Identifying Content and Levels of Representation in Scientific Data

Karen M. Wickett, Simone Sacchi, David Dubin, and Allen H. Renear
{wickett2, sacchi1, ddubin, renear}@illinois.edu
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211 USA

ABSTRACT
Heterogeneous digital data that has been produced by different communities with varying practices and assumptions, and that is organized according to different representation schemes, encodings, and file formats, presents substantial obstacles to efficient integration, analysis, and preservation. This is a particular impediment to data reuse and interdisciplinary science. An underlying problem is that we have no shared formal conceptual model of information representation that is both accurate and sufficiently detailed to accommodate the management and analysis of real world digital data in varying formats. Developing such a model involves confronting extremely challenging foundational problems in information science. We present two complementary conceptual models for data representation, the Basic Representation Model and the Systematic Assertion Model. We show how these models work together to provide an analytical account of digitally encoded scientific data. These models will provide a better foundation for understanding and supporting a wide range of data curation activities, including format migration, data integration, data reuse, digital preservation strategies, and assessment of identity and scientific equivalence.

KEYWORDS
Data Curation, Conceptual Modeling, Information Organization, Representation, Identity, Scientific Equivalence.

INTRODUCTION
Advances in the application of digital technologies in science have not only increased our ability to collect data in digital form (resulting in the "data deluge"), but also promise advances in techniques for analyzing this data. This promise however has not yet been fully realized. Heterogeneous digital data that has been produced by different communities with varying practices and assumptions, and that is organized according to different representation schemes, encodings, and file formats, presents substantial obstacles to efficient integration, analysis, and preservation. This is a particular impediment to the data reuse and interdisciplinary science that is needed to address complex real-world problems (Renear, Sacchi, & Wickett, 2010; Palmer, Weber, & Cragin, 2011). The traditional issues of heterogeneity are well known, if far from resolved. But there are even more fundamental, and troublesome problems. Many of these pertain to *identity*. What does it mean to say that two files in a computer system "contain" the same data, but in different formats? It is a commonplace in science to distinguish information from particular representations, but there is no adequately developed shared understanding of what this means.

The same information content can be represented differently in different data description languages. Moreover, not only can assertions in a data description language themselves be encoded in different encoding formats, but encodings themselves may be encoded. An abstract RDF triple may be encoded in any one of several RDF formats (RDF/XML, N3, Turtle, etc.) and those representations in turn may have different character encodings, which in turn may have different mappings into bit streams. Currently we do not have a well–defined framework for distinguishing and relating concepts such as these, and considerable confusion ensues. Within the earth sciences community for instance this problem has recently emerged in connection with dataset identifiers and has become known as the problem of "scientific equivalence" (Tilmes, Yesha, & Halem, 2010, 2011).

However developing a conceptual model for these notions raises additional deep and challenging issues in ontology and the semantics of languages for the expression of scientific data. In virtue of exactly what properties does a particular symbol structure express a particular piece of information? In virtue of what properties does a particular symbol structure encode a particular statement in a representation language? What sorts of information–preserving transformations are possible? Or, more generally, what features at what level are preserved by what transformations?

In what follows we present the Basic Representation Model and the Systematic Assertion Model, and show how these models work together to provide an analytical account of

ASIS&T 2012, October 26–31, 2012, Baltimore, MD, USA.

digitally–encoded scientific data. This account can serve as a foundation for answering the questions above. These models will provide a better foundation for understanding and supporting a wide range of data curation activities, including format migration, data integration, data reuse, digital preservation strategies, and assessment of data identity and scientific equivalence.

THE DATA CONSERVANCY

The National Science Foundation DataNet program has funded a number of projects charged with contributing to the development of a national infrastructure for data curation; supporting the documentation, integration, preservation, sharing, access, and analysis of scientific data in digital formats. Of particular concern is the integration of heterogeneous data from multiple sources, which is considered essential not only to realize full value from collected data, but to make progress on challenging complex problems facing society, problems which are typically interdisciplinary in nature.

The *Data Conservancy*, hosted at Johns Hopkins University Sheridan Libraries, is a multi-institutional project funded under the NSF DataNet program. It is tasked with researching, designing, implementing, and sustaining a data curation infrastructure for cross-disciplinary discovery, with an emphasis on observational data and an initial focus on data from astronomy, earth sciences, life sciences, and social sciences: "transforming the ability of scientists to answer grand challenge questions that are important to the nation and the world."

At the Center for Informatics Research in Science and Scholarship (CIRSS), Graduate School of Library and Information Science, University of Illinois at Urbana Champaign, two Data Conservancy projects are underway. The first, *Data Practices*, is studying the information behavior of scientists around the creation, management, sharing, and use of scientific data. The second group, *Data Concepts*, is developing a conceptual model of fundamental concepts related to scientific datasets. The premise of the Data Concepts agenda is that the reliable application of semantic technologies to scientific data curation and integration requires precisely defined shared understanding of key notions, such as dataset, format, encoding, version, file, and collection. The work reported on here is from the Data Concepts group.

TYPES OF MODELS

We can understand data modeling as a means to achieve specific objectives in data curation and data management. For data to be "preserved" and available for meaningful use and reuse over time, we have to precisely describe what is to be preserved, how it is to be preserved and for what purposes, and how to correctly interpret its content. Two major classes of models participate in achieving this goal: digital preservation models and scientific data models.

