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)

Funded by NSF Award # 1448107

NSF Solicitation

NSF 14-059

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

DDI3.2 Citation

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):

contributor, coverage, creator, date, dc:abstract, dc:accessRights, dc:alternative, dc:audience, dc:available, dc:bibliographicCitation, dc:conformsTo, dc:created, dc:dateAccepted, dc:dateCopyrighted, dc:dateSubmitted, dc:educationLevel, dc:extent, dc:hasFormat, dc:hasPart, dc:hasVersion, dc:isFormatOf, dc:isPartOf, dc:isReferencedBy, dc:isReplacedBy, dc:isRequiredBy, dc:isVersionOf, dc:issued, dc:mediator, dc:medium, dc:modified, dc:references, dc:replaces, dc:requires, dc:spatial, dc:tableOfContents, dc:temporal, dc:valid, description, format, identifier, language, publisher, relation, rights, source, subject, title, type

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

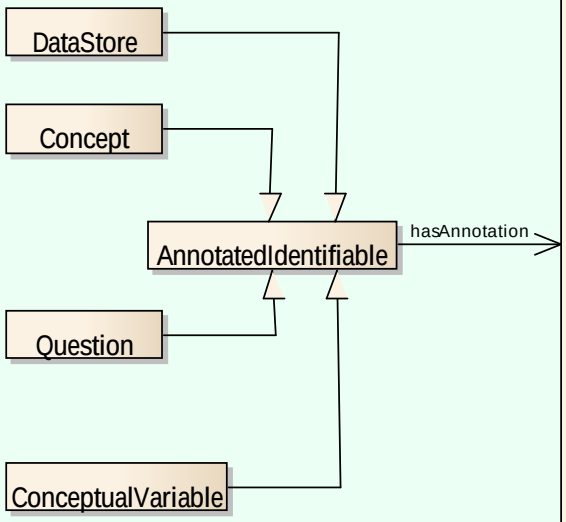
Dataset type properties support traditional citation content and help to facilitate data reuse:

- Who
- What
- When
- Where
- Whether
- How
- Structure
- Provenance

DDI4

Annotations on Almost Everything

Most objects inherit from AnnotatedIdentifiable. Examples:

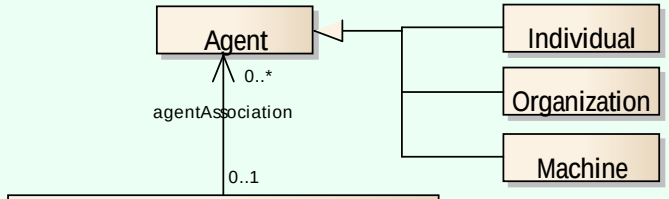


hasAnnotation

Annotation

- 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]

