FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign

DATA CONCEPTS

⑥

WHAT IS DATA?

What is data?

Now we can finally answer this question

We give an answer based on our ontology.

And we note that data is a *role*, and so it is *relative* in two senses:

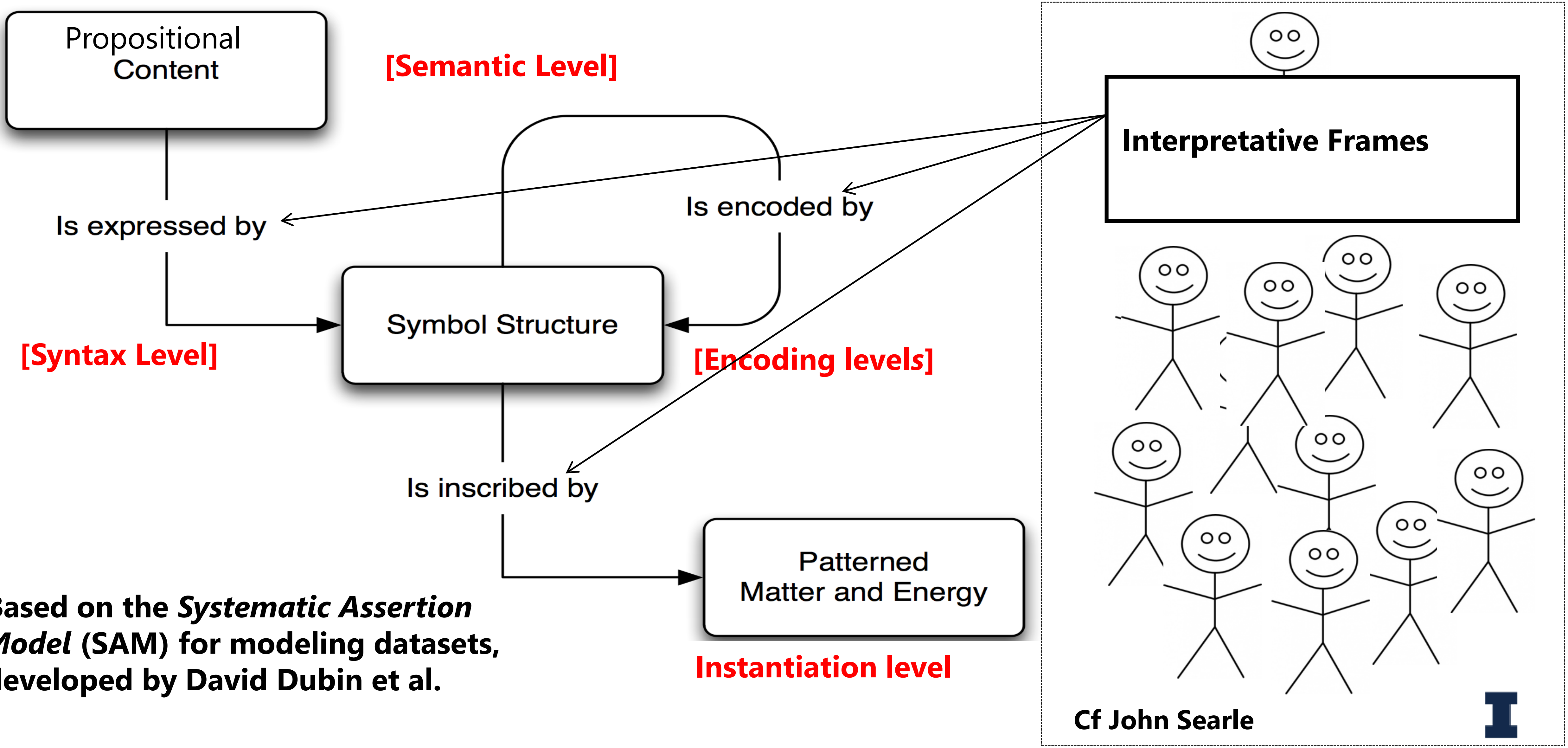
-- nothing is data intrinsically, only with respect to use

-- the same propositions can be data in one circumstance,
and a claim supported by data in another

And it is routinely the case in science that

one person’s data is another person’s theory

[Recall:] our final slide from before



Based on the *Systematic Assertion Model* (SAM) for modeling datasets, developed by David Dubin et al.