Scientific data models identify and describe entities and processes involved in the creation of datasets. The purpose of these models is to support retrieval and meaningful use and reuse of data, by making explicitly and computationally available meaning, context and provenance. These conceptual models, usually specified as ontologies, support the semantic annotation of data through their expression in knowledge representation languages like RDF, RDF/S and OWL. Examples of scientific data models are the Extensible Observation Ontology (OBOE) (Madin et al., 2007) and the Semantic Sensor Network ontology (SSN) (Lefort et al., 2011).

Digital preservation models identify and describe entities and processes involved in a preservation ecosystem — a digital information system or the broader socio-technical environment where preservation transactions occur — to support preservation planning. These models are usually general and flexible enough to be applied in a broad range of digital preservation scenarios: they specify how digital objects must be represented in an information system and how they interact with the preservation ecosystem. Examples of digital preservation models are the Open Archival Information System (OAIS) (CCSDS, 2002) and the Preservation and Long-term Access through Networked Services (PLANETS) (Farquhar & Hockx-Yu, 2008) models.

Both types of model are essential to the development of systems that support effective data curation. However, preservation models and scientific data models do not cover the entire spectrum of modeling requirements. Preservation models represent only a facet of the representational stack for data. They deal with digital objects, which is not — strictly speaking — what data and datasets are (Renear et al., 2010; Sacchi, Wickett, Renear, & Dubin, 2011a). On the other hand, scientific data models focus on supporting the scientific use of data, rather than the question of what data really are. Since neither preservation models nor scientific data models address the basic nature of data and datasets, a gap remains.

The Basic Representation Model and the Systematic Assertion Model (SAM) (Dubin, 2010) bridge this gap between preservation models and scientific data models. The Basic Representation Model is a summary of the key entities and relationships involved in the representation of digital objects. The Systematic Assertion Model provides an account of how these relationships come to be established for scientific data as well as more detail on their nature and interrelationships. An application of the two models is presented using a working example of biodiversity data.

A WORKING EXAMPLE

Our working example of a digital object that carries scientific data is a species occurrence record that describes the collection of a specimen of a *Mola mola*, or ocean sunfish. This record appears as a row in a file of occurrence records expressed using the Darwin Core schema (Darwin Core Task Group, 2009), a vocabulary of terms for describing biological specimens and species observations. The occurrence content of the Darwin Core Archive is a text file of tab-separated values, received through VertNet[1]. The archive also contains files describing the source of the record and

[1] `http://vertnet.org`

linking the fields used in the occurrence file to definitions from Darwin Core.

id	1821
scientificName	Mola mola
family	Molidae
order	Tetraodontiformes
class	Actinopterygii
genus	Mola
specificEpithet	mola
identifiedBy	Wiley, Martin
basisOfRecord	preserved specimen
collectionCode	KUI
institutionCode	KU
catalogNumber	32586
fieldNumber	MLW 34
preparations	1 EtOH
verbatimEventDate	1/8/65
verbatimLongitude	75.9000 W
verbatimLatitude	34.1217 N
locality	Atlantic Ocean, about 100 mi. E of Carolina Beach, North Carolina
minimumDepth	31
maximumDepth	31
recordedBy	Wiley, Martin L
year	1965
month	1
day	8
eventDate	1/8/65
decimalLongitude	-75.9000
decimalLatitude	34.1217
continent	Atlantic Ocean
country	USA
StateProvince	North Carolina

Table 1. A species occurrence record

Table 1 shows our species occurrence record for a *Mola mola* specimen as a table. For presentational reasons, we have ordered the attributes into meaningful groups and arranged the table vertically. As we argue later, the occurrence record row in the archive file and Table 1 express the same *data content*. In fact we argue that these are two different encodings of the same *primary symbol structure*.

Biodiversity records like this one are frequently shared, processed and re-aggregated within distributed networks for data sharing, such as VertNet, or the Global Biodiversity Information Facility [2]. We can imagine a research lab wishing to incorporate these text-based occurrence records into a XML-based system. The Darwin Core Task Group provides XML schemas for both formats, meaning that creating a transformation would be straightforward. The result of such a transformation would be an XML file that contained the record as an XML element, as opposed to the tab-delimited file, where the record appears as a row. We can also imagine that the lab, having an interest in Linked Open Data, produces an RDF expression of the occurrence data, using Darwin-SW[3].

In a sense, these transformations are straightforward. But what shall we say about the how the products of these transformations are related to one another and to the "original" text-based occurrence file? It seems natural to think that they are in some sense "the same", but examining bit-level representations of the files, or even character string representations, will not reflect this sameness (Renear & Dubin, 2003).

In order to make sense of the *sameness* relationship between the results of these transformation actions, we need a model that formally discriminates among the levels of representation of digital objects and supplies an account of data that is independent of any particular file-level instantiation.

THE BASIC REPRESENTATION MODEL

The Basic Representation Model identifies the core entities and relationships that are involved in representing the information carried by digital objects. This model has its roots in work designed to establish a general model for the preservation of resources in digital repositories (Sandore & Unsworth, 2010).

The Functional Requirements for Bibliographic Records provides a conceptual model that appears to identify entities and relationships relevant to our problems (IFLA, 2009). FRBR distinguishes common intellectual content (the work, or information), from varied symbolic expressions; expressions from their particular varied manifestations (such as variations in typeface, layout etc); and manifestations from their individual concrete physical instances. Not surprisingly FRBR has been recommended as a model for understanding representation issues for scientific data (Hourclé, 2008). However, although promising in some respects FRBR has a number of weaknesses, and attempting to systematically apply FRBR to scientific data in digital form makes these weaknesses evident.