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



AgentAssociation

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

e.g. role:
codeValue=Conceptualization

PairedCodeValueType

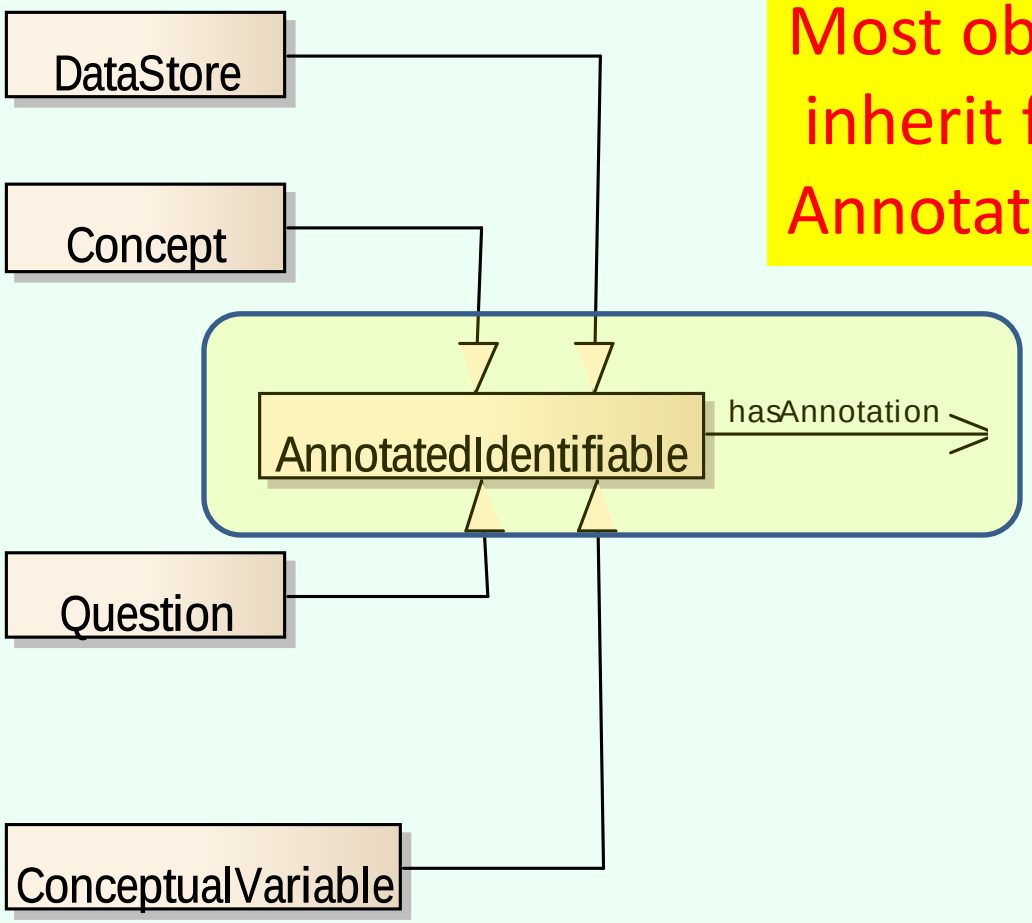
- extent :CodeValueType [0..1]

CodeValueType

- codeValue :xs:string [0..1]
- codeListID :xs:string [0..1]
- codeListName :xs:string [0..1]
- codeListAgencyName :xs:string [0..1]
- codeListVersionID :xs:string [0..1]
- otherValue :xs:string [0..1]
- codeListURN :xs:string [0..1]
- codeListSchemeURN :xs:string [0..1]