Data, our definition

So our answer to the vexed question “What is data?” is:

Data are **propositions**

(i) [systematically] *asserted*. . .

(ii) as *evidence*

Dubin et al. 2009-2014

Propositions?

By illustration, and very roughly:

Two declarative sentences that say the same thing *express the same proposition*.

Asserted?

In order to be *asserted* **propositions** must be

expressed in a **language**

encoded in **symbols**

inscribed in **material form**

So we see that human *intentionality* is fundamental to the idea of data:

Why? Because *expressing, encoding,* and *inscribing* are not things that just happen naturally.

They require conventions created and maintained by *communities of persons,* as well as particular *beliefs, intentions, and practices* based on those conventions.

Evidence?

What is it for proposition(s) to be evidence for something?

We dodge this epistemological question

[but we know it when we see it, right?]

Is being data a *type* or a *role*?

All data consists of propositions

But not all propositions are data

Only propositions that are asserted are data

And not asserted propositions are data either

Only those intended to serve as evidence

Data is not a type of thing, it is a role

Just as persons are students when enrolled in a school

propositions are data when asserted as evidence

Being asserted as evidence is contingent (and social) circumstance

And so data is role that propositions have
in certain contingent social circumstances

What kind of thing is data evidence for?

We think of data as being evidence for things such as

theories, hypotheses, conjectures, *claims*, *assertions* . . .

Let’s call all of those things *claims*.[*]

But what kind of thing is a claim?

Both data and claims appear to be (ontologically)
the same kind of thing: *propositions in a role*

Data are propositions in the role of being asserted as evidence
Claims are propositions in the role of being asserted as supported by evidence

*In order to be considered data data need not be believed sufficient to *confirm* or *justify* the claim it is offered as evidence for. And at least arguably it need not even be believed to increase likelihood of those claims: being *asserted as evidence* could be taken as meaning *being considered as potentially evidence,* this would allow the assertion to be agnostic with respect to normative evidentiary weight.

Data and claims, both are roles

So what makes, in some scenario,
some propositions data (evidence) and other propositions claims?

Only the actions and intentions of a particular person or persons

There is nothing intrinsic to data that makes data data,
and nothing intrinsic to claims that makes claims claims.

Being data is a <u>role</u> that some propositions
have in certain contingent social circumstances.

And the same for claims.

Data is *relative*

So whether propositions are data or claims depends upon what is intended.
And propositions can be data in one circumstance, claims in another.

In fact, science as a whole depends on this. For instance:

For a climate scientist,

growth rings on tree rounds may be evidence
for theories about temperature changes

But for an evolutionary ecologist

those theories about temperature changes may in turn be evidence
for theories about competitive advantages

In a slogan:
one person’s data is another person’s theory

Some relevant publications

The SAM Model for Datasets

“Definitions of Dataset in the Scientific and Technical Literature.” *Proceedings of the 73rd Annual Meeting of the American Society for Information in Science and Scholarship*, Renear, Allen H., Simone Sacchi, Karen M Wickett (2010).

“A Framework for Applying the Concept of Significant Properties to Datasets”. *Proceedings of the 74th Annual Meeting of the American Society for Information in Science and Scholarship*, Sacchi, Simone, Karen M Wickett, David Dubin, Allen H. Renear (2011).

Content, format, and Interpretation*, Proceedings of Balisage: The Markup Conference,* Dubin, David, Wickett Karen, Sacchi, Simone (2011).

“Identifying Content and Levels of Representation in Scientific Data.” *Proceedings of the 75th Annual Meeting of the American Society for Information in Science and Scholarship,* Wickett Karen M, Simone Sacchi, David Dubin Allen Renear. (2012)

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.