DDI and Enhanced Data Citation

by Larry Hoyle, Mary Vardigan, Jay Greenfield, Sam Hume, Sanda Ionescu, Jeremy Iverson, John Kunze, Barry Radler, Wendy Thomas, Stuart Weibel, Michael Witt[1]

Abstract

In October 2014 at the fifth DDI Moving Forward Sprint a subgroup met[2] to focus on adding structure to DDI4 to support enhanced citation of data. A principal question was how to record the role(s) and degree of contribution of those contributing to the creation and curation of data. We also considered the question of which information objects associated with data creation might need enhanced citation information. We chose to think broadly about this, moving beyond the notion of citing a dataset to explore other types of intellectual objects that might merit some form of citation or annotation and reuse – for example, a new data collection method or a constructed variable. In thinking about roles we reviewed the CRediT taxonomy (Allen et al. 2014) and decided that it would serve as a good foundation in DDI4 for an extensible vocabulary for roles. Further, we determined that all DDI4 versionable objects should allow for the attachment of an annotation supporting citation along with role and degree of contribution. As a result of the Dagstuhl meeting the initial releases of DDI4 will have an annotation object allowing for the attribution of roles and associated degree of contribution for creators and contributors to the creation of versionable objects. Attribution information has also been proposed as a CDISC ODM-XML extension planned for development in 2015.

Keywords

Attribution, contributor role, data citation, DDI, Dublin Core, CDISC, enhanced citation, metadata

About the Data Documentation Initiative (DDI)

The DDI initiative was established by the Inter-university Consortium for Political and Social Research (ICPSR) in 1995 with support from NSF and in 2003 transitioned to become a self-supporting membership Alliance with over 40 current members who contribute to shaping the standard. DDI has two major development lines: DDI Codebook, intended to document simple quantitative survey data, and DDI Lifecycle, which is broader in scope, covering the research data life cycle from conceptualization to collection and processing to data publication and beyond.

DDI's primary goal is to document research datasets and processes thoroughly so that data are independently understandable. Advantages of the DDI approach are that metadata are machine-actionable and reusable. With origins in the quantitative survey-based social sciences, DDI can be used by researchers in other disciplines and can describe other types of data, such as experimental, observational, biological, administrative, and transaction data. Originally expressed in XML schemas, DDI is now evolving as a model-based specification that will enable a variety of renderings.

Introduction

It is common to cite traditional scholarly literature such as conference papers and journal articles, and the mechanism for doing this is widely accepted and practiced. Citing research data, which also represent significant intellectual effort, is becoming more common, but best practices and norms for citing data are not yet widely accepted (Borgman 2012). A related

About the Clinical Data Interchange Standards Consortium (CDISC)

CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission, and archiving of clinical research data and metadata. CDISC is member-supported by approximately 350 biopharma, academic, and service provider organizations. CDISC's mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform-independent and freely available via the CDISC website.

CDISC standards cover the full clinical research lifecycle from protocol through analysis and reporting, including regulatory submissions. The CDISC data exchange standards provide support for data and metadata exchange and archiving of the foundational and therapeutic area content standards.

issue is that to make data usable, there is a need to provide more than simple attribution and location information.

The practice of citing other constructs related to data objects such as instruments, questions, and variables is rare and, in general, lacks the consensus needed for it to emerge as a practice. Yet citing at this more granular level is increasingly viewed as important both in terms of provenance chains and awarding credit (CDC 2013).

Contributorship is another important part of the picture. Processes and procedures for attribution around data are just getting established, and it is clear that the life cycle of research data presents new possibilities for how we view contributions to the creation of research data. The development of a dataset has many stages and can involve a multitude of actors who make significant contributions to the final product but have traditionally gone unacknowledged. The idea of extending credit beyond the principal investigators and authors to others who have played critical roles is also being explored in research done by the Harvard/Wellcome Trust (Allen et al. 2014). This synergy offers an opportunity to think about contributorship with respect to research data in new ways.

The Data Documentation Initiative (DDI) is an open structured metadata standard for documenting and managing research data, and as such, it needs to address these key issues of data citation and contributorship. The standard needs to include all of the metadata elements necessary to cite and describe a data object and to support that citation. Ideally all of these citation-related metadata are machine-actionable and can facilitate additional data discovery.

To explore these related issues, a group of experts on data citation funded by the National Science Foundation (#1448107) met in October 2014 at Schloss Dagstuhl and worked alongside the DDI4: Moving Forward sprint. Representatives from the Dublin Core Metadata Initiative and CDISC, the Clinical Data Interchange Standards Consortium, were also in attendance. The group sought to answer some key questions:

- What objects documented by DDI should be citable?
- What elements are needed in DDI and CDISC to cite data and describe data sources in a comprehensive way across the lifecycle?
- Given the lifecycle focus of DDI, how can we support broad attribution for contributions, and how can we describe the level of contribution?

This paper reports on the group's consensus around these questions and sets out a list of recommendations to improve citation coverage in DDI. To test our decisions and assumptions, we created a sample dataset and went through the exercise of citing the dataset and related information objects using enhanced citation information.

The need to attach other kinds of annotations having a function similar to citation to objects was another focus of the meeting. This might include administrative information such as the OMB-required information about the provenance of survey questions or characterization information such as parameter settings for an instrument. The structure for such information is not generally well known enough to be formalized in DDI, as it is typically defined and revised by some community of interest. This highlighted the need for DDI4 to include a structured information object capable of being validated from some external vocabulary. Plans are underway for the development of a DDI4 object to be structured by an external vocabulary.

Current Status of Data Citation

The changing nature of scholarly and research communication is not new; for example, a notable effort reporting on this subject met at Dagstuhl in 2011 [Bourne et al. 2012]. The rapidly evolving research landscape requires us to revisit some of the traditional paradigms that have characterized citation and attribution, particularly when it comes to publication and citation of non-traditional research products, such as research data.

The overall purpose of data citation has been articulated in two similar sets of data citation principles [Force 11, CODATA-ICSTI Task Group]. They both state that data should be considered legitimate, citable products of research and be accorded the same importance in the scholarly record as citations of research publications. Noting that no single style or mechanism may apply to all data or all disciplines, the groups advocating for data citation also mention the following general requirements:

- Credit: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data.
- Evidence: Whenever and wherever a claim relies upon data, the corresponding data should be cited.
- Unique Identification: A data citation should include a persistent method for identification that is machine-actionable and globally unique.
- Specificity, Verifiability, and Utility: A data citation should lead to the specific data subset used (timeslice, version, etc.), to sufficient context to verify that the data used were the same as the data retrieved (fixity, provenance, etc.), and to code, documentation, and methods adequate for making informed use of the data.
- Interoperability and Flexibility: Data citation methods should be sufficiently flexible to accommodate the variant practices

> among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

An international consortium called DataCite[3] was created in 2010 to establish citation of data and other non-traditional research products as normal, mainstream research activities and to provide an infrastructure for the registration of Digital Object Identifiers (DOIs) for data. DataCite has published a list of minimal elements that should be part of a data citation (while acknowledging that data publication practices can vary across disciplines). These include Creator, PublicationYear, Title, Publisher, and Identifier with optional properties of Version and ResourceType.

While these citation elements may be formatted in different ways, DataCite recommends the format: Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier. (The major style guides each have specified formats for data citations.)

ORCID (Open Researcher and Contributor ID) is a complementary effort designed to disambiguate contributors' names by assigning globally unique researcher identifiers to link researchers to records of their scholarly output that are accurate and complete.

Citable Objects in DDI

In our discussions of citations and annotations more broadly, the Enhanced Citation Working Group at the Dagstuhl DDI Sprint (October, 2014) identified several generalized use cases for annotations as well as the need to distinguish among varieties and purposes of such annotations. For example, we might think about attribution (manifested through citations), administration, and characterization as forms of annotation.

We use the term *description* to signify structured annotations (as opposed to unstructured annotations such as notes). A *description-type* is a specific set of metadata elements intended to support an identified functional requirement. Just as declared *data-types* rationalize the management of variables in a programming language, declared *description-types* will help rationalize the management of classes of metadata sets within a data management application. We suggest that this notion of description-types be explored as the DDI Alliance models annotations.

Generalized use cases for description-types

We identify four general use cases that we believe justify different description-types. It is expected that additional use cases will emerge.

Citation:

A form of attribution, this *description-type* is the familiar bibliographic notion of establishing the relationship of one or more individuals with a manifestation of an intellectual product, disambiguating that intellectual product from others, and, where practicable, facilitating access.

Citation answers (at a minimum) the following questions:

- Who is credited with creating the product?
- What is the product named?
- When was the product created?
- Where can the product be accessed?
- Whether: Are there constraints on access to the product?

