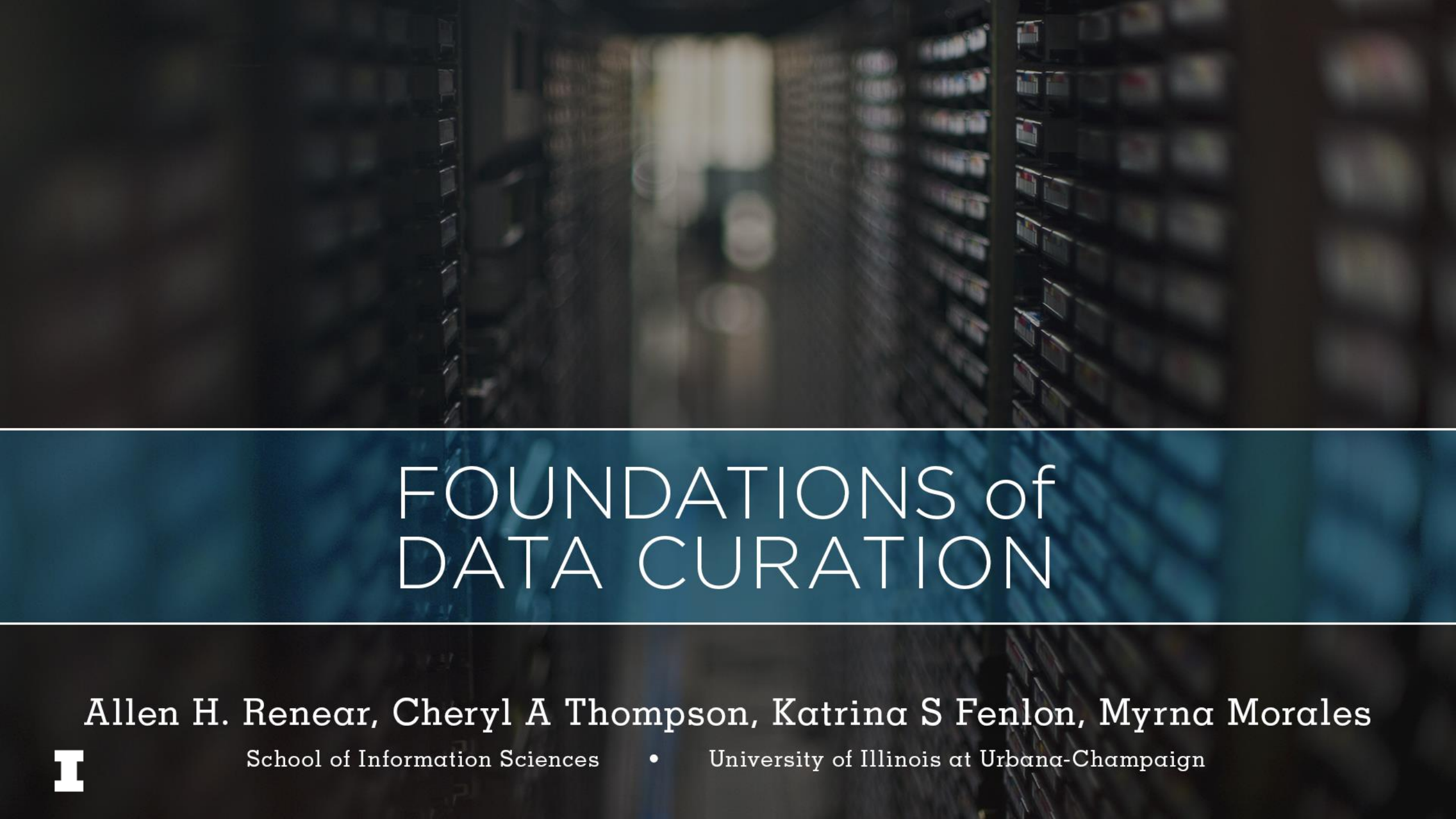


FOUNDATIONS of DATA CURATION

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

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


DATA
CONCEPTS

1. What Is Data? A First Attempt.

Some dictionary definitions.

These are useful, but not formal enough, and conceal problems.

An empirical approach to the question.

This yields some additional insights,

but no rigor, and plenty of variation, and inconsistencies.

Scientists still insist formal definitions are needed.

So we aren’t going to give up.

The Question

What is data?

???

No, we want to know *what* is data, not *who* is Data.

What we want for an answer

What we are after here is a *formal definition* of data.

And one that is part of a *conceptual model (or ontology) of data concepts*.

This is not just of theoretical interest.

> A conceptual model for data concepts would provide a rigorous formal foundation for the design of systems supporting all aspects of data curation.