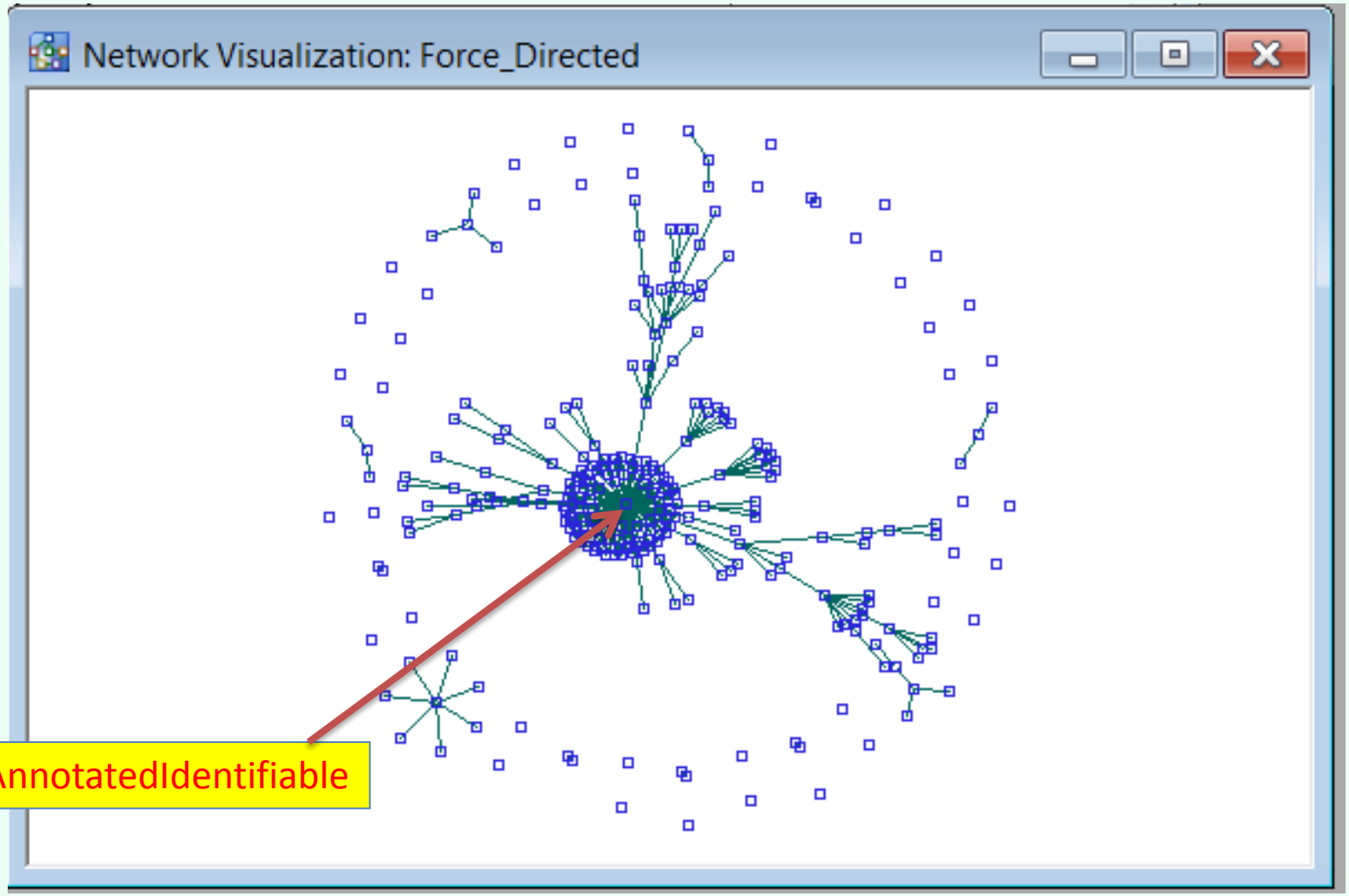
e.g.
extent:
codeValue = Lead

DDI4
Annotations on Almost Everything

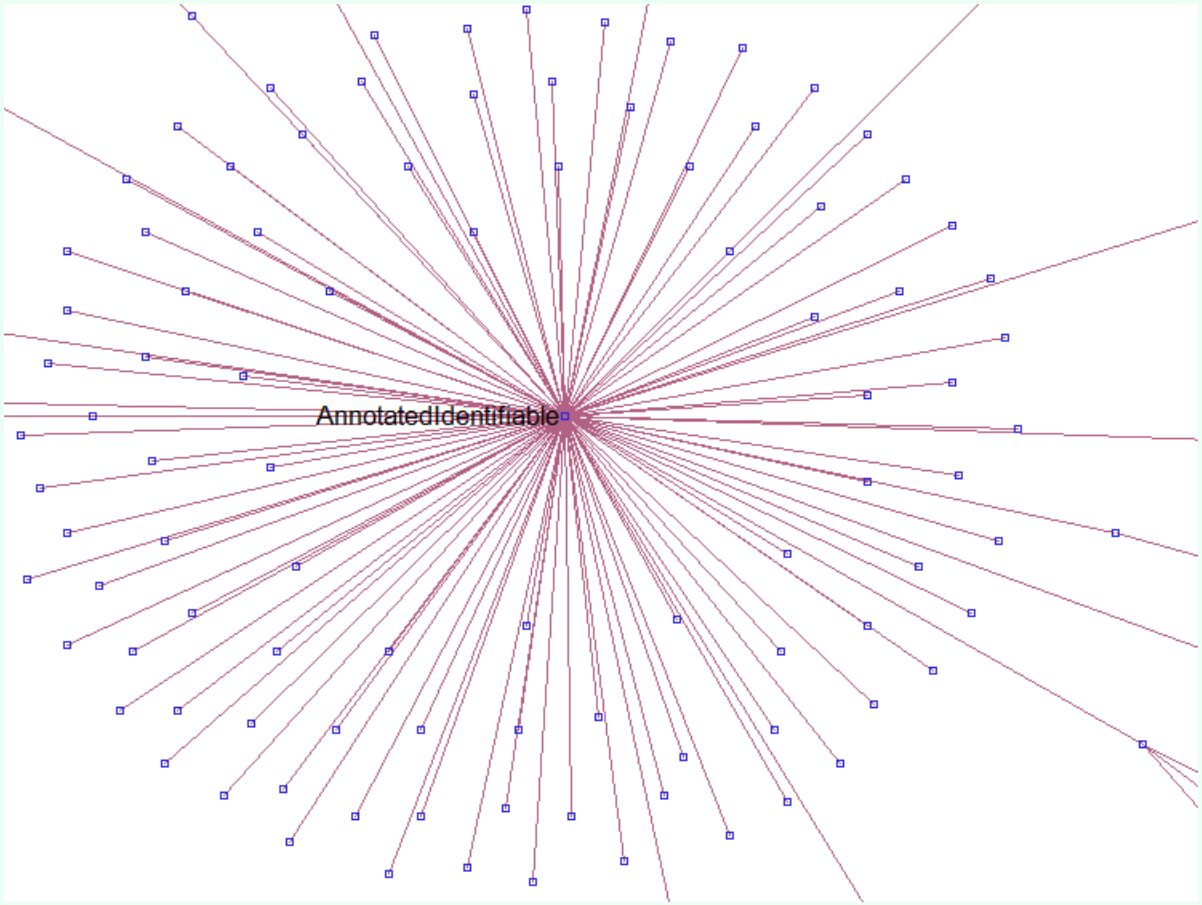


Most objects inherit from AnnotatedIdentifiable

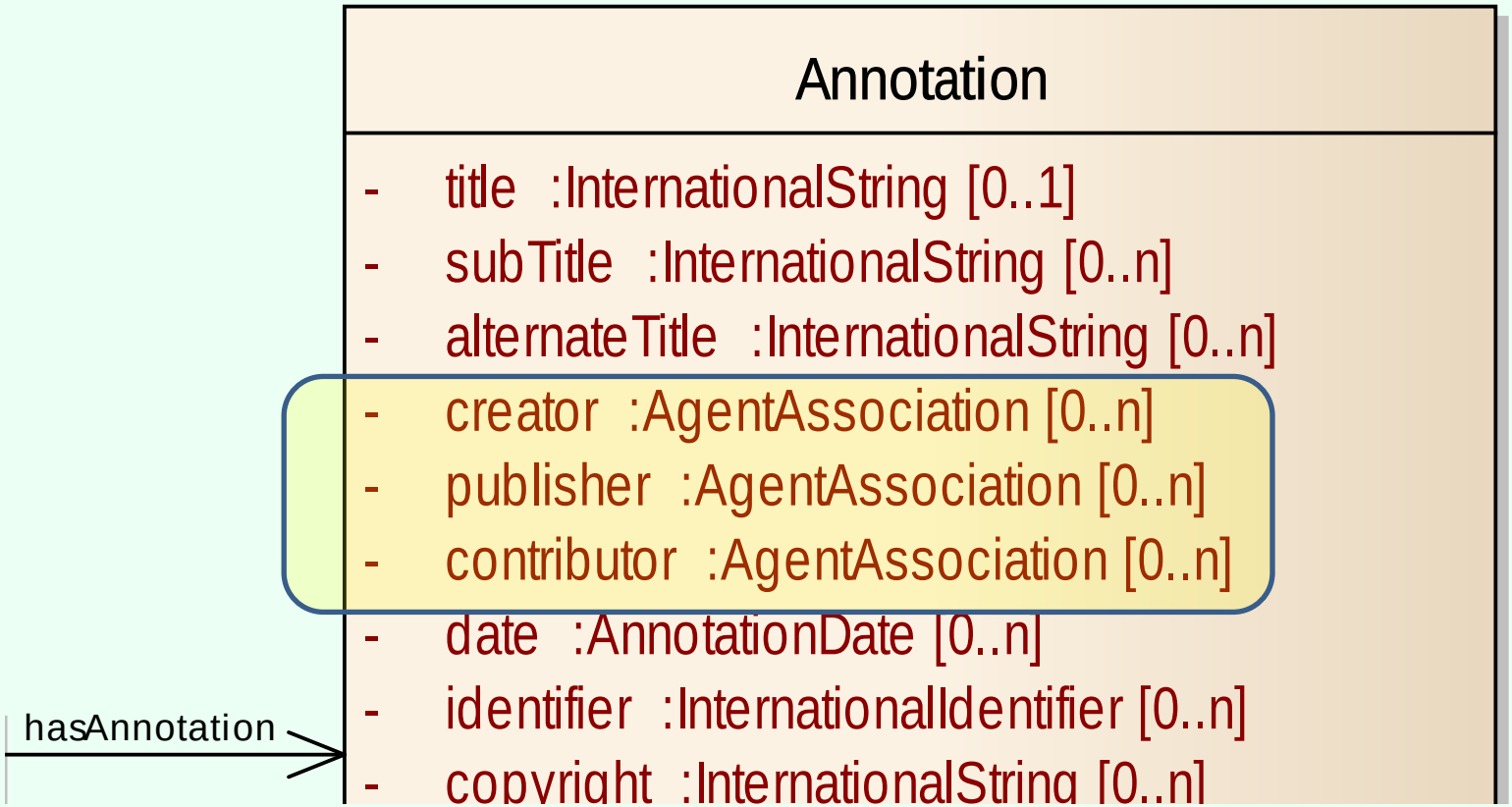
DDI4 Inheritance



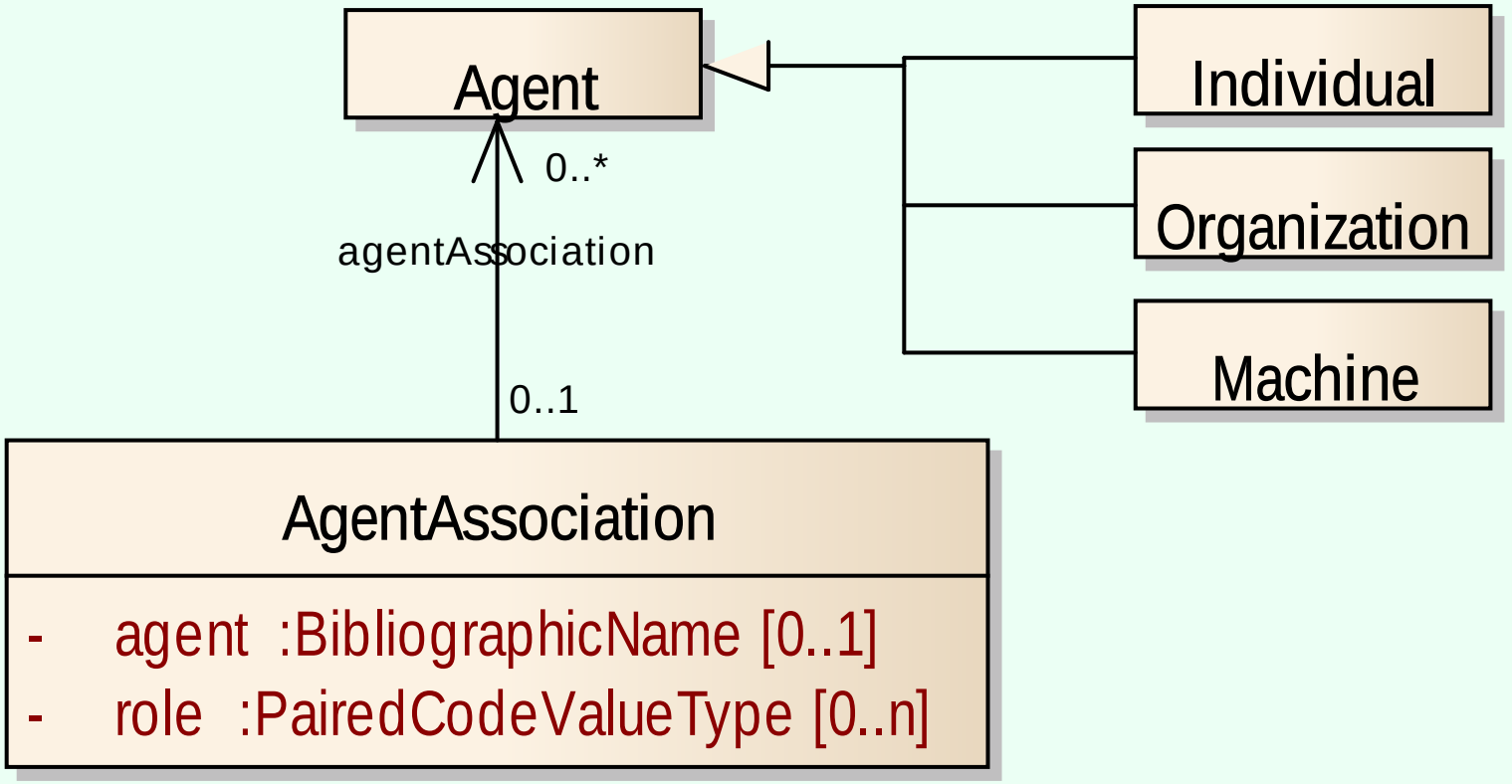
AnnotatedIdentifiable



Creators and Contributors Are AgentAssociations

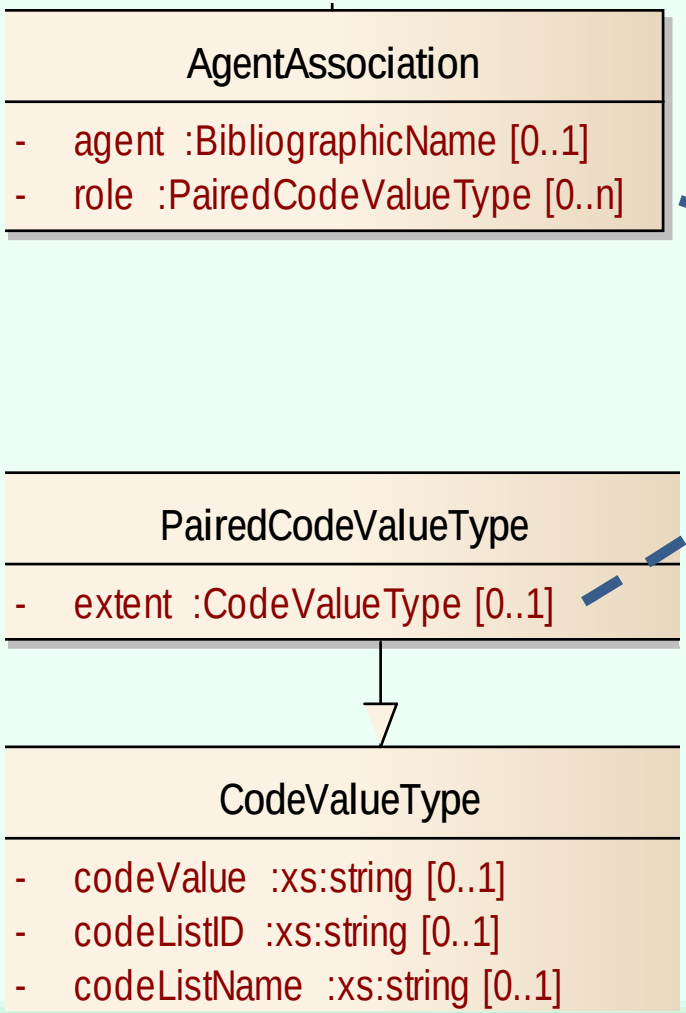


A Link to an Agent Object



e.g. role:
codeValue=Conceptualization

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

COMMENT

Writing

Investigation

Study conception

Formal analysis