First, the notion of an expression seems inadequately developed. FRBR would classify an XML/RDF serialization and an N3/RDF serialization as different expressions, but these are nevertheless both serializations of the same set of RDF triples. Those triples themselves however are not identified in the FRBR Group 1 ER diagram, despite the fact that they would seem to be what is most immediately and directly expressing information. This problem is not unique to digital objects: a printed text and an audio recording of that text both involve the same sequence of sentences, and those are the sentences that directly realize the work. But most applications of FRBR consider the printed text and the recorded text different expressions and do not separately identify their common sequence of sentences at all.

This problem is related to a more general one. FRBR makes a simple distinction between expressions and manifestations, but it appears that there are typically a number of expression-like layers of embodiment at the manifestation level. For

[2] `http://gbif.org`

[3] `http://code.google.com/p/darwin-sw/`

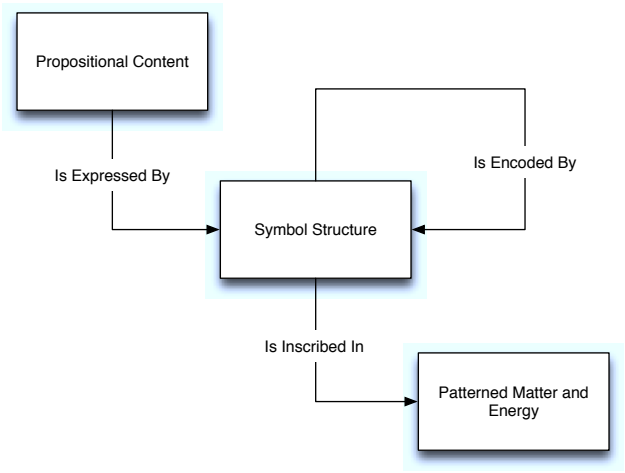

Figure 1. The Basic Representation Model

instance, RDF triples can be mapped to a particular serialization which in turn is mapped to some sequence of scalar values, which in turn can be mapped to octets, etc. Each layer appears to be a "notational entity", but only the first is directly expressing information, the rest are better described as encoding other symbol structures. And any number of layers is possible. We need a model that recognizes that in digital objects, there may be a cascade of representational layers.

Another problem with FRBR is that its entity types appear to not represent fundamental *types* of things, but rather *roles* that fundamental things enter into in particular circumstances (Renear & Dubin, 2007; Guarino & Welty, 2000). This makes it hard to identify what features are contingent properties and what features apply to fundamental types, as well as making extension and refinement of the model convoluted. As a criterion for distinguishing types from roles we adapt Guarino and Welty and apply this rule: If it is possible that something that is an F might not have been an F, then being an F is a role that things have; otherwise F is a type of thing. So, using their example, since it is possible that something that is student might not have been a student (i.e., might not have enrolled this year), *student* is role. But since it is not possible that something that is a person might not have been a person (and still exist), *person* is a type of thing.

Entity types

One of the goals of this research is to clearly discriminate between the genuine entities (*types*) and the relationships (*roles*) those entities those type enter into in the context of scientific data.

Distinguishing types from roles allows us to reduce the number of first class entities in our model — classes that are asserted as entity types. This modeling strategy provides a more parsimonious and consistent ontological treatment of digital objects. Only three kinds of things seem to participate in the representation of digital objects in general, and of digitally–encoded data in particular: *Propositional Content*, *Symbol Structure*, and *Patterned Matter and Energy*.

Propositional Content In our model propositions appear as the language–independent content expressed by symbol structures. In the sense intended propositions may be defined as *all and only those things that are either possibly true or possibly false*. That is, they are the proper subjects of truth values. The symbol structure that expresses a proposition may also be considered true or false, but only in a derivative sense: derivatively "true" if the proposition it expresses is true, and derivatively "false" if the proposition it expresses is false. A common alternative account of propositions defines them as the proper objects of epistemic attitudes, such as belief or doubt. For our purposes these two accounts of proposition may be considered coextensive: the class of things that can be true or false is identical with the class of things that can be the object of epistemic attitudes. Although the significant role of propositions in our model is as the expressed content of symbol structures, the definitions just given allow propositions to exist independently of symbol structures.

In our Mola mola species occurrence record, the propositional content might be understood as including *a Mola mola was collected on 1/8/65*. The proposition per se is independent of the scientific language used to express this fact. A record written in natural language (such as the sentence we just used) can express the exact same propositional content as the Darwin Core record. Many different records can therefore express the same proposition, and the proposition can be true (or false) even if it is not expressed by any record at all.

Symbol Structure In our model symbol structures are abstract arrangements of symbols that, in a given context, express propositions. Individual symbols themselves are the atomic components of symbol structures. Although the symbol structures in our examples are in some language with a determinate semantics, our model allows symbols and symbol structures to express different propositions in different languages or different contexts. Examples of abstract objects that can serve as symbol structures include graphs, relations, and sequences, along with more familiar kinds of symbol structures like strings of characters.

Patterned Matter and Energy Whereas both propositions and symbol structures are abstract objects, patterned matter and energy is a concrete quantity of matter and energy that manifests a physical arrangement that is the physical inscription of an (abstract) symbol structure. In order for a digital object to effectively communicate information, there must be some instantiation of the symbol structures in a physical medium that an agent can interact with.

Relationship types

These entity types participate in the representation of digital objects through a set of contingent relationships. How these relationships come to be established as well as more detail on their structure is provided in Systematic Assertion Model.

