Identifying Content and Levels of Representation in Scientific Data

Karen Wickett, Simone Sacchi, David Dubin, Allen 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

October 29, 2012

Did you come to the right session?

Did you come to the right session?

- The session on conceptualizing data and identity is taking place now in the Johnson Room.

Did you come to the right session?

- The session on conceptualizing data and identity is taking place now in the Johnson Room.
- But you came here for papers on using information analytics.

Did you come to the right session?

- The session on conceptualizing data and identity is taking place now in the Johnson Room.
- But you came here for papers on using information analytics.
- Too bad: this talk is on conceptualizing data and identity.

Did you come to the right session?

- The session on conceptualizing data and identity is taking place now in the Johnson Room.
- But you came here for papers on using information analytics.
- Too bad: this talk is on conceptualizing data and identity.
- For a great paper on using information analytics, see Weber et al. "Value and Context in Data Use."

Did you come to the right session?

- The session on conceptualizing data and identity is taking place now in the Johnson Room.
- But you came here for papers on using information analytics.
- Too bad: this talk is on conceptualizing data and identity.
- For a great paper on using information analytics, see Weber et al. "Value and Context in Data Use."
- Nic just finished presenting that paper in the Johnson Room.

Background: the Data Conservancy

The *Data Conservancy*, hosted at Johns Hopkins University Sheridan Libraries, is a multi-institutional project funded under the NSF DataNet program. 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.

Problems we're addressing

These are hard:

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.
2. Interpreting the structure of digital data for integration and reuse.

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.
2. Interpreting the structure of digital data for integration and reuse.
3. Tracking the provenance of digital data for trust and verification.

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.
2. Interpreting the structure of digital data for integration and reuse.
3. Tracking the provenance of digital data for trust and verification.
4. Acknowledging and crediting the contributions to creating a data set.

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.
2. Interpreting the structure of digital data for integration and reuse.
3. Tracking the provenance of digital data for trust and verification.
4. Acknowledging and crediting the contributions to creating a data set.
5. Balancing local encoding control with standardization for reuse.

Problems we're addressing

These are hard:

1. Establishing that two different digital resources encode the same scientific data.
2. Interpreting the structure of digital data for integration and reuse.
3. Tracking the provenance of digital data for trust and verification.
4. Acknowledging and crediting the contributions to creating a data set.
5. Balancing local encoding control with standardization for reuse.
6. Understanding what data really is.

Our contribution

The problems on the last slide are hard:

Our contribution

The problems on the last slide are hard:

- We don't have a magic technology that solves those problems.

Our contribution

The problems on the last slide are hard:

- We don't have a magic technology that solves those problems.
- We're not pushing a specific metadata format or standard notation.

Our contribution

The problems on the last slide are hard:

- We don't have a magic technology that solves those problems.
- We're not pushing a specific metadata format or standard notation.
- We're offering a logical framework that could serve as a basis for metadata languages.

Our contribution

The problems on the last slide are hard:

- We don't have a magic technology that solves those problems.
- We're not pushing a specific metadata format or standard notation.
- We're offering a logical framework that could serve as a basis for metadata languages.
- Descriptions conforming to our model contain information that software tools can use to solve problems.

Good practice

Some things we think you'll agree are just good practice:

Good practice

Some things we think you'll agree are just good practice:

1. Documenting your scientific data file format.

Good practice

Some things we think you'll agree are just good practice:

1. Documenting your scientific data file format.
2. Keeping logs of data transformations.

Good practice

Some things we think you'll agree are just good practice:

1. Documenting your scientific data file format.
2. Keeping logs of data transformations.
3. Citing evidence that justifies belief in the data.

Good practice

Some things we think you'll agree are just good practice:

1. Documenting your scientific data file format.
2. Keeping logs of data transformations.
3. Citing evidence that justifies belief in the data.
4. Taking responsibility for your interpretation of that evidence.

Good practice

Some things we think you'll agree are just good practice:

1. Documenting your scientific data file format.
2. Keeping logs of data transformations.
3. Citing evidence that justifies belief in the data.
4. Taking responsibility for your interpretation of that evidence.

Our models use all those things in a theory of what it means to be data, and how we understand different senses of "the same data."