The intellectual product referenced by a citation can take many forms, including books, articles, and reports. The creation of other intellectual products may also be credited. Examples could include data, an algorithm, an instrument, and more. A citation recognizes the creation of the intellectual product and may also serve to help locate information about the product.

There are needs, though, for description types that go beyond the simple attribution and location use case (see below).

Sourcing:

Within the social science and biomedical research communities, other object types need to be referenced. For example, in the US, questionnaire questions administered to more than 10 persons by a federal agency must be assessed by the Office of Management Budget (OMB) for the degree of burden imposed on respondents. OMB requires that each such question be 'sourced' to facilitate review. The structured data for such sourcing will look very much like a citation, but should be typed differently to facilitate discovery and administration. The functional requirement is not intellectual attribution, but rather administrative responsibility:

- Where does a question come from?
- Is it an accepted and tested component of an existing instrument?
- Does it require further analysis or vetting?

Instrument Description:

Data collected across a large sample may rely on multiple physical instruments manufactured by different manufacturers. The set of all such instruments may be thought of as a conceptual instrument, but differences among instruments from different manufacturers may yield systematic differences in data that can be normalized after the fact based upon known operational differences among instruments. One can imagine a rich variety of such problems that require instrument-description metadata to identify the manufacturer, operational provenance, operational characteristics, and more. The characterization metadata in this case may also serve to document the use of the instrument to create data rather than just its creation. An infrared thermometer, for example might have a switch allowing for either fast response with less accuracy, or slow response with more accuracy. A second thermometer might also have a switch for Fahrenheit vs. Celsius. Knowing the switch settings used to collect a set of data could be important. A questionnaire, another type of instrument, might be administered on paper, or on a computer.

Instruments can also be seen as intellectual products and thus can be cited to attribute them to specific creators.

Dataset Description:

The use case for a structured dataset description is broad-based and fundamental to the work of DDI and other data initiatives. Promoting appropriate conventions for such descriptions is a core responsibility of DDI.

We define a dataset as:

> *a discrete collection of measurements collected via observation, experiments, or analysis, using specified methodologies and instruments, and structured in a manner documented by formal schemas.*

The unbounded diversity of datasets mitigates against any single means of characterizing or cataloging them. However, communities of practice exist and can be encouraged to coalesce around common conventions for structuring their data and dataset descriptions so as to promote discovery, reuse, and preservation.

See, for example, the Data Discovery Index that NIH proposes as part of its larger BD2K initiative. Such an index will provide pointers (actionable links) to dataset metadata that reside (and are managed) elsewhere. These pointers would form a 'data catalog'. From the BD2K Data Discovery Index Workshop Summary Report (emphasis added):

> 'A catalog could in some cases be a human-viewable database analogous to a traditional paper catalog and in other cases, could be a set of functions to serve both human users and, increasingly, machine interfaces (i.e., 'computers talking to computers') to support the needs of scientific data discovery, exchange, and analysis […] Unlike the printed catalog of the past, it can be predicted that a new resource that enables locating, characterizing, and accessing NIH-funded data will have to evolve in an agile way to serve both data producers and data consumers to keep pace with the ever-changing, networked world. Technical approaches to describing, finding and providing access to the broad variety of 'data objects' that are the output of contemporary biomedical science will likely continue to evolve and improve.'

The abstract requirements of a dataset description might include the following:

- Who (institutional responsibility, authorship, funding sources)
- What (Title[s] and project description)
- When (date of publication)
- Where (access points)
- Whether (management of access: who can use, edit, reference a dataset)
- How (how was data collected, and what additional information may be necessary for its interpretation)
- Structure specifications (formal machine-interpretable schemas necessary for parsing the dataset)
- Provenance (reuse and modification history)

The characterization information in this dataset description supports the traditional citation content and helps to facilitate data reuse.

In CDISC, the Define-XML standard (http://cdisc.org/define-xml) provides the metadata to describe CDISC datasets such as SDTM or ADaM, and is required when these datasets are part of an FDA submission. Define-XML v2.0 meets most of the stated requirements with the exception of *whether*.

We emphasize that these requirements are speculative and will evolve in the context of communities of practice.

Rationale for identifying and promoting typed references

These generalized description-types (citation, sourcing, instrument-description, dataset-description) share a requirement for structured metadata, some elements of which will be common to several description-types. Other description-types will require elements that may be specific to the particular description-type, or even to the specifics of a given instrument or experiment (e.g., sensor type or calibration history or a schema specifying the structure of a dataset). Additional description-types will likely emerge as well.

DDI should support the import of metadata elements from established metadata dictionaries such that description-types reflect the needs of existing projects and established communities.

It should be noted that there is inherent tension between the objectives of (a) incorporating metadata practices established in existing communities and (b) encouraging the reuse of metadata elements by promoting standardization across communities. This is a socialization function that DDI can encourage, but cannot expect to achieve entirely. The metadata world is a messy place.

As description-types evolve, the DDI community has a role in encouraging the reuse of elements from one description-type in another while avoiding the overloading of semantics that might create ambiguities. For example, a manufacturer can be mapped to a creator and date of manufacture can be mapped to *publication date*, but structured descriptions appropriate for an instrument-description will quickly diverge from the *who-what-when-where-whether* of citation metadata. Care must be taken to avoid conflation of semantics that will cause ambiguity or confusion.

One of the most important benefits of description-typing will be to help create shared understanding among users. Complicated systems such as DDI require shared understandings of functional requirements, component definitions, and relationships among the parts. Designers, modelers, software developers, practitioners (system managers and creators of metadata), and end-users achieve shared understandings through natural language. To overload widely understood concepts (such as citation) with other description-types such as an *instrument-description* is to risk obfuscation of both description categories and violate shared user-models.

Finally, a search-view of DDI will benefit from distinguishing among the functional requirements implied in description-typing. The conventional notion of citation metadata leads users to expect to find such data in a coherent collection of item records of the *who-what-when-where-whether* form. Those searching for *instrument-description* metadata or *dataset descriptions* will expect to find it in collections of records organized to reflect respective functional requirements.

To summarize, we propose a description-typing approach that has the following characteristics:

- Metadata element sets exist in communities of practice, and should be welcomed into DDI, even when not formally sanctioned or managed by DDI.
- DDI should promote, but cannot enforce, the standardization of metadata practices that support cross-community discovery.
- Description-types are discrete sets of structured annotations that serve specific functional requirements and are characterized by carefully selected metadata elements that meet these functional requirements and promote coherent discovery. We identify four such types here, but expect others to emerge.

W5 HSP	Proposed DDI Property	Dublin Core Mapping	DDI3.2 Mapping
What	Label (type = Title)	Title	Label and/or Citation/Title
Who	Creator Role DegreeOfContribution	Creator	Citation/Creator
When	PublicationDate	Date	Citation/PublicationDate
What	UserID	Identifier	URN and UserID and/or Citation/ InternationalIdentifier and/or Citation/ dc:identifier
Where	Publisher	Publisher	Citation/Publisher and/or Citation/ dc:publisher
Who	Contributor Role DegreeOfContribution	Contributor	Citation/Contributor (with role) and/or Citation/dc:contributor
What	Language	Language	Citation/Language and/or Citation/ dc:language
Whether	Copyright	Rights	Citation/Copyright and/or Citation/dc:rights
Whether	License	Rights? None?	Archive/Item/Access/ AccessConditions, dc:AccessRights? *Note: dc:license to be added in DDI3.3.*
What	Description	Description	Description and/or Citation/dc:description, Abstract?
HSP	UserAttribute	dc:any, or None	*Citation/dc:any* or None, UserAttributePair
What	Label (type = SubTitle)	SubTitle	Citation/SubTitle and/or Citation/dc:subTitle
What	Label (type = AlternateTitle)	AlternateTitle	Citation/AlternateTitle and/or Citation/ dc:alternative
When/P	DateCreated	Created	Citation/dc:created
When/P	DateModified	Modified	Citation/dc:modified
What/P	Version	isVersionOf ?	PhysicalInstance/pi:DataFileVersion or r:Version of the containing element
What	Resource Type		KindOfData
W5 HSP	Pointer to metadata		PhysicalInstance/ r:DataRelationshipReference and r:RecordLayoutReference
Where	Actionable link to the dataset		DataFileIdentification/ DataFileURI and/or r:Location

***Table 1**. DDI Properties Supporting Data Citation (W5 plus **H**ow, **S**tructure, and **P**rovenance)*

We believe this approach can account for existing citation metadata approaches and provide a platform for developing emerging standards for referencing data objects. It also supports special purpose annotations (such as *sourcing and instrument-descriptions*) and affords a flexible foundation for the evolution of description-types that are currently unforeseen.