Credit where credit is due

Liz Allen, Amy Brand, Jo Scott, Micah Altman and Marjorie Hlava are trialling digital taxonomies to help researchers to identify their contributions to collaborative projects.

Research today is rarely a one-person job. Original research papers with a single author are — particularly in the life sciences — a vanishing breed. Partly, the inflation of author numbers on papers has been driven by national research-assessment exercises. Partly, it is the emergence of big and collaborative science, assisted by technology, that is changing the research landscape.

What we cannot tell easily by reading a

Through the endorsement of individuals' contributions, researchers can start to move beyond 'authorship' as the dominant measure of esteem. For funding agencies, better information about the contributions of grant applicants would aid the decision-making process. Greater precision could also enable automated analysis of the role and potential outputs of those being funded, especially if those contributions were linked to an open

journal articles could be classified using a 14-role taxonomy (see 'Who did what?'). The survey was sent to 1,200 corresponding authors of work published in PLOS journals, Nature Publishing Group journals, Elsevier journals, *Science* and *eLife*. Corresponding authors were asked to indicate the contribution of each author of their article according to the roles in the taxonomy, and to comment on its comprehensiveness; whether there

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

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

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

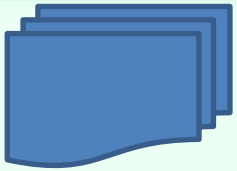
Other Outcomes

- 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

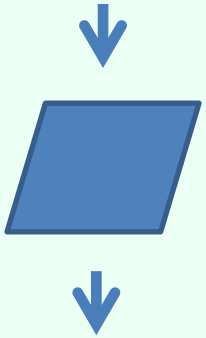
Drinking Our Own Champagne (or Eating Our Own Dogfood?)