Our model has three key relationship types:

Is Expressed By Every meaningful digital object will use symbol structures to express propositions. For instance, a digital object may use RDF triples to express propositions about species occurrence. We use the *is Expressed By* relationship type for this technical sense of "express". The *Is Expressed By* relationship type represents the fact that the propositional content of a digital object is understood as being expressed by a symbol structure that is the primary expression — the *Primary Symbol Structure* — for that content in a particular context. *Is Expressed By* represents a general relationship that is instantiated between specific propositional content and a specific symbol structure. An event–based account of how this relationship is actually instantiated for scientific data is provided by the Systematic Assertion Model.

Is Encoded By A digital object will typically map the symbol structures that express propositions into other symbol structures. We call this mapping from symbol structure to symbol structure an *encoding* of one symbol structure by (or into) another. For instance, a digital object may map RDF triples into the XML/RDF serialization language. Or those same triples might be encoded in the N3 serialization language. In each case we have the same *Primary Symbol Structure* – the RDF triples that express propositional content – but a different *encoding* of that primary symbol structure. Symbol structures that are encodings of other symbol structures may in turn be encoded by still other symbol structures. For instance the N3 symbol structure may itself be encoded in a UTF-8 byte sequence. Unpacking the encoding levels provides a more complete and consistent way to represent what changes when digital objects undergo transformations, like format migrations.

Is Inscribed In The *Is Inscribed In* relationship type represents the fact that a particular symbol structure is represented in a physical medium through a mapping between the symbol structure and a particular concrete arrangement of matter and energy.

Applying the model

Below, Figure 2 shows an extract from a text file of tab–delimited values that contains the *Mola mola* record presented earlier, and Figure 3 is an extract from an XML version of the record, transformed according to available guidelines for Simple Darwin Core. We can use the entities and relationships from the Basic Representation Model to analyze these two versions of the record and to prepare for an analysis according to the Systematic Assertion Model.

```
id minimumDepthInMeters year scientificName
1821 31 1965 Mola mola
```

Figure 2. Extract from the text-based occurrence file

```
<SimpleDarwinRecordSet>
 <SimpleDarwinRecord>
  <dc:identifier>1821</dc:identifier>
  <dwc:minimumDepthInMeters>31</dwc:minimumDepthInMeters>
  <dwc:year>1965</dwc:year>
  <dwc:scientificName>Mola mola</dwc:scientificName>
```

Figure 3. Extract from an XML version of the example

The XML record expresses propositional content pertaining to the collection of the *Mola mola* specimen in the year 1965. This propositional content is the same for both versions of the record. In addition, both versions of the record share the same primary symbol structure for expressing the propositional content.

The primary symbol structure that these two versions of the record share in common is a graph structure that assigns values from chosen domain vocabularies to attributes that are defined as elements of the Darwin Core Schema. The graph takes the form of a set of subject-predicate-object triples that matches each attribute of the Darwin Core Occurrence with the associated value[4]. The two different versions of the data arise from different encodings of this single graph structure. Figure 2 shows the graph structure encoded as tab-delimited values (each row refers to the same subject, the header gives the predicate and the object is given by the cell value), while Figure 3 shows the triples encoded as XML elements. In addition, Table 1 shows another encoding of these triples into an arrangement of rows and columns.

These encodings may also stand in a variety of encoding relationships with other symbol structures. In the record we received from VertNet, the text-based occurrence file was encoded as ASCII characters. The XML file would likely be encoded as UTF-8 characters, or according to UTF-16. These encodings may be inscribed into patterned matter and energy by being stored on a disk-based storage device, or by being written onto magnetic tape. Alternatively, the characters may be inscribed by printing the associated graphemes on to paper with ink.

The Basic Representation Model provides an account of digital information objects in terms of levels of representation and representation relationships. It does not however indicate how these objects came to enter into these relationships, spell out what entities and events are involved in creating and sustaining these relationships, or provide full details on how these events are situated in the context of scientific observation and communication. The Systematic Assertion Model, presented in the next section, addresses these critical aspects of scientific communication.

THE SYSTEMATIC ASSERTION MODEL

The Systematic Assertion Model focuses on key provenance events through which propositional content and symbol structures acquire the status of "data content" and "data," respectively. Propositional data content serves as a locus for tracking identity across different encodings of data, or the combination of two or more datasets together.

Entities of direct interest to scientists (i.e., the subjects of data content) may sometimes be directly observed specimens or phenomena. More often, though, observations serve as evidence for claims about properties not directly observed or for generalization from observed particulars. For example, an instrument reading may be recorded as evidence for a

[4]This triple structure can equivalently be viewed as assigning an attribute/value pair to an individual.

particular temperature or voltage. Evidence for claims about a species of fish are drawn from observation of particular fish, and so forth.

Justification for scientific claims can also come from computations performed over existing data. Execution of such procedures can create new data content, or put existing content into a new expression. As with observations, computational events should be documented for purposes of replication and verification. Symbols are connected to their meanings via contingent interpretive conventions (Dubin, Wickett, & Sacchi, 2011), and the layered encodings of digital information objects call for special care in making these conventions clear and explicit.

As defined in the previous section, propositions and symbol structures are, in our model, rigid types that supply identity conditions (Guarino & Welty, 2000). Their instances can acquire relational properties, such as a proposition being the substance of a scientist's assertion, or a symbol structure serving as the primary expression for that same assertion.

SAM Axioms

The key definitions of our model are summarized by the following axioms, expressed according to the conventions of the ALC language for description logics (Baader & Nutt, 2003).