Unresolved Issues:

- Who has responsibility for introducing, defining, registering, and managing description-types?
- How is a description-type declared in data instances?
- Do all description-types have a set of obligatory and optional metadata elements (as has been proposed for citations)?
- Are there constraints on the sources of metadata elements?
- Are there means by which DDI can help to 'socialize' metadata best practices within its domain?
- How can description-types defined by a user community be validated? How can relationships among the metadata elements in the description-type be described?

Elements for Citing and Describing Data

DDI elements

DDI Lifecycle Version 3.2 currently allows structured annotation metadata to be provided for several versionable object types, including StudyUnit and PhysicalInstance. However, we recommend that these fields be available on all versionable object types because, as noted previously, there is a growing need to recognize effort for non-traditional objects. Moreover, having all versionable object types contain the same basic annotation properties increases consistency in the DDI standard and can also reduce redundancies in the current DDI Lifecycle model, where, for example, the same identifying information can be recorded in multiple places.

Beyond the minimal set of metadata comprising a citation, there are other elements often used to administer, characterize, and validate data objects. For example, an author may want to provide copyright and license information for a question response scale.

Of course, the determination of when or what objects to cite is determined by researchers electing to re-use existing data, and is not a function of the standards. However, DDI needs to make it possible to cite and describe objects comprehensively.

We propose the following properties be added to each versionable object. The properties are categorized as who, what, when, where, whether, how, structure, and provenance (W5 HSP) as suggested by ISO 19773.

We also suggest that some additional properties and elements be further explored for possible inclusion. Examples include a permanence or stability indicator (e.g., as in the National Library of Medicine vocabulary) and Data Fingerprint.

When adding structured annotation metadata, the attributes should be added at the highest applicable level in the hierarchy, and this information will apply to lower levels in the hierarchy that do not explicitly include such information. When an object does not have certain information, relationships can be followed to other objects in order to discover that information. For example, if a variable does not specify a creator, a user or system could look for Creator information in the data file that contains the variable. If the data file metadata does not specify a creator, a user or system could look at the study description that contains the data file. This type of relationship traversal can be performed in a machine-actionable manner. For example, a Web application could show creator information on a variable's page, even if the information was pulled from the study description.

CDISC elements

CDISC proposes to support data citations for key CDISC ODM-XML and Define-XML data and metadata elements (see Table 2). The CDISC data exchange standards have not implemented the Dublin Core Metadata Element Set, but do include several attributes that correspond to terms from Dublin Core including the Operational Data Model (ODM-XML) element attributes CreationDateTime as Date, Originator as Creator, Description as Title, and FileOID as Identifier. The missing attributes will be added as an ODM-XML extension, based on the Dublin Core metadata terms where appropriate.

To ensure that citation metadata does not demand the addition of redundant metadata that could cause integrity issues, the core attributes are populated using existing ODM-XML attributes where available. In cases where ODM-XML does not support the required attributes, an extension containing the new elements and attributes has been proposed. Table 3 shows examples of ODM-XML elements and attributes that map to the proposed CDISC data citation properties.

Contributors and Contribution

The practice of attributing credit through the citation of datasets and other non-traditional scholarly objects is becoming more common but lacks the maturity of the traditional scholarly literature citation paradigm. One issue is that the number and

Standard	Element
ODM-XML	ODM, Study, Protocol, StudyEventDef, FormDef, ItemGroupDef, ItemDef, CodeList, CodeListItem, EnumeratedItem, MethodDef, ConditionDef, User, ClinicalData, SubjectData, StudyEventData, FormData, ItemGroupData, AuditRecord
Define-XML	ValueList, CommentDef

Table 2. *CDISC Elements to be Used to Support Data Citation*

W5 HSP	Proposed CDISC Properties	Element / Attribute
Who	Creator	/ODM/@Originator
When	Date	/ODM/@CreationDateTime
What	Title	/ODM/@Description
What	Identifier	/ODM/@FileOID
Where	Publisher	/ODM/@dc:Publisher
Who	Contributor	/ODM/Study/MetaDataVersion/dc:ContributorDef
What	Language	/ODM/Study/@xml:lang
Whether	Rights	/ODM/Study/MetaDataVersion/dc:RightsDef
Who	Creator	/ODM/Study/MetaDataVersion/ItemGroupDef/@dc:Originator
What	Title	/ODM/Study/MetaDataVersion/ItemGroupDef/@Name
What	Identifier	/ODM/Study/MetaDataVersion/ItemGroupDef/@OID
Where	Publisher	/ODM/Study/MetaDataVersion/ItemGroupDef/@dc:Publisher
Whether	Rights	/ODM/Study/MetaDataVersion/ItemGroupDef/@dc:RightsOID

Table 3. *Examples of ODM-XML Data Citation Information*

nuance of contributions to datasets are much greater than can be adequately expressed when reduced to an ordered list of authors, such as for a scholarly paper. Pressure to give credit and attribution is expected to increase as federal research funders require data sharing and management plans with grant proposals and intend to track their outputs. The scholarship of data does not fit neatly within the model of traditional scholarly publication and requires recognition of new contributor roles and contributions.

DDI Lifecycle defines different stages of the research process from concept to data collection, processing, archiving, distribution, discovery, analysis, and repurposing of data[9] The DDI Controlled Vocabulary Group (DDI-CVG) has been developing a controlled vocabulary, the DDI Controlled Vocabulary for Lifecycle Events[10], which names and defines a set of actions that constitute recognizable contributions to the entire data life cycle from project inception to data use and reuse (research process). The DDI Lifecycle Events Controlled Vocabulary primarily focuses on contributions related to research data in the context of the social and behavioral sciences. In addition, the CVG drafted the DDI Controlled Vocabulary for Contributor Roles, a controlled vocabulary of agents who perform specific actions that make up contributions[11]. To give an example using both vocabularies, the agent may be a Data Collector, whereas the DDI Lifecycle Event in which the action takes place may be Data Collection.

Similar activities have taken place outside of DDI. Haeussler and Sauermann (2014) analyzed patterns of contribution using the five-level categorization of contribution type requested by PLOS ONE (*conceived, performed, analyzed, materials, wrote*).

Allen et al. conducted a workshop and initiated a series of studies to begin to define and standardize a taxonomy (see Appendix A) to help researchers identify their contributions to collaborative projects, primarily in the context of the preparation and publication of scholarly papers. Categories of contribution were classified and defined by giving high-level examples of contribution actions. This taxonomy was evaluated by authors of scholarly papers and generally accepted with 85 percent of them saying that it was easy to use and covered all the roles of contributors to their papers. Eighty-two percent of respondents reported that the taxonomy was at least the same or better in terms of accuracy than how contributions to their paper had actually been recorded. A follow-up study asked authors to reconstitute the submission of their original papers using these contributor roles, and this experiment was deemed successful. Feedback indicated that a weight for contribution was missing, so a simple scale was created—lead, equal, and supporting—to augment the taxonomy, which was further revised and named the Contributor Roles Taxonomy (CRediT). The project's leaders are currently pursuing formal standardization through Consortia Advancing Standards in Research Administration Information (CASRAI) in conjunction with the National Information Standards Organization (NISO).

If DDI were to adopt the CRediT taxonomy, it could increase and improve associations between objects in DDI and those outside of DDI that also share these terms. To investigate this possibility, similarities and differences between the DDI Lifecycle Events and CRediT were explored through a 'stub' mapping of the vocabularies to each other. Early results suggested that DDI may benefit by adding some concepts to its vocabulary from CRediT such as Software, Formal Analysis, Resources, and Funding Acquisition. Many if not most terms from DDI Lifecycle Events could be seen to fit into CRediT, but some gaps or mismatches were evident and warrant more thorough analysis. Also the grounding of the CRediT taxonomy is a scholarly paper, and while it is not exclusive of data, there are potential gaps and alignment issues if the primary scholarly work is a dataset and is not assumed to be a paper. It is a good time for exploring these options because DDI is in the process of designing a new version (DDI4), and the DDI Controlled Vocabulary for Contributor Roles has been approved by the CVG but has not yet been published and could be extended.