- 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

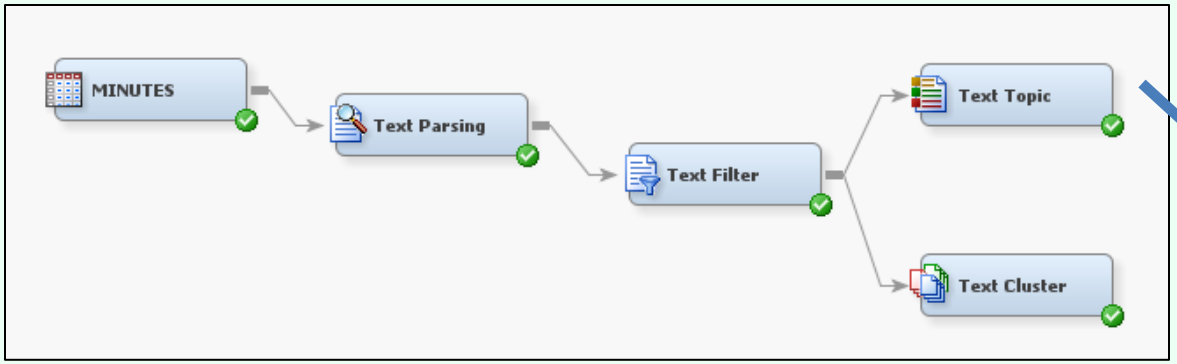
Creating a Dataset



Minutes in Google Docs



SAS Dataset, one row per paragraph



Text Mining Process
(SAS Text Miner)

