for the text mining procedure
Creating a Dataset

Minutes in Google Docs

SAS Dataset, one row per paragraph

Text Mining Process (SAS Text Miner)

Topics Dataset
Role and Degree for Our Dataset

Contributors: Larry Hoyle (conceptualization, lead; methodology, lead; software, lead; formal analysis, lead; data curation, lead), Mary Vardigan (conceptualization, equal), Sam Hume (conceptualization, equal), Sanda Ionescu (conceptualization, equal), Jay Greenfield (conceptualization, equal), Jeremy Iverson (conceptualization, equal), John Kunze (conceptualization, equal), Barry Radler (conceptualization, equal), Wendy Thomas (conceptualization, equal), Stuart Weibel (conceptualization, equal), Michael C. Witt (conceptualization, equal)

Gets quite long, not likely to appear in citation or author line. – Where then? And how to harvest?
Harvesting Citation Information

Contributors: Larry Hoyle (conceptualization, lead; methodology, lead; software, lead; formal analysis, lead; data curation, lead), Mary Vardigan (conceptualization, equal), Sam Hume (conceptualization, equal), Sanda Ionescu (conceptualization, equal), Jay Greenfield (conceptualization, equal), Jeremy Iverson (conceptualization, equal), John Kunze (conceptualization, equal), Barry Radler (conceptualization, equal), Wendy Thomas (conceptualization, equal), Stuart Weibel (conceptualization, equal), Michael C. Witt (conceptualization, equal)

Pointing to structured information could make harvesting easier

The Annotation object will also have an additional property capable of containing administrative, characterizing, and other information structured by an external vocabulary.
Text Mining Topics Generation as an Instrument

- Text Miner is “point and click” - very much an instrument
- Each node has a set of parameter settings
- Single values
- and tables
How Do We Preserve These Metadata?

- What are the parameters and what do they mean?
- How are they structured?
- What were the values for this analysis?

<table>
<thead>
<tr>
<th>Property</th>
<th>Node_TextParsing</th>
<th>Node_TextFilter</th>
<th>Node_TextTopic</th>
<th>Node_TextCluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>delimit</td>
<td>Std</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bCapitalize</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bPartOfSpeech</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NounGroups</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multiDS</td>
<td>SASHELP.ENG_MULTI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bPatterns</td>
<td>NONE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stopList</td>
<td>SASHELP.ENGSTOP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ignorePOS</td>
<td>'AUX' 'CONJ' 'DET' 'INTERJ' 'PART' 'PREP' 'PRON'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ignoreAttrib</td>
<td>'NUM' 'PUNCT'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bStems</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>synonymDS</td>
<td>SASHELP.ENGSYNMS</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A “Metamodel” Object

- When structure is not well known or agreed upon
- A DDI object which takes structure from an external vocabulary
- Encourages sharing of structure
- Allows validation against the vocabulary
Resources

• Project archive:  
  http://kuscholarworks.ku.edu/handle/1808/15746