Data, some lexicographical definitions

information, especially information organized for analysis
-- American Heritage Dictionary

factual information (such as measurements and statistics) used as a basis for reasoning, discussion, or calculation.
-- Merriam Webster Dictionary

a collection of facts from which conclusions may be drawn
-- Wordnet (Princeton)

a collection of observations . . .
-- [common]

a collection of organized information, usually the result of experience, observation or experiment, … may consist of numbers, words, or images, particularly as measurements or observations …

-- State of Maryland, Department of Information Technology

Are we done?

No, these definitions useful, but they are much too casual for rigorous modeling.

What we want is a conceptual model, or ontology, for data concepts

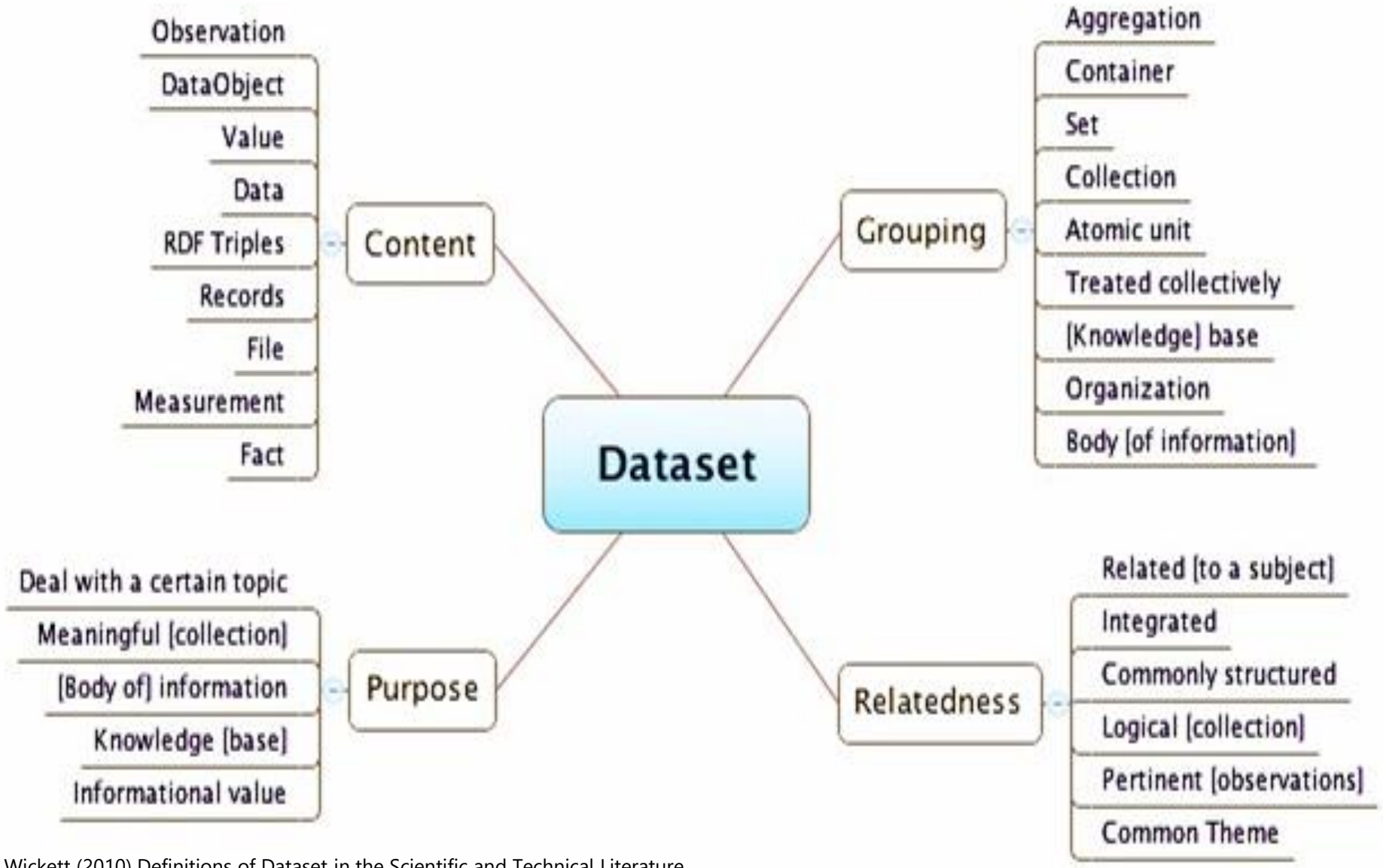
An empirical approach

Maybe we should ask a scientist

— they should have an answer, right?

[Well, actually they don’t; but let’s ask anyway. . .]