In terms of a mechanism to record role of contributor and weight of contribution in DDI, the following recommendations were made:

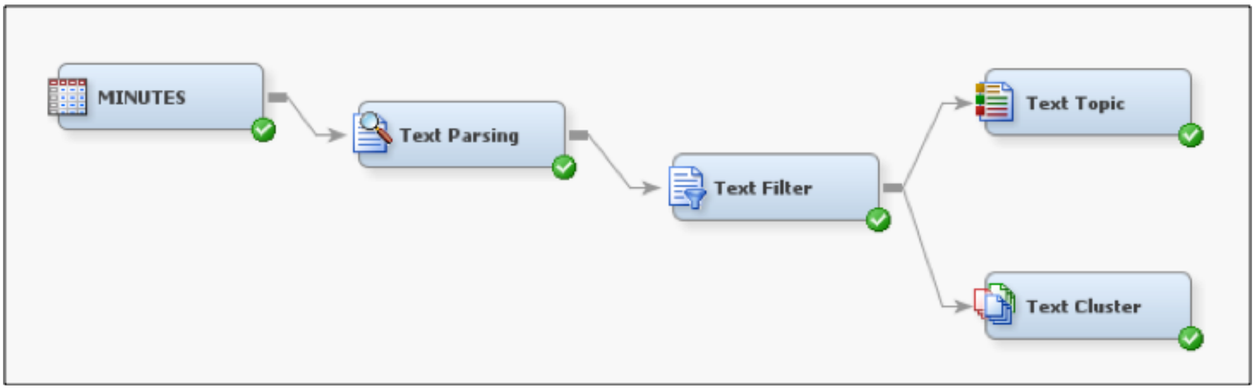
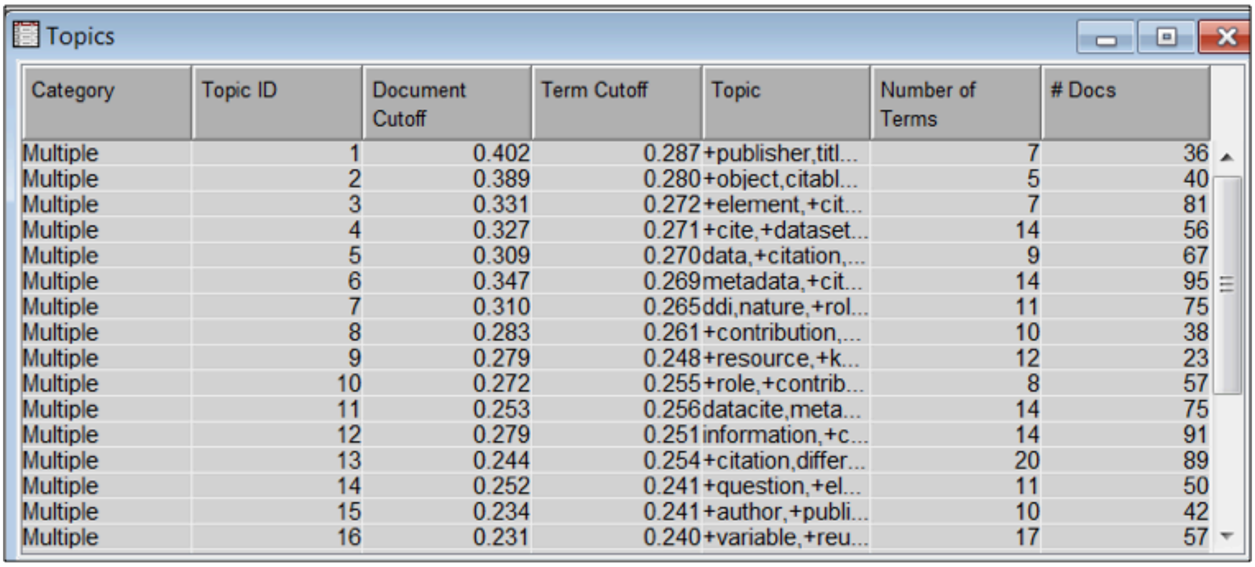


Figure 1. Screenshot of the Text Miner Process

1. A citable object in DDI should provide sufficient information to build a citation that can include one or more contributors, e.g., the name of the contributor.
2. Contributors can be classified by using a reference to the CRediT taxonomy, including weight of contribution (lead, equal, and supporting) as properties of Contributor. This is the recommended practice, although it should be possible for a different taxonomy that includes contributor roles to be referenced and used.
3. The same weight scale of contribution from CRediT can be added as a property of Creator. Creator may also have a property of Role.
4. DDI should collaborate with the Harvard-Wellcome initiative and give input to close any gaps and align mismatches such that a shared taxonomy is applicable to the DDI Lifecycle in particular and scholarship of data in general.
5. The DDI CVG should consider expanding its DDI Controlled Vocabulary for Contributor Roles and continue its examination of other similar lists of roles such as those created by DataCite and ONIX. It should situate its Controlled Vocabulary for Contributor Roles into its Lifecycle Events for internally consistent mapping.

A SAS Dataset Use Case

To apply the principles and practices being discussed, the group created several use cases, including one in which a quantitative dataset was created from a qualitative dataset through text mining. For the latter we chose the raw minutes created as Google Docs during the first three days of the meeting. The derived dataset had


Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.402	0.287	+publisher,titl...	7	36
Multiple	2	0.389	0.280	+object,citabl...	5	40
Multiple	3	0.331	0.272	+element,+cit...	7	81
Multiple	4	0.327	0.271	+cite,+dataset...	14	56
Multiple	5	0.309	0.270	data,+citation,...	9	67
Multiple	6	0.347	0.269	metadata,+cit...	14	95
Multiple	7	0.310	0.265	ddi,nature,+rol...	11	75
Multiple	8	0.283	0.261	+contribution,...	10	38
Multiple	9	0.279	0.248	+resource,+k...	12	23
Multiple	10	0.272	0.255	+role,+contrib...	8	57
Multiple	11	0.253	0.256	datacite,meta...	14	75
Multiple	12	0.279	0.251	information,+c...	14	91
Multiple	13	0.244	0.254	+citation,differ...	20	89
Multiple	14	0.252	0.241	+question,+el...	11	50
Multiple	15	0.234	0.241	+author,+publi...	10	42
Multiple	16	0.231	0.240	+variable,+reu...	17	57

Figure 2. The Topics Table

as its unit of analysis the topics computed by the text mining software. We decided to also create an example variable to show how it could have structured annotation information useful for a citation. It became clear that the text mining procedure could itself serve as an example instrument, in that it is essentially a 'black box' with a set of inputs – data and parameters -- and an output – a dataset. A structured annotation for this procedure would include documentation of all of these inputs. Since the set of parameters for the text mining procedure is unique to this software (SAS), this annotation would need its own Instrument description-type. An external vocabulary could be developed and referenced allowing the recording of the parameters used to generate any other dataset with the same software.

We exported the minutes of the first three days of the workshop from three separate Google Docs into Microsoft Word and concatenated them into a single text file in UltraEdit. This process made each paragraph in the original documents into a single line in the text file. Then we wrote a SAS program to read the minutes into a SAS dataset.

A SAS Enterprise Miner, Text Miner process produced a Topics dataset and a Clusters dataset from the Minutes dataset using the default options (see Figure 1). All of the parameter values for the default options were exported to an XML file to allow for future replication of the process.

From the Results Window, we saved the Topics results as a SAS Dataset (Figure 2), and the resulting dataset was then modified in SAS Enterprise Guide. A new variable was computed, combining the topic number, the number of documents using the topic, and the list of most highly weighted terms for the topic. Since the variable names for the Topic Result table are standard, this is a reusable variable that can be recomputed. We used the Topics2 dataset and the new variable (TopicDescription) as objects to be cited.