Propositional Content and Symbol Structures

The following axioms describe classes for modeling propositional content in the Systematic Assertion Model.

$$Proposition \sqsubseteq AbstractThing \tag{1}$$

$$Conjunction \sqsubseteq Proposition \tag{2}$$

$$SymbolStructure \sqsubseteq AbstractThing \tag{3}$$

Description logic classes are sets of individuals in the overall domain of description (sometimes called "top class"), and the "$\sqsubseteq$" symbol expresses a subclass relationship between classes. So axioms 1-3 give us a class hierarchy for propositional content in SAM and tell us that in this model, propositions and symbol structures are abstract things, and conjunctions are a kind of proposition.

Data content can be modeled at a per-datum level of granularity, or as the propositional content of an entire data set, including data where part or all of two or more data sets have been combined together. In order to represent the collective content of datasets, we use the relations *conjunctOf* and *hasConjunct*.

A symbol structure is an abstract arrangement that contingently expresses propositional content in the context of an assertion event. Examples include familiar data structures, such as strings and floating point numbers, but also abstract sets, vector spaces, natural language sentences, and the arrangement of rows and columns in a printed table.

Events

The Systematic Assertion Model focuses on key provenance events in the creation and recording of scientific data. In particular, attention is on events such as a scientist indicating a proposition on the basis of an observation, or a scientist using some symbol structure to express that propositional content. SAM has three core event classes to represent these provenance events in the life-cycle of scientific data: $Observation$, $Computation$, and $Assertion$.

$$Observation \sqsubseteq Event \tag{4}$$

$$Computation \sqsubseteq Event \tag{5}$$

$$Assertion \sqsubseteq Event \tag{6}$$

Observations are individual events, during which cognitive agents attend to directly observable objects (e.g., the moon) or phenomena (e.g., the evaporation of a quantity of water). Computations are individual events in which an agent executes an effective procedure or algorithm. Computations act upon symbol structures, and typically yield a different symbol structure as output.

Assertions are individual events in which an agent advances a claim, either by means of an ephemeral utterance or expressed in more durable media. The primary expressive form need not be a natural language sentence, but could be any abstract arrangement of symbols. Such symbol structures are often linked in contingent representation relationships governed by interpretive conventions (Dubin et al., 2011).

Claims, Systematic Assertions, and Data Content

The final set of axioms use the core classes that define events, propositional content and symbol structures to define data and data content.

$$Claim \equiv Proposition \sqcap \exists\, substanceOf.Assertion \tag{7}$$

The $substanceOf$ relation stands between an assertion event and some proposition that is the substance of the assertion. Axiom 7 uses the "existential restriction" form for description logic axioms, and states that some thing x is a claim if and only if there exists some assertion event and x is the substance of that assertion. So, in SAM, claims are propositional content that are the substance of assertions.

$$SystAssertion \equiv Assertion \sqcap \exists\, warrantedBy.(Observation \sqcup Computation) \tag{8}$$

The $warrantedBy$ relation stands between an assertion and some evidence that justifies the assertion. Systematic assertions are assertions where the asserting agent appeals for justification to an observation or computation event. Evidence for such an appeal might take the form of instrument readings recorded and time-stamped in a lab notebook.

$$(Proposition \sqcap \exists\, substanceOf.SystAssertion) \sqsubseteq DataContent \tag{9}$$

$$(Proposition \sqcap \exists\, conjunctOf.DataContent) \sqsubseteq DataContent \tag{10}$$

Data content are propositions that are the substance of systematic assertions. That is, they are the propositions expressed during assertions warranted by observations or computations. A proposition is also considered data content if it is a conjunct of a complex proposition that is data content.

So we can view the data content of a dataset as a whole, or at the granularity of individual facts.

$$Data \equiv SymbolStructure \sqcap \exists primaryExpressionFor.SystAssertion \quad (11)$$

Data are the symbol structures that are the primary form of expression for a systematic assertion event. A primary symbol structure is a symbol structure that stands in a *primary Expression For* relationship with a systematic assertion, and therefore expresses data content. It is encoded by symbol structures at lower levels in a representional stack, but does not encode any other symbol structure.

The *Mola mola* record

We illustrate how SAM can be used to describe data by giving an RDF account of a portion of the *Mola mola* species record, focusing on only one attribute/value pair: the species identification. The data content for this pairing is represented by an abstract proposition that we call *ex:speciesID*. The first line below is an RDF statement (*s1*) expressing that proposition, using the RDF *type* property to express the fact that the specimen belongs to a particular species. The identifier *kui:32596.0* denotes the specimen, and *eunis:124279* is a species identifier from the namespace of the European Nature Information System[5]. The next clause below describes s1, giving the RDF subject, predicate and object, in addition to noting that *s1* expresses the species identification. The final clause states that the species identification is a proposition according to SAM, is expressed by *s1*, and is a conjunct of the entire propositional content of the record.

```
kui:32586.0 rdf:type eunis:124279 .

_:s1 a rdf:Statement;
     rdf:subject kui:32586.0 ;
     rdf:predicate rdf:type;
     rdf:object eunis:124279
     sam:expresses ex:speciesID

ex:speciesID a sam:Proposition;
   sam:expressedBy _:s1 ;
   sam:conjunctOf ex:recordContent .
```