Understanding our models

Understanding our models

- Data content, in our models, is propositional.

Understanding our models

- Data content, in our models, is propositional.
- Propositions are the language independent bearers of truth values and the objects of epistemic attitudes.

Understanding our models

- Data content, in our models, is propositional.
- Propositions are the language independent bearers of truth values and the objects of epistemic attitudes.
- Propositions can stand in conjunctive relations with each other, so the content of a single spreadsheet cell or an entire data set is a proposition.

Understanding our models

- Data content, in our models, is propositional.
- Propositions are the language independent bearers of truth values and the objects of epistemic attitudes.
- Propositions can stand in conjunctive relations with each other, so the content of a single spreadsheet cell or an entire data set is a proposition.
- Our Basic Representation Model (BRM) defines an encoding stack from propositions, through layers of symbolic expression, to physical inscriptions.

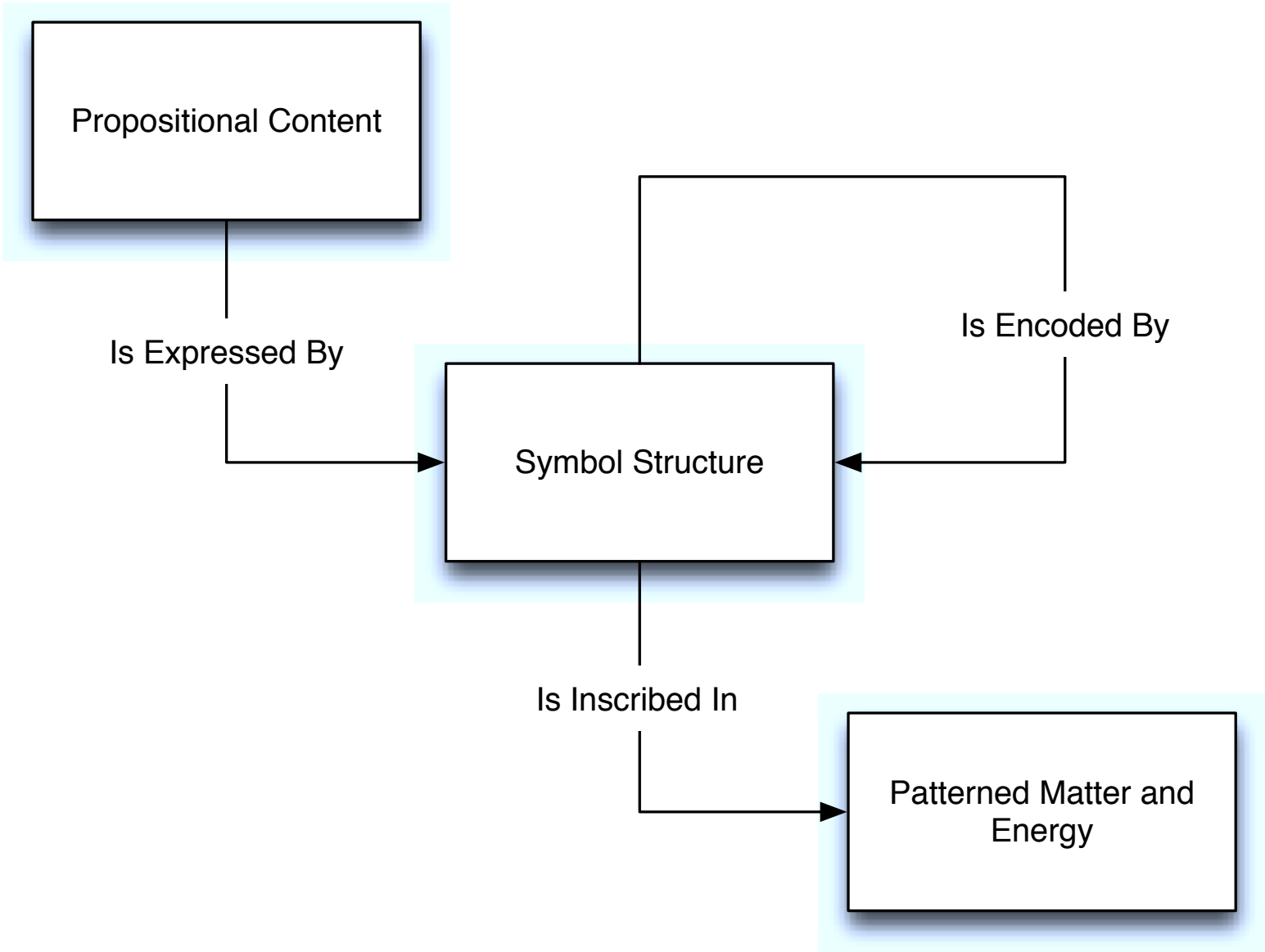
Understanding our models

- Data content, in our models, is propositional.
- Propositions are the language independent bearers of truth values and the objects of epistemic attitudes.
- Propositions can stand in conjunctive relations with each other, so the content of a single spreadsheet cell or an entire data set is a proposition.
- Our Basic Representation Model (BRM) defines an encoding stack from propositions, through layers of symbolic expression, to physical inscriptions.
- The highest level of encoding is understood as the *primary expression* (a simplification).

Understanding our models

- Data content, in our models, is propositional.
- Propositions are the language independent bearers of truth values and the objects of epistemic attitudes.
- Propositions can stand in conjunctive relations with each other, so the content of a single spreadsheet cell or an entire data set is a proposition.
- Our Basic Representation Model (BRM) defines an encoding stack from propositions, through layers of symbolic expression, to physical inscriptions.
- The highest level of encoding is understood as the *primary expression* (a simplification).
- So these are similar to (but not exactly like) FRBR Group 1 entities.

The Basic Representation Model



The Systematic Assertion Model

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.
- This happens through their participation in provenance events.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.
- This happens through their participation in provenance events.
- Observation and computation are two broad superclasses for events in the conduct of science.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.
- This happens through their participation in provenance events.
- Observation and computation are two broad superclasses for events in the conduct of science.