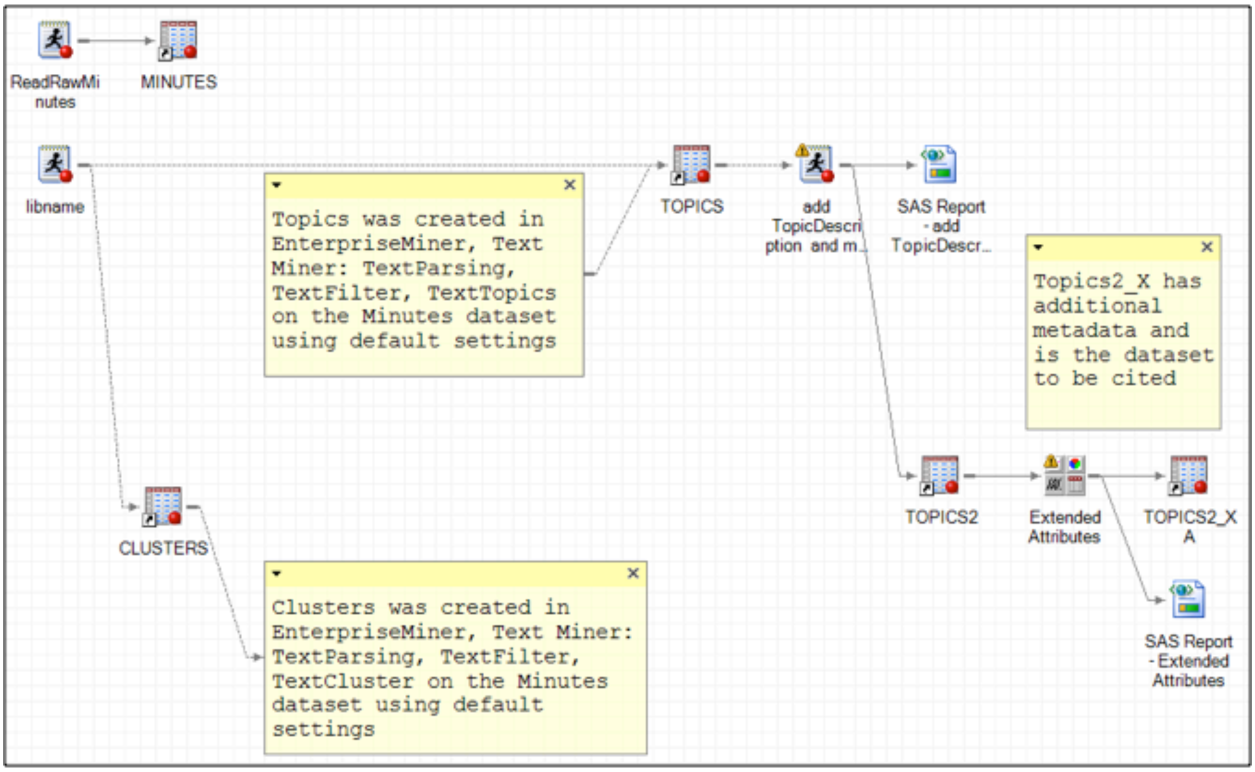
Using an Enterprise Guide (EG) add-in (Hoyle 2013), we added additional metadata to the SAS dataset and generated a DDI3.2 instance and a codebook from that. This additional metadata included attribution information (creator, contributors, funding information, etc.) and other descriptive information (coverage, methodology, etc.). See Appendix 1 and Appendix 4 of the full use case and documentation, available in KU Scholarworks. The complete set of extended attributes for the dataset is listed in Appendix 5. Extended attributes for the variable TopicDescription are listed in Appendix 6.

Figure 3 shows the Enterprise Guide process flow diagram. Text Miner is a separate application so that is represented in the flow by a note.

Representing structured annotation information in DDI3.2

The DDI3.2 instance for our example dataset (Topics2) includes the following elements, reflecting the need for information on attribution (who, what, when, where, whether), and characterization (how, structure, and provenance).

Who – Creator, Contributor, FundingInformation
We did not include institutional responsibility or a reference to a persistent researcher identifier. DDI3.2 allows both Creator and Contributor to reference a structure which


ReadRawMinutes
MINUTES
libname
TOPICS
add TopicDescription and m...
SAS Report - add TopicDescr...
CLUSTERS
TOPICS2
Extended Attributes
TOPICS2_XA
SAS Report - Extended Attributes

Topics was created in EnterpriseMiner, Text Miner: TextParsing, TextFilter, TextTopics on the Minutes dataset using default settings

Topics2_X has additional metadata and is the dataset to be cited

Clusters was created in EnterpriseMiner, Text Miner: TextParsing, TextFilter, TextCluster on the Minutes dataset using default settings

Figure 3. Enterprise Guide Process Flow

can contain references to external persistent identifiers (like ORCID). It is not clear, though, how to document institutional responsibility for different phases of the data lifecycle in DDI3.2.

What - Title, Description, Abstract, Version

When - Creation Date, Modification Date, and Publication Date
In some datasets there may be other date/time references required. Retrospective studies may ask respondents about some time period in the past. These can be documented in r:TemporalCoverage.

Where - Publisher, Pointer to metadata, Actionable link to dataset
We provided a Handle (Hoyle et al. 2014) pointing to a landing page having links to the data files and associated metadata. This (ScholarWorks) landing page is not really structured to provide a persistent actionable link to each dataset, although it does provide a separate URL to each.

Whether - Access Rights, Copyright, License, Permanence
It is not clear whether AccessRights and License are both required.

How - ProcessingDescription, GenerationInstruction, Language, CollectionMethodology, RelatedResource
The metadata includes descriptive text about the method used to generate the dataset, its source, and collection methodology.

Structure - LogicalProduct, PhysicalInstance
A DDI3.2 description of the data was harvested from the SAS dataset. This should allow machine-actionable interpretation of the structure of the dataset.

Provenance – Provenance is present only as unstructured descriptions in CollectionMethodology, ProcessingDescription, and GenerationInstruction.

Citation-related information for the dataset in DDI3.2

All of the elements we propose for versionable objects (see Table 1) could be documented in DDI3.2 with the following caveats.

- Contributor was entered as a structured string including both role and degree of contribution. DDI3.2 would allow multiple <r:Contributor> elements, each of which could include a r:<ContributorRole>, but not a degree of contribution.
- Copyright information was not provided but could be structured in a <r:Copyright> element.
- License appears as <dc:accessRights>. In DDI 3.3 <dc:license> will be available.
- In many small research projects UserID may not be clear in the context of a dataset which is being developed. A dataset has a name, unique within a file system folder, but not necessarily unique outside of that context. A DDI identifier is not certain to remain the same during development of the file. Once archived, a dataset will probably have a unique identifier.
- The description also included coverage information – spatial, temporal, and topical (subject). Note also that creation date, modification date, and publication date were all included.

Citation-related information for a variable in DDI3.2

We created a new variable (TopicDescription) that could be reusable with the Topic Result table from any SAS Text Miner Text Topic node instance. The Citation information for that variable is different from the dataset. DDI3.2 doesn't allow an r:Citation element to be attached to a variable, so the citation information was structured in a set of r:UserAttributePair elements. These are listed in Appendix 6. Note that the Creator and Contributor information for the variable is different than for the dataset as a whole. Allowing structured annotation information to be attached to any versionable object in DDI4 will make the structure of this information more consistent.

Instrument description

The Text Mining procedure used to produce the Topics dataset can be considered as a 'black box' instrument that takes a text dataset and produces a quantitative dataset. Documenting the use of this instrument to allow someone to reproduce the results requires recording all of the parameter choices made in using this instrument. This instrument description description-type can require a large number of information objects unique to the particular instrument. Appendix 2 shows the values of the 96 properties set for the run of Text Miner used to generate the topics dataset. At the time of this analysis these were the 'default' choices, but there is no guarantee that the default values will remain the same for future versions of the software so listing them is important for replication. SAS Enterprise Miner allows the export of diagram properties as an XML file. The tables shown in the appendix were processed from that file.

In the case of Text Miner many properties are relevant to only one node in the process. These properties, for example, are only relevant for the TextParsing node:

- Delimit = Std,
- bCapitalize=Y,
- stopList=SASHELP.ENGSTOP,

Each of the nodes in this process has its own set of inputs and outputs and might be considered sub-instruments linked by their inputs and outputs.

Some of these properties, like stoplist, point to a data file (SASHELP.ENGSTOP). This is a list of terms that will be ignored in the computations within and following the text parsing node. In this case, then, an input parameter can be complex – e.g., the contents of another dataset.

Sample citations[16]

Here we show how citations for the dataset and the variable might be listed in three common styles.

Use case dataset
APA – Hoyle, Larry (2014). *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432* [data file, codebook, DDI metadata] http://hdl.handle.net/1808/15746.

MLA - Hoyle, Larry. *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432*. University of Kansas, 2014. Web. 17 Nov 2014.

Chicago - Hoyle, Larry. *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432*. Lawrence Kansas: University of Kansas. 2014. http://hdl.handle.net/1808/15746.

All three styles for citing a dataset leave out contributors and cited author roles:

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)

Variable – TopicDescription
APA – Hoyle, Larry (2014). Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From a SAS Text Miner Text Topics Node Result Table [variable]. http://hdl.handle.net/1808/15746.

MLA - Hoyle, Larry. Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table. University of Kansas, 2014. Web. 17 Nov 2014.

Chicago - Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table. Lawrence Kansas: University of Kansas. 2014. http://hdl.handle.net/1808/15746.

In the three examples above only the APA style indicates that the object being cited is a variable. The MLA style doesn't yield a persistent identifier. The handle shown above points to a landing page (Hoyle et al. 2014) with a description of the collection and more than a dozen URLs to objects within the collection (original raw data, software code, a codebook, a DDI instance). An explicit link to the data file and another explicit link to the structured metadata for the data would be much more machine-actionable.

None of these styles allow for designation of a role or degree of contribution for the creator or a listing of contributors and their roles. If the standard citation styles included an explicit reference to structured metadata, including some mechanism for identifying the structure style, both of these problems could be handled by machine-actionable searching of the metadata.