Topics Dataset

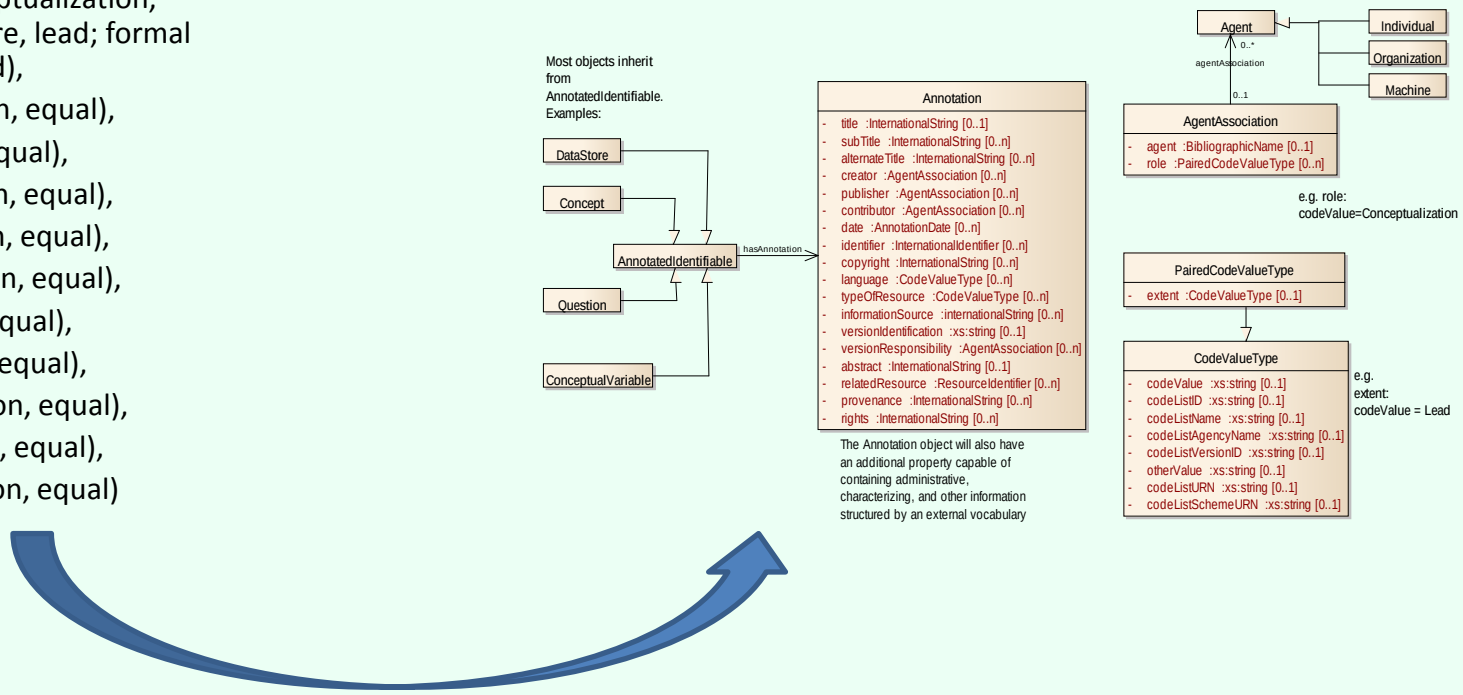
Role and Degree for Our Dataset

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

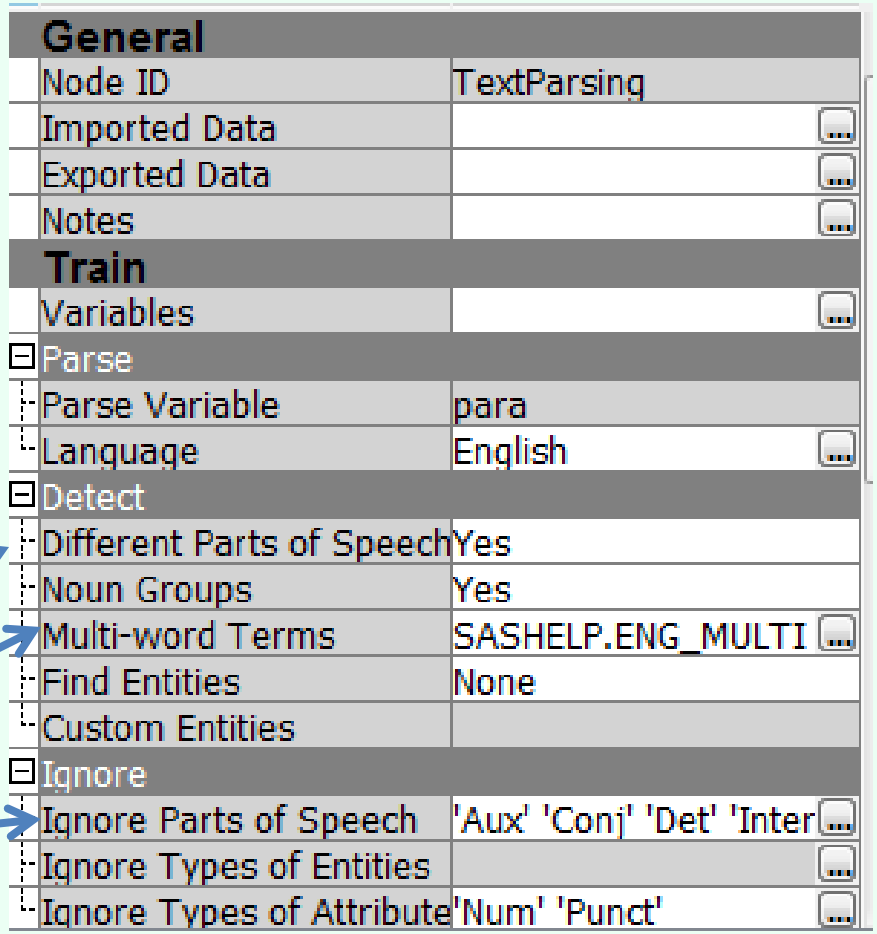
Contributors: Larry Hoyle (conceptualization, lead; methodology, lead; software, lead; formal analysis, lead; data curation, lead),
Mary Vardigan (conceptualization, equal),
Sam Hume (conceptualization, equal),
Sanda Ionescu (conceptualization, equal),
Jay Greenfield (conceptualization, equal),
Jeremy 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

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



General	
Node ID	TextParsing
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Parse	
Parse Variable	para
Language	English
Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inter
Ignore Types of Entities	
Ignore Types of Attribute	'Num' 'Punct'

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?

Property	Node_ TextParsing	Node_ TextFilter	Node_ TextTopic	Node_ TextCluster
delimit	Std			
bCapitalize	Y			
bPartOfSpeech	Y			
NounGroups	Y			
multiDS	SASHELP.ENG_MULTI			
bPatterns	NONE			
stopList	SASHELP.ENGSTOP			
ignorePOS	'AUX' 'CONJ' 'DET' 'INTERJ' 'PART' 'PREP' 'PRON'			
ignoreAttrib	'NUM' 'PUNCT'			
bStems	Y			
synonymDS	SASHELP.ENGSYNMS			

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