The identification of the specimen as a member of the species *Mola mola* is part of the conjunctive content of the entire Darwin Core occurrence record. The *Conjunction* class in SAM is a subclass of *Proposition*, and the Darwin Core occurrence record itself is expressed below in abbreviated form as the named graph *Desc1*. Although literal reading of the record suggests that each metadata statement describes the same Darwin Core *Occurrence* instance, we understand the record to express a conjunction of different propositions concerning the specimen, the collection event, the collection record, etc.

```
ex:recordContent a sam:Conjunction ;
   sam:substanceOf ex:kuiRecordAssert ;
   sam:expressedBy ex:Desc1 ;
   sam:hasConjunct ex:speciesID .
```

[5] http://eunis.eea.europa.eu

```
ex:Desc1 = {ex:id1821 a dwc:Occurrence ;
   dwc:minimumDepthInMeters "31" ;
   dwc:year "1965" ;
   dwc:scientificName "Mola mola" ;
   dwc:collectionCode "KUI" ;
 [...]
   dwc:identifiedBy "Wiley, Martin" ;
   dwc:catalogNumber "32586" ;
   dwc:continent "Atlantic Ocean" ;
   dwc:verbatimEventDate "1/8/65" ;
   dwc:verbatimLatitude "34.1217 N" ;
   dwc:fieldNumber "MLW 34" ;} .
```

The TRiG[6] serialization shown here is a different syntax than the delimited occurrence record described earlier. But we understand the primary expressive form of this record to be an abstract DCAM description (which is to say, a graph from the RDF perspective[7]). So any serialization of this graph would be an encoding of precisely the same data. We model the KUBI recording of this data as an *assertion* of the conjunctive record content, with the DCAM description (*Desc1*) serving as the primary expression:

```
ex:kuiRecordAssert a sam:Assertion ;
   sam:hasSubstance ex:recordContent ;
   sam:warrantedBy ex:mlwObserv ;
   sam:hasPrimaryExpression ex:Desc1;
   event:agent "KU Biodiversity Institute" .
```

This assertion event is the key node in this example for two reasons. First, one can only interpret the meaning of data symbols in the context of a particular expressive event. In some contexts, for example, "scientificName: *Mola mola*" would be understood as connecting a name to a species, rather than a species to a particular specimen. Second, in our model it is the warrant or justification for an assertion that determines what symbols are data and what propositions are data content. In this example, the assertion is justified on the basis of observational evidence: the *warrantedBy* relation connects this assertion to an observation record (*ex:mlwObserv*) representing Martin Wiley's observations of the specimen on January 8, 1965:

```
ex:mlwObserv a sam:Observation ;
   sam:warrants ex:kuiRecordAssert ;
   event:agent "Wiley, Martin L." ;
   event:time "1965-01-08"^^xsd:date .
```

It is through this appeal to observational evidence that the Darwin Core description gains the status of data (since it is an assertion that is warranted by an observation) and that the record content and its conjuncts are understood as data content.

DISCUSSION AND FUTURE WORK

The analysis of the example in the previous section highlights several points about the representation of digitally–encoded scientific data. In particular, the Systematic Assertion Model explicitly represents how the same data content can be expressed using different symbol structures. The original field notes and the Darwin Core triples are different primary expressions of the same data content. In SAM, data are symbol structures, and so different symbols are different data. But two distinct symbol structures (such as the Darwin Core triples and the statements in the original field

[6] http://www.wiwiss.fu-berlin.de/suhl/bizer/TriG/Spec/

[7] http://dublincore.org/document/dc-rdf/

notes) can, in appropriate contexts, mean exactly the same thing. Also, one might encode the Darwin Core triples in any number of encodings or serializations.

The expressive relationship between a symbol structure and the expressed content is not an essential property of the symbol structure itself, but is contingent on the intentions of the asserting agent and the interpretive conventions shared within a scientific community (Dubin et al., 2011). In the current example, the strings "scientificName" and "Mola mola" are used to advance a claim about the species of a particular specimen, but in a different context that same attribute/value pair might only express a fact about the name of that species.

SAM's definition of data also distinguishes information that is essential to scientific identity from information that is auxiliary. For example, in Table 1, the record number 1821 may be important enough to preserve across migrations and transformations of the data, but the number itself was assigned by fiat, and isn't the kind of fact that requires observation or calculation to justify it, and therefore we do not consider it to be expressing data content.

Extending the analysis to other digital object types

According to the Basic Representation Model, three kinds of things are involved in the representation of digital objects: *Propositional Content*, *Symbol Structure*, and *Patterned Matter and Energy*. The specific features that characterize digitally–encoded scientific data are all properties that things of these types acquire when they participate in events described by the Systematic Assertion Model. These events — assertions, observations, computations, etc. — are related to empirical scientific investigation and support a detailed account of the means by which content and symbol structures participate in the representation of a specific type of digital objects, namely digitally–encoded data.

However the Basic Representation Model itself is not limited to scientific data, but applies to other kinds of information–bearing digital objects as well. The contingent relationships between content, symbol structures, and physical objects can be instantiated in virtue of other indication events for other digital object types (e.g. textual documents).

While it is plausible to understand all digital objects as expressing some sort of content, that content may not always be strictly propositional in nature — the content of a digital image is most likely to be characterized as a set of features a person can experience looking at the actual rendered image (Sacchi, Wickett, Renear, & Dubin, 2011b). Extending the Basic Representation Model to include digital objects with non-propositional content remains a challenging open problem.