Implications for DDI4

Following the Dagstuhl meeting the DDI4 model will have the following features supporting enhanced citation:

- All objects in DDI4 are identifiable (except for a few primitives)
- All versionable objects will have an annotation
- An annotation can have attribution information including Creator and Contributor
- Creator and Contributor can have a list of role, degree of contribution pairs
- The annotation should include the possibility of an additional set of information structured from an external vocabulary (scheduled for release 2 of DDI4)

This last feature addresses the need for DDI4 to have a mechanism to allow the incorporation of a set of information objects with a vocabulary drawn from an external controlled vocabulary. Ideally this mechanism will include the capability of validating those objects and also indicating relationships among the objects. Designing this mechanism will be a task for the DDI4 modeling group.

This need comes up both for instrument parameters and for the vocabulary for Creator and Contributor roles. The latter might have a hierarchical structure. At the top level of this hierarchy we propose using the CRediT taxonomy (Appendix A). In a hierarchy a 'software' role, for example, might be more specifically described as 'algorithm development.' DDI3.2 allows for attaching role to Contributor but not Creator. In a large study co-principal investigators may have specialized roles that should be documented. In each case role should also be paired with a 'degree of contribution' measure. We propose using the CRediT taxonomy as the top level of a taxonomy for describing role and a three-level category (e.g., 'lead', 'equal', and 'supporting') for degree as in the CRediT proposed standard.

Input parameters may also be complex objects, including datasets, as noted for the stoplist dataset. Parameters might also come from stream sources at specific times.

The group recommendation to allow structured annotation information to be attached to any versionable object will yield a more consistent structure for this information. For citation type information the addition of Role and DegreeOfContribution to Creator and Contributor, along with the elements already present in DDI3.2, should allow for a usable set of information.
For other description types, though, DDI4 will need to support external controlled vocabularies for attribute names and complex data types (including datasets) for attribute properties.

Conclusion

We began the meeting at Dagstuhl with a set of questions related to enhancing the citation information available in DDI, with the goal of informing the initial releases of DDI4. Our initial thoughts were focused mainly on attribution of credit for intellectual creation (a traditional citation). Along the way, however, we realized that there are other classes of information that can be referenced like a citation. Some of these are fairly well understood, like the information characterizing a dataset. This is the information that the DDI initiative has been developing over its 20-year history – what the data represent, when they were collected, how they were collected, why they were collected, and whether they can be used. Structure for other information may not be so well developed, or may be known only to a special community. We realized the need for DDI to be able to point to external structures and to incorporate that information for specific cases as needed. There may be many cases in which attribution information will need to be recorded along with this additional, externally structured, information.

Other questions cannot be addressed by the structure of DDI. Common adoption of conventions for contributor role and degree of contribution will result in a significant expansion of the information expected to be provided in a citation. Imagine reference sections and curricula vitae in fields where large scale multi-authorship is common if all the roles and degrees of contribution were to be listed. Clearly some sort of common infrastructure for locating structured annotation information is needed. There are such efforts under way (e.g., CASRAI[17], DataCite).

Requesting or requiring additional attribution information will increase the demand on researchers to provide metadata, already often seen as a burden. Future work might address the kinds of tools that could lessen this burden. Incorporating the collection of metadata, including attribution-related information, into the research workflow rather than considering it an additional task to be undertaken at the conclusion of a project might help ease the friction and improve the quality of the metadata. Properly designed tools might help. Training in research practices might also encourage better practices (e.g., Long 2009).

Finally, we leave it to others to develop algorithms and software to generate metrics for enhanced citation and to determine the best way to store, harvest, and display this information in a machine-actionable way. Multi-dimensional information could be collected – roles by degrees by numbers of collaborators, as well as the traditional creator vs. contributor distinction. Will a univariate metric be adequate, or should this complex of information be represented in a more nuanced simplification? Contributorship is clearly a fruitful topic for further research.

Contributors

Contributors to the project leading to this paper are listed below along with their roles. Attribution of roles was self-assigned using a web based form. Roles and degree of contribution are structured using the CRediT taxonomy.

Micah Altman – **Conceptualization**, Supporting; **Methodology**, Supporting; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, None; **Resources**, None; **Data Curation**, None; **Writing – original draft**, None; **Writing – review & editing**, None; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

Jay Greenfield - **Conceptualization**, Equal; **Methodology**, Equal; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation,** Equal**; Resources**, None; **Data Curation**, None; **Writing – original draft**, Equal; **Writing – review & editing,** Equal; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

Larry Hoyle (ORCID http://orcid.org/0000-0002-8262-2393) - **Conceptualization**, Lead; **Methodology**, Lead; **Software**, Lead; **Validation**, None; **Formal analysis**, Lead; **Investigation**, Equal; **Resources**, None; **Data Curation**, Lead; **Writing – original draft**, Equal; **Writing – review & editing**, Lead; **Visualization**, Lead; **Supervision**, Lead; **Project administration**, Lead; **Funding acquisition**, Lead

Sam Hume - **Conceptualization**, Equal; **Methodology,** Equal; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, Equal; **Resources**, None; **Data Curation**, None; **Writing – original draft**, Equal; **Writing – review & editing**, Equal; **Visualization**, None; **Supervision,** None; **Project administration**, None; **Funding acquisition**, None

Sanda Ionescu - **Conceptualization**, Equal; **Methodology**, Equal; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, Equal; **Resources**, Equal; **Data Curation**, None; **Writing – original draft**, Supporting; **Writing – review & editing**, None; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

Jeremy Iverson (ORCID http://orcid.org/0000-0003-3002-9245) - **Conceptualization**, Equal; **Methodology**, Supporting; **Software**, None; **Validation**, Supporting; **Formal analysis**, None; **Investigation**, None; **Resources,** None; **Data Curation**, None; **Writing – original draft**, Supporting; **Writing – review & editing**, Supporting; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

John Kunze (ORCID http://orcid.org/0000-0001-7604-8041) - **Conceptualization**, Equal; **Methodology**, Equal; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, Equal; **Resources**, None; **Data Curation**, None; **Writing – original draft**, Equal; **Writing – review & editing**, Supporting; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

Nancy Cayton Myers - **Conceptualization**, None; **Methodology**, None; **Software**, None; **Validation,** None; **Formal analysis**, None; **Investigation**, None; **Resources**, None; **Data Curation**, None; **Writing – original draft**, None; **Writing – review & editing,** None; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, Supporting

Barry Radler - **Conceptualization**, Supporting; **Methodology**, Supporting; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, None; **Resources**, None; **Data Curation**, Supporting; **Writing – original draft**, Equal; **Writing – review & editing**, Supporting; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition**, None

Wendy Thomas - **Conceptualization**, Supporting; **Methodology**, Supporting; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation,** Supporting; **Resources**, None; **Data Curation**, None; **Writing – original draft**, None; **Writing – review & editing**, Equal; **Visualization**, None; **Supervision**, Supporting; **Project administration**, None; **Funding acquisition**, None

Mary Vardigan (ORCID http://orcid.org/0000-0002-6168-6531) - **Conceptualization**, Lead; **Methodology**, Lead; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, Equal; **Resources**, None; **Data Curation**, None; **Writing – original draft**, Lead; **Writing – review & editing**, Lead; **Visualization**, None; **Supervision**, Equal; **Project administration**, Lead; **Funding acquisition**, Lead

Joachim Wackerow - **Conceptualization**, Equal; **Methodology**, None; **Software,** None; **Validation**, None; **Formal analysis**, None; **Investigation**, None; **Resources**, Supporting; **Data Curation**, None; **Writing – original draft**, None; **Writing – review & editing**, Supporting; **Visualization**, Supporting; **Supervision**, Supporting; **Project administration**, None; **Funding acquisition**, Supporting

Stuart Weibel - **Conceptualization**, Equal; **Methodology**, Equal; **Software**, None; **Validation**, None; **Formal analysis**, None; **Investigation**, Equal; **Resources**, None; **Data Curation**, None; **Writing – original draft**, Equal; **Writing – review & editing**, Equal; **Visualization**, None; **Supervision**, None; **Project administration**, None; **Funding acquisition,** None

Travis Weller - **Conceptualization**, None; **Methodology**, None; **Software**, None; **Validation,** None; **Formal analysis**, None; **Investigation**, None; **Resources**, None; **Data Curation**, None; **Writing – original draft**, None; **Writing – review & editing,** None; **Visualization**, None; **Supervision**, None; **Project administration,** None; **Funding acquisition**, Supporting

Michael Witt (ORCID http://orcid.org/0000-0003-4221-7956) - **Conceptualization,** Equal; **Methodology**, Equal; **Software**, None; **Validation**, None; **Formal analysis**, Supporting; **Investigation**, Equal; **Resources**, Supporting; **Data Curation**, Supporting; **Writing – original draft**, Equal; **Writing – review & editing**, Supporting; **Visualization**, Supporting; **Supervision**, Suppporting; **Project administration**, Supporting; **Funding acquisition**, None

References

Allen, L., Scott, J., Brand, A., Hlava, M. & Micah Altman (2014) Publishing: Credit where credit is due? Nature. 508 (April). p. 312–313. doi:10.1038/508312a. Available from: http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033. [Accessed: 28/01/2015]

Borgman, C. (2012) Why Are the Attribution and Citation of Scientific Data Important? In Ulhlir, P. (rapporteur). For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. ISBN: 978-0-309-26728-1. Available at: http://nap.edu/catalog.php?record_id=13564. Washington, DC: The National Academies Press. [Accessed: 28/01/2015]

Bourne, P.E., Clark, T.W., Dale, R., de Waard, A., Herman, I., Hovy, E.H. & David Shotton (2012) Improving the Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 1133). In Dagstuhl Manifestos. ISBN 2193-2433. 1 (1). doi: 10.4230/DagMan.1.1.41.