How scientists define datasets



Renear, Sacchi, Wickett (2010) Definitions of Dataset in the Scientific and Technical Literature

Fits our dictionary definitions well enough

information, especially information organized for analysis
-- American Heritage Dictionary

factual information (such as measurements and statistics) used as a basis for reasoning, discussion, or calculation.
-- Merriam Webster Dictionary

a collection of facts from which conclusions may be drawn
-- Wordnet (Princeton)

a collection of observations . . .
-- [common]

a collection of organized information, usually the result of experience, observation or experiment, … may consist of numbers, words, or images, particularly as measurements or observations …
-- State of Maryland, Department of Information Technology

Content
Grouping
Relatedness
Purpose

But is there really enough agreement for modeling?

There is too much variation here:

Purpose:	analysis, evidence, explaining, being explained
Relatedness:	same subject, same attributes, same syntax
Grouping:	set, aggregation, collection

And most challenging for modeling:

Content:	observations, values, facts, numbers, records, expressions, triples, tuples, information . . .

While we see patterns, colloquial definitions are still, from a modelling point of view, too informal, too varied, too inconsistent.

Well, maybe scientists just don't care much about definitions?

Cries from the heart. . .

"There is ambiguity in what type of object a dataset is;
with different groups of users applying different connotations

*There needs to be an explicit statement of what
the intended preservation of a dataset will imply."*

Sam Pepler, National Center for Atmospheric Science, UK

"While there has been substantial work in the IT community regarding metadata and file identifier schemas, there appears to be relatively little work on the organization of the file collections .

One symptom . . . appears in nomenclature describing collections: the terms `Data Product,' `Data Set,' and `Version' are overlaid with multiple meanings between communities."

Bruce Barkstrom, NASA


2
THE IDENTITY PROBLEM

The Identity Problem

Identity problems in data curation

Identity problems and representation levels

Identity problems in data curation

Archiving:	Is this dataset already in the archive?
Preservation:	Was the information preserved in the new file format?
Security:	Has this dataset been tampered with?
Authentication:	Is this the data we think it is?
Reproducibility:	Does this XML file have the same information as that JSON file?
Provenance:	Were these datasets derived from the same data?
Conversions:	Does the converted file have the same data as the original?

and on and on. . .

> "… there are an unknown number of transformations that are invariant in the sense of preserving the scientific meaning . . . different scientific communities use different tools that require different representations.

Ruth Duerr, National Snow and Ice Data Center

Data Conservancy wiki, December 2010

Same, different, same, different. (But same/different what?)

Consider conversions*

[DC to ISO-Bib, TEI P2 to TEI P3, mzData to mzML, JSON to XML . . . and on . .]

Some conversions are simple format changes
some involve a change in model type
some have schema integration challenges
[and some have profound heterogeneity problems]

In a successful conversion we'd probably say
"the data is the same . . . it is only in a different format, encoding, etc.

So in a successful conversion something changes;

and something remains the same.

But what exactly changes? And what remains the same?

*Transformations, transcodings, etc.

Identity problems

Two biologists, Jill and John, used the same data

What does that mean?
And how can we tell?

Compare:

Two biologists, Jill and John, used the same statistician.

Identity and representation levels

Consider two files with the same data

but relational tables in one case

and RDF triples in another

Same data, different representations

Identity and representation levels

Consider two files with

… same data and the same RDF triples,

but an XML serialization in one case,

and an N3 serialization in the other

Identity and representation levels

Consider two files

with the same data, same RDF triples, same N3 serialization,

but an ASCII character encoding in one case
vs an EBCDIC encoding in another

Identity and representation levels

How many of these levels are there ?

How do we name, define, and manage them?

How can they be identified and re-identified?

Identity conditions

So, underlying many data curation issues is the problem of identity.

Is x is the same [data, document, text, image . . .] as y?

The conceptual question: What do we mean?

The operational question: How do we tell?

3

SOME ONTOLOGICAL ANALYSIS

Some ontological analysis

We now embark on building *a conceptual model for data concepts*

FRBR seems relevant here,

so we will try to generalize FRBR to *representation in general*.

[After all, data seems to be, in part, *a representation of how things are*.]

We develop fragment of a FRBR-like conceptual model for representation

Two problems appear when we try to apply this model to data.

The two middle entity types appear to simultaneously:

1) collapse into a single entity type
2) explode into an indefinite number of entity types

Generalizing FRBR

The cascade of representations we saw in the last video
(**data, RDF triples, N3 serialization, ASCII encoding**)
seems somewhat similar to FRBR's **Work, Expression, Manifestation, Item**)

the same ***data*** can be realized in triples, relations, or a tree
just as the same ***work*** can be realized in English or French ***expressions*** ***(text)***. .
.

the same ***triples*** can be encoded in RDF/XML, N3, or turtle
just as the same ***expression*** can encoded in Helvetica or Times ***manifestation***

and so on

And a *work* is a little like data isn't it? Aren't both of them information?
And we need languages to express both works and data, right?
And those languages are realized in different ways
And eventually instantiated in the physical world.