Data, Metadata, and Datasets

Being *data* is not, strictly speaking, the same as being a *dataset*. In the analysis of our example and in a recent poster (Wickett, Thomer, Sacchi, Baker, & Dubin, 2012) we made a distinction between elements of a symbol structure that express *data content* — those elements that are *data* — and elements that express contextual information about the data — e.g., information about the collection event, the scientific method used, or the record itself. For example, as mentioned above, the record number 1821 is information about the record in the context of the Darwin Core archive, but it does not express data content. Since the elements that express contextual information function as a sort of embedded metadata, the symbol structure that is our record is therefore not just data. It can be seen as including both data and metadata. This common feature can be observed in a wide variety of scenarios where data are involved: observation results are recorded and information is added to facilitate their use, reuse and meaningful interpretation. We use the term "datasets" for symbol structures that express data content together with, in many cases, auxiliary information.

This distinction between the components of a symbol structure that express data content and those that express auxiliary information, suggests a revision and integration of the role–based Conceptual Model for Datasets we presented in the last ASIS&T Annual Meeting. In that context the model served as a framework to support the identification of significant properties of scientific datasets (Sacchi et al., 2011a). A dataset was defined as "the primary symbol structure for a systematic assertion, i.e., an assertion justified by observation or computation. For the Intended Community it [1] expresses Dataset Content, and [2] supports operations appropriate to its Dataset Type". However, as we see from the analysis above, not everything included in a dataset can be identified strictly as data under the Systematic Assertion Model.

Assuming that a symbol structure is composed of discrete elements, we can identify the set of elements that we consider *data* as a subset of the elements that constitute the entire dataset. In our example, we can identify the Darwin Core Archive, which contains the occurrence records as well as information about the provenance of the records and how to interpret the fields used in occurrence records, as a dataset. At the level of the primary symbol structure, this dataset is a set of triples expressing all of that information according to the conventions established by the Darwin Core Task Group. The set of triples expressing *data content* and the set of triples expressing metadata information are both subsets of the dataset, which is the complete set of triples.

In our case, data and metadata components are interwoven: no structural distinction allows an immediate discrimination between data and metadata. However, depending on the data representation language in use, there could be a dedicated structural section, like a header, for the components that are explicitly intended to express metadata information.

In terms of *propositional content* we can make a parallel observation: data content is a subset of the entire conjunctive propositional content of a dataset. The propositional content of a dataset can also be composed of auxiliary information that is not *data content*, like contextual, descriptive, or other information expressed by metadata attribute-value pairs.

Identity and Scientific Equivalence

One of the goals of our work is to develop models that let us give precise accounts of information objects carrying "the same data". Handling this kind of identity question is an obvious issue for preservation, conversion and reuse of scientific data, and has received recent attention in the Earth sciences domain (Tilmes et al., 2011). A recent analysis of identifier schemes in that domain presents the problem directly in a use case focused on the "ability to tell that two data instances contain the same information even if their formats are different" (Duerr et al., 2011). The approach taken within this community attempts to avoid the problems of identity by instead developing the notion of *scientific equivalence*.

Tilmes, et al. (2011) characterize two digital objects (such as files) as scientifically equivalent if they are "sufficiently similar that their use in a scientific investigation would result in the same results or conclusions." This definition is suggestive and something along these lines is no doubt true. However, as the sort of account that could be part of a formal conceptual framework for dataset concepts it is problematic.

For one thing the definition takes the form of a subjunctive conditional ("their use in a scientific investigation would result..."). The semantics of subjunctive, or contrary to fact, conditionals has proven challenging and it is unclear how to represent them in standard logical languages (Lewis, 1973). Furthermore, as we have argued elsewhere (Dubin et al., 2011), the propositional content that data express is not determined absolutely by the symbols used in an expression, but is always determined in reference to a set of conventions, or an *interpretive frame*, that supplies the relevant mapping between symbol structures and the intended propositional content. The generation of results and synthesis of conclusions on the basis of data will also be subject to the interpretive frames of an investigator.

This problem goes beyond the specification of interpretative frames. For instance, consider two digital objects, A and B, that, in a given context, are understood as carrying the same data, but using different encodings. Now let's suppose that the encoding in digital object A is considerably easier to manage and exploit than the encoding in digital object B. In such a situation, it is unlikely that the use of these digital objects in a scientific investigation "would result in the same results or conclusions" (Tilmes et al., 2011). Since the availability and cost of software tools has an effect on whether some encodings are easier or harder to exploit than others this definition would seem to have scientific equivalence be the kind thing that could come and go as software strategies or other techniques gained or lost prominence.

One approach that avoids the effects of contingent circumstances is to specify that two digital objects are scientifically equivalent if and only if it is *logically impossible* that they could generate different scientific results. But a definition relying on this kind of necessity will also present significant challenges. Obviously what would be intended by such a definition is that it is logically impossible for scientifically equivalent digital objects to have different scientific results *in the same circumstances*, as different circumstances (calculations, additional information, interpretative frames, etc.) will lead to different conclusions. But if we specify that the circumstances must be exactly the same it now becomes unclear whether two different digital objects, however trivial their differences, could ever be shown to be equivalent, since the conditions of processing must be different to accommodate the differences in encoding.

These may not yet be decisive objections as presented here, but they do show that the subjunctive definition of scientific equivalence requires significant refinement before it will provide an account that is adequate for resolving identity problems for scientific data.

The Basic Representation Model and SAM support a more fine-grained approach to the equivalences that might hold between the objects that carry scientific data. In the case of the example *Mola mola* record, we are able to say that the various digital objects (the tab-delimited text record, and the XML record) generated from the Darwin Core Archive are encodings of the same *data*. Since these objects encode the same data and that data expresses the data content pertaining to the collection and identification of the specimen, we can also say that these objects carry the same *data content*. However, while the original field notes that were the basis of the Darwin Core record express the same data content, they are different data since they do not use the Darwin Core vocabulary.