CASRAI (2015). CRediT -- An open standard for expressing roles intrinsic to research. Available from: http://credit.casrai.org/. [Accessed: 28/01/2015]

Center for Disease Control (CDC)--Behavioral Risk Factor Surveillance System (BRFSS) (2015) Suggested Citation Styles. http://www.cdc.gov/brfss/questionnaires.htm#citation. [Accessed: 29/10/2014]

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013) Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal 12 (2013). p. CIDCR1-CIDCR75. Available from: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf. [Accessed: 29/10/2014]

DataCite. DataCite Metadata Kernel (2014) Available from: http://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf [Accessed: 29/10/2014]

FORCE11-The Future of Research and Communications E-Scholarship (2014) Data Citation Principles. Available from: https://www.force11.org/datacitation [Accessed: 29/10/2014]

Häussler, C., & Sauermann, H. (2014) The Anatomy of Teams: Division of Labor and Allocation of Credit in Collaborative Knowledge Production (May 7, 2014). Available from SSRN: http://ssrn.com/abstract=2434327 or doi10.2139/ssrn.2434327.

Hoyle, L. (2013) Using Extended Attributes in Data Analysis Software - Controlled Vocabularies, Tools and DDI. In: Proceedings of the 5th Annual DDI Users Conference December 2013, Paris, France. Available from: http://www.eddi-conferences.eu/ocs/index.php/eddi/EDDI13/paper/viewFile/86/86. [Accessed: 28/01/2015]

Hoyle, L. (2014) A SAS Dataset Use Case for Enhanced Data Citation. Available from: https://kuscholarworks.ku.edu/bitstream/handle/1808/15746/ASASDatasetUseCaseForEnhancedCitation.pdf?sequence=19&isAllowed=y. [Accessed: 28/01/2015]

Hoyle, L., Vardigan, M., Hume, S., Ionescu, S., Greenfield, J., Iverson, J., Kunze, J., Radler, B., Thomas, W., Weibel, S. & Witt, M.C. (2014) Comprehensive Citation Across the Data Life Cycle Using DDI - Work products from the NSF1448107 sponsored group attending Dagstuhl event 14432 in October 2014. Available from: http://hdl.handle.net/1808/15746

Long, J. Scott (2009) The Workflow of Data Analysis Using Stata. College Station, TX: Stata Press.

Mooney, Hailey How to Cite Data: General Info. Available from: http://libguides.lib.msu.edu/citedata. [Accessed: 28/01/2015]

Notes

1. Jay Greenfield began his professional career developing artificial intelligence applications in a university environment. These apps could both listen and speak. They were expert systems that assisted in medical diagnosis and circuit board trouble shooting.

In 1992 Jay joined Westat where eventually he became the leading technologist. At Westat Jay used his knowledge of artificial intelligence to create actionable metadata. He used actionable metadata to spawn survey research questionnaires, conduct data management and create data dictionaries.
While at Westat Jay led the development of data collection and data analysis systems for many major health and nutrition studies including the Medicare Current Beneficiary Survey (MCBS), the Medical Expenditure Panel Survey (MEPS), the integration of the Continuing Survey of Food Intake by Individuals (CSFII) with the National Health and Nutrition Examination Survey (NHANES) and the development of the original legislation for the Children's Health Insurance Program (CHIP). In 2008 Jay joined Booz Allen as a study and data management expert on the National Children's Study. Currently, Jay is a health informatics architect working with data standards, data standard groups and terminologies to annotate medical data with metadata in order to facilitate search, research and discovery.

Larry Hoyle (ORCID http://orcid.org/0000-0002-8262-2393) is a Senior Scientist at the Institute for Policy & Social Research at the University of Kansas and was Principal Investigator for the NSF grant (1448107) which funded the enhanced citation group at Dagstuhl event 14432. He is a member of the DDI Moving Forward Advisory Group. He was also the first chair of the North American Data Documentation Initiative Conference (NADDI). Correspondence may be addressed to LarryHoyle@ku.edu or 1541 Lilac Ln. Suite 607 Blake, Lawrence KS 66045-3129.

Sam Hume is Vice President of SHARE Technology and Services at CDISC. At CDISC he leads the SHARE project and co-leads the XML Technologies team. Sam has over 20 years of work experience in clinical research informatics. Previously, he worked as Director of IS Architecture at AstraZeneca, VP of Technical Operations at Phoenix Data Systems and Chief Technology Officer at CB Technologies. Sam has an MS in Information Science, MS in Telecommunications, and is completing his doctorate in Healthcare Informatics.

Sanda Ionescu has been with ICPSR since 1999, working to implement and support the Data Documentation Initiative (DDI) — an XML-based specification for social science data documentation. She manages DDI-related projects and serves as the ICPSR representative in the Expert Committee of the DDI Alliance. At ICPSR she is also part of a team that provides user support for data and documentation issues. Within the DDI Alliance she participates in the efforts to develop and promote the DDI standard, and leads the Controlled Vocabularies working group that produces classifications for study- and variable-level metadata. She holds an MA in Communication from the University of Massachusetts-Amherst, and a BA in English and French from the University of Bucharest, Romania.

Jeremy Iverson (ORCID http://orcid.org/0000-0003-3002-9245) is a co-founder and partner at Colectica where he helps build software to document statistical data using open standards. Previously, he was a programmer at the Wisconsin Longitudinal Study, working to process, document, and disseminate data for the long-running study. Jeremy is currently an invited expert on the Data Documentation Initiative's Technical Committee.

John Kunze (ORCID http://orcid.org/0000-0001-7604-8041) is an Identifier Systems Architect at the California Digital Library. A former BSD Unix hacker who helped standardize URLs and Dublin Core metadata, his current work focuses on the EZID service, the N2T resolver, ARK identifiers, dataset citation, and lightweight metadata dictionaries.

Barry Radler is a Researcher at the University of Wisconsin Institute on Aging. His research interests focus on understanding how human beings process information, make decisions, and behave in social, political, and marketing contexts. For the last 20 years he has explored, advocated, and implemented the use of information technologies to improve research processes and data. Past and ongoing projects include an optical character recognition system for survey data entry, investigating mode effects between mail and online surveys, using custom computer applications to identify the processes and output of different cognitive systems, and, most recently, the application of web-based metadata standards to document complex longitudinal datasets.

Wendy Thomas is the Director of the Data Access Core in the Minnesota Population Center (MPC) at the University of Minnesota and has been active in the data and information technology community for over 25 years providing data research and support services to academic, governmental, non-profit, and for-profit researchers. She has been a Coordinating Member of the State Data Center program since 1990 and is a former President of the U.S. Association of Public Data Users. She has been active in the work of the Data Documentation Initiative (DDI) since 1997, chairs the DDI Technical Committee and is a member of the DDI Moving Forward Advisory Group. Her work in the MPC covers the preservation and documentation of historical census data and supporting materials for the IPUMS International projects. Her major publications focus on data documentation and the impact of standards on data dissemination and preservation. For more information, http://users.pop.umn.edu/~wlt/

Mary Vardigan (ORCID http://orcid.org/0000-0002-6168-6531) holds the position of Archivist at the Inter-university Consortium for Political and Social Research (ICPSR) where she directs the ICPSR Collection Delivery Unit, which involves oversight of activities in the areas of Metadata, Publications, Web Site Development, User Support, and Membership Development. She also serves as Director of the Data Documentation Initiative (DDI), an international collaboration to establish a metadata standard for the social and behavioral sciences. She is involved in other projects related to data stewardship, including the Data Seal of Approval, the Research Data Alliance, and various efforts to promote data citation. She was co-PI for the NSF grant (1448107) that funded the enhanced citation group at Dagstuhl event 14432.