- Assertions are linguistic acts, bringing illocutionary force to a proposition.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.
- This happens through their participation in provenance events.
- Observation and computation are two broad superclasses for events in the conduct of science.
- Assertions are linguistic acts, bringing illocutionary force to a proposition.
- Systematic assertions appeal for justification to observations or computations.

The Systematic Assertion Model

- The BRM defines context-free relationships between content, symbols, and physical media.
- The Systematic Assertion Model (SAM) defines how some symbol structures become data.
- This happens through their participation in provenance events.
- Observation and computation are two broad superclasses for events in the conduct of science.
- Assertions are linguistic acts, bringing illocutionary force to a proposition.
- Systematic assertions appeal for justification to observations or computations.
- Data content are the substance of systematic assertions, and data are their primary expressions.

A DCAM Description as a named graph

```
ex:recordContent a sam:Conjunction ;
    sam:substanceOf ex:kuiRecordAssert ;
    brm:isExpressedBy ex:Desc1 ;

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" ;} .
```

SAM-based Metadata

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

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

Caveats and Clarifications

Caveats and Clarifications

- Remember that the RDF vocabulary used in the slides and paper is just an example.

Caveats and Clarifications

- Remember that the RDF vocabulary used in the slides and paper is just an example.
- What's important is the information captured and how one understands it.

Caveats and Clarifications

- Remember that the RDF vocabulary used in the slides and paper is just an example.
- What's important is the information captured and how one understands it.
- A more representative expression of the Darwin Core record would connect with the abstract graph via an *interpretive frame*.

Caveats and Clarifications

- Remember that the RDF vocabulary used in the slides and paper is just an example.
- What's important is the information captured and how one understands it.
- A more representative expression of the Darwin Core record would connect with the abstract graph via an *interpretive frame*.
- Different senses of scientific equivalence amount to strict identity at different levels of the model.

Caveats and Clarifications

- Remember that the RDF vocabulary used in the slides and paper is just an example.
- What's important is the information captured and how one understands it.
- A more representative expression of the Darwin Core record would connect with the abstract graph via an *interpretive frame*.
- Different senses of scientific equivalence amount to strict identity at different levels of the model.
- For example, same data content, same primary symbol structure, or same justification.