CONCLUSION

We have shown how the Basic Representation Model and the Systematic Assertion Model together provide the missing account of the entities and relationships involved in the creation and representation of scientific data. The Basic Representation Model accounts for layered encodings of digital information objects, connecting agents' expressions of content to the final inscription in a physical storage device. This model is in some respects similar to the Functional Requirements for Bibliographic Records in its separation of intellectual content and embodying structures, but is more flexible and designed to accommodate the curation of datasets and other resources in digital environments. The Systematic Assertion Model draws attention to the core provenance events in the recording of scientific data, thereby placing focus on essential details in the context of the creation of those data. These models can inform the design of systems for the curation, preservation, and sharing of scientific data, and guide the development of cross–domain metadata vocabularies.

ACKNOWLEDGMENTS

The research reported here is being carried out at the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois at Urbana-Champaign. It is funded by the National Science Foundation as part of the Data Conservancy, a multi–institutional NSF funded project (OCI/ITR-DataNet 0830976) hosted at Johns Hopkins University Sheridan Libraries. This work reflects discussions with members of the Data Practices group at CIRSS, in particular Karen

Baker and Andrea Thomer. Assistance with the example record was provided by Laura Russell at VertNet.

References

Baader, F., & Nutt, W. (2003). Basic description logics. In F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, & P. F. Patel-Schneider (Eds.), *Description logic handbook* (p. 43-95). New York: Cambridge University Press.

CCSDS. (2002). *Reference model for an open archival information system (OAIS)* (Tech. Rep.). CCSDS 650.0-B-1, Blue Book.

Darwin Core Task Group. (2009, October). *Darwin core.* Published on the World Wide Web at `http://rs.tdwg.org/dwc/`.

Dubin, D. (2010, October). Encoded descriptions at face value. In A. Grove (Ed.), *Proceedings of the american society for information science and techonology* (Vol. 47). Pittsburgh, PA.

Dubin, D., Wickett, K. M., & Sacchi, S. (2011, August). Content, format, and interpretation. In B. T. Usdin (Ed.), *Proceedings of balisage: the markup conference 2011* (Vol. 7). Montréal, Canada.

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W., Glassy, J., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, *4*, 139-160. Available from `http://dx.doi.org/10.1007/s12145-011-0083-6` (10.1007/s12145-011-0083-6)

Farquhar, A., & Hockx-Yu, H. (2008). Planets: Integrated services for digital preservation. *Serials: The Journal for the Serials Community*, *21*(2), 140–145.

Guarino, N., & Welty, C. A. (2000). A formal ontology of properties. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management* (pp. 97–112). London, UK: Springer-Verlag.

Hourclé, J. A. (2008). FRBR applied to scientific data. In *Proceedings of the american society for information science and techonology* (Vol. 45, pp. 1–4). Available from `http://dx.doi.org/10.1002/meet.2008.14504503102`

IFLA. (2009). *Functional requirements for bibliographic records: Final report* (Tech. Rep.). International Federation of Library Associations and Institutions. Available from `http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf`

Lefort, L., Henson, C., Taylor, K., Barnaghi, P., Compton, M., Corcho, O., et al. (2011, June). *Semantic sensor network XG final report, W3C incubator group report (2011)* (Tech. Rep.). W3C. Available from `http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/`

Lewis, D. (1973). *Counterfactuals.* Oxford: Basil Blackwell.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, *2*(3), 279 - 296.

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011, October). The analytic potential of scientific data: Understanding re-use value. In A. Grove (Ed.), *Proceedings of ASIS&T 2011: the 74rd annual meeting of the american society for information science and technology* (Vol. 48). Silver Spring, MD.

Renear, A. H., & Dubin, D. (2003). Towards identity conditions for digital documents. In *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications* (p. 19).

Renear, A. H., & Dubin, D. (2007). Three of the four FRBR group 1 entity types are roles, not types. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–19.

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010, October). Definitions of dataset in the scientific and technical literature. In A. Grove (Ed.), *Proceedings of the american society for information science and techonology* (Vol. 47). Silver Spring, MD.

Sacchi, S., Wickett, K. M., Renear, A. H., & Dubin, D. (2011b). One thing is missing or two things are confused: An analysis of OAIS Representation Information. *Poster presented at the Seventh Internation Digital Curation Conference*.

Sacchi, S., Wickett, K. M., Renear, A. H., & Dubin, D. S. (2011a). A framework for applying the concept of significant properties to datasets. In *Proceedings of the american society for information science and techonology.* New Orleans, LA.

Sandore, B., & Unsworth, J. (2010, June). ECHO DEPository — phase 2: 2008–2010 final report of project activities [section]. In (pp. 30–37). University of Illinois at Urbana-Champaign.

Tilmes, C., Yesha, Y., & Halem, M. (2010). Tracking provenance of earth science data. *Earth Science Informatics*, *3*(1-2), 59-65.

Tilmes, C., Yesha, Y., & Halem, M. (2011). Distinguishing provenance equivalence of earth science data. *Procedia Computer Science*, *4*, 548-557.

Wickett, K. M., Thomer, A., Sacchi, S., Baker, K. S., & Dubin, D. (2012). What dataset descriptions actually describe: Using the systematic assertion model to connect theory and practice. *Poster presented at Third Annual Research Data Access and Preservation Summit*.