Stuart Weibel worked in OCLC Research for 25 years, where he contributed to research in web standards for libraries, digital libraries, and convened workshops that led to the formation of the Dublin Core Metadata Initiative.

Michael Witt (ORCID http://orcid.org/0000-0003-4221-7956) is the head of the Distributed Data Curation Center (D2C2) and an Associate Professor of Library Science at Purdue University. He is also the Project Director for the Purdue University Research Repository and Editor-in-Chief of Databib. Witt serves on the Organizational Advisory Board of the Research Data Alliance, the editorial board of Information Technology and Libraries, and the DMPTool Steering Committee. Sponsors for his research include the Institute for Museum and Library Services, Microsoft Research, and the Alfred P. Sloan Foundation. For more information, http://www.lib.purdue.edu/research/witt.

2. This meeting was funded in part by NSF grant 1448107. Facilities support was provided by Schloss Dagstuhl – Leibniz Center for Informatics, the site of the meeting (http://www.dagstuhl.de/14432).
3. DataCite https://www.datacite.org/
4. ORCID. http://orcid.org/
5. http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf
6. ISO 19773:2011, http://www.iso.org/iso/catalogue_detail.htm?csnumber=41769
7. National Library of Medicine Permanence Levels, http://www.nlm.nih.gov/psd/pcm/devpermanence.html
8. Universal Numeric Fingerprint, http://thedata.org/book/universal-numerical-fingerprint
9. DDI Lifecycle, http://www.ddialliance.org/what
10. DDI Lifecycle Events CV, http://www.ddialliance.org/Specification/DDI-CV/LifecycleEventType_1.0.html
11. Draft DDI Roles CV, [unpublished] https://docs.google.com/file/d/0B5aS3-mIMlfkQ3dpTnZIQ2hxTnM/edit
12. Report on the International Workshop on Contributorship and Scholarly Attribution (16 May 2012) http://projects.iq.harvard.edu/attribution_workshop/files/iwcsa_report_final_18sept12.pdf
13. Nature 508, 312–313 (17 April 2014) http://dx.doi.org/10.1038/508312a
14. Appendices - Numbered appendices are available in Hoyle 2014. Appendices A and B follow in this document.
15. URL for data - http://kuscholarworks.ku.edu/bitstream/handle/1808/15746/topics2.sas7bdat?sequence=11&isAllowed=y
URL for Extended Attributes companion file - http://kuscholarworks.ku.edu/bitstream/handle/1808/15746/topics2.sas7bxat?sequence=12&isAllowed=y

URL for DDI3.2 metadata instance - http://kuscholarworks.ku.edu/bitstream/handle/1808/15746/NSF1448107TopicsUseCase2014_11_09.xml?sequence=10&isAllowed=y

16. Example styles taken from From How to Cite Data: General Info, http://libguides.lib.msu.edu/citedata
17. CASRAI http://casrai.org/

A Note about Appendices: All numbered appendices are available online in the document: Hoyle 2014. Project data are archived at: https://kuscholarworks.ku.edu/handle/1808/15746.

Appendix A

Harvard/Wellcome Trust Taxonomy (CRediT Taxonomy)

A classification of the diverse roles played in the work leading to a research output. The classification includes, but is not limited to, traditional authorship roles. When there are multiple people serving in the same role a 'degree of contribution' should be further specified as either 'lead', 'equal', or 'supporting'. Roles are intended to apply to all those who contribute to a project — and it is recommended that, if possible, all contributors be listed, whether or not they are formally listed as authors. It is also intended that multiple roles be assigned to a single person where appropriate. Roles and their descriptions are listed below from http://credit.casrai.org/proposed-taxonomy/.

#1 conceptualization
Ideas; formulation or evolution of overarching research goals and aims.

#2 methodology
Development or design of methodology; creation of models.

#3 software
Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.

#4 validation
Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.

#5 formal analysis
Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data.

#6 investigation
Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.

#7 resources
Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.

#8 data curation
Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.

#9 writing – original draft
Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).

#10 writing – review & editing
Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

#11 visualization
Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.

#12 supervision
Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.

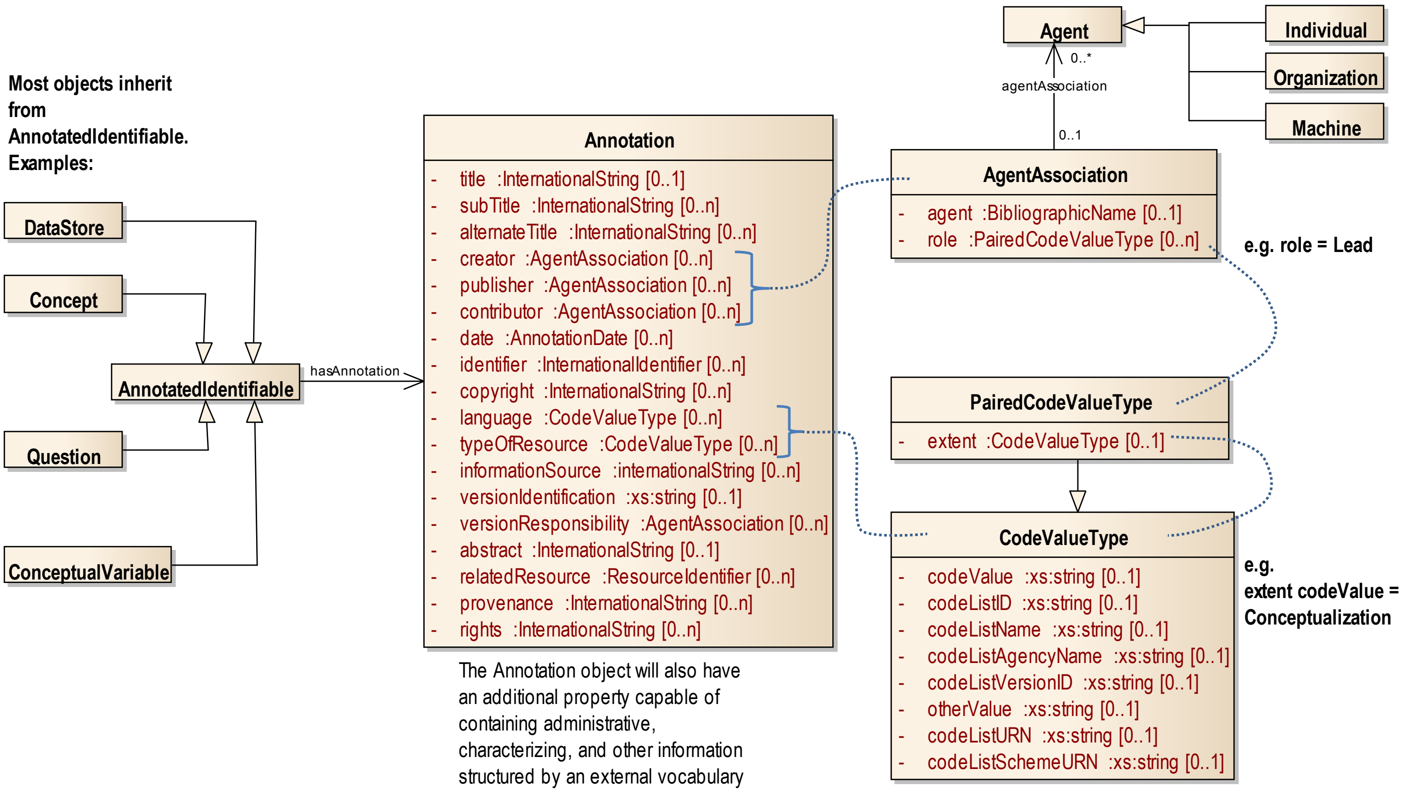
#13 project administration
Management and coordination responsibility for the research activity planning and execution.

#14 funding acquisition
Acquisition of the financial support for the project leading to this publication.

Appendix B

Citation Related Objects in the DDI4 Model

The diagram below shows the elements added to the DDI4 model during the Dagstuhl sprint and its immediate follow-up. All objects except for primitives and complex data types inherit from AnnotatedIdentifiable which, in turn, has an Annotation. An Annotation contains attributes of creator, contributor, and publisher of type AgentAssociation. An AgentAssociation has a role attribute of type PairedCodeValueType that allows a codeValue (a role from a specified set of roles) to be paired with an extent (a degree of contribution) also drawn from a specified vocabulary.


Agent
Individual
Organization
Machine
0..*
agentAssociation
0..1
AgentAssociation
- agent :BibliographicName [0..1]
- role :PairedCodeValueType [0..n]
DataStore
Concept
AnnotatedIdentifiable
hasAnnotation
Question
ConceptualVariable
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]
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]

Most objects inherit from AnnotatedIdentifiable. Examples:

e.g. role = Lead

e.g. extent codeValue = Conceptualization

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