No, the parallels aren't perfect, but it is a start.

A model for *representation in general*

In order to cover data concepts we need something more general than FRBR.

We need a conceptual model for *representation*

And for concreteness we start with linguistic representation

In the conceptual model of linguistic representation that follows there will be four key entity types

proposition	(meaning)
sentence	(a linguistic expression of a proposition)
encoding	(a representation of a sentence)
inscription	(a physical instantiation of an encoding)

The plan is to identify the ontological kind for each

and then replace the linguistic model with a more general one.

Let's examine each of these entity types in turn . . .

Propositions

Representation typically involves
the presentation of ***propositional content***

Propositions may be defined variously as:

-- the content of assertions

-- the objects of belief, doubt, etc.

-- the (proper) bearers of truth and falsity

-- the meanings of declarative sentences

For our purposes today these are all the same things

Propositions Vs Sentences

Sue: "Snow is white"
Astrid: "Schnee ist Weiss"

Sue and Astrid are using different sentences to say the same thing

So, *same* **proposition**, *different* **sentences**

Sentences Vs Encodings

Jill's paper: *Snow is white.*

Allen's exclamation: [listen while I say it out loud]

One is in a particular *writing system*
consisting of graphemes, punctuation, etc.

another in a *particular speech system*
consisting of phonemes, stress, pitch, etc.

They are using different symbols to encode the same sentence.

So, *same* **sentence**, *different* **encodings**

Encodings Vs Inscriptions

Snow is white
Snow is white
Snow is white

Same **encoding**, *different* **inscriptions**

Comparison with FRBR

So maybe,

Work	=	proposition
Expression	=	sentence
Manifestation	=	encoding
Item	=	inscription

Looks about right, no?

Now let's identify the general ontological kinds . . .

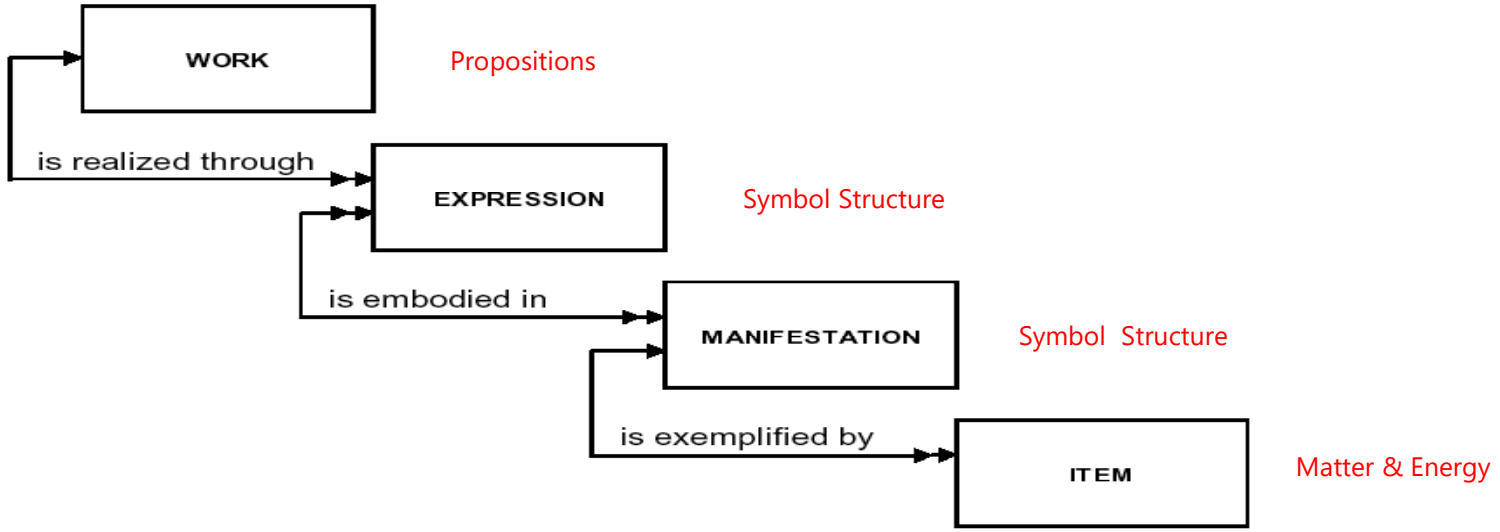
Ok, where have we got to?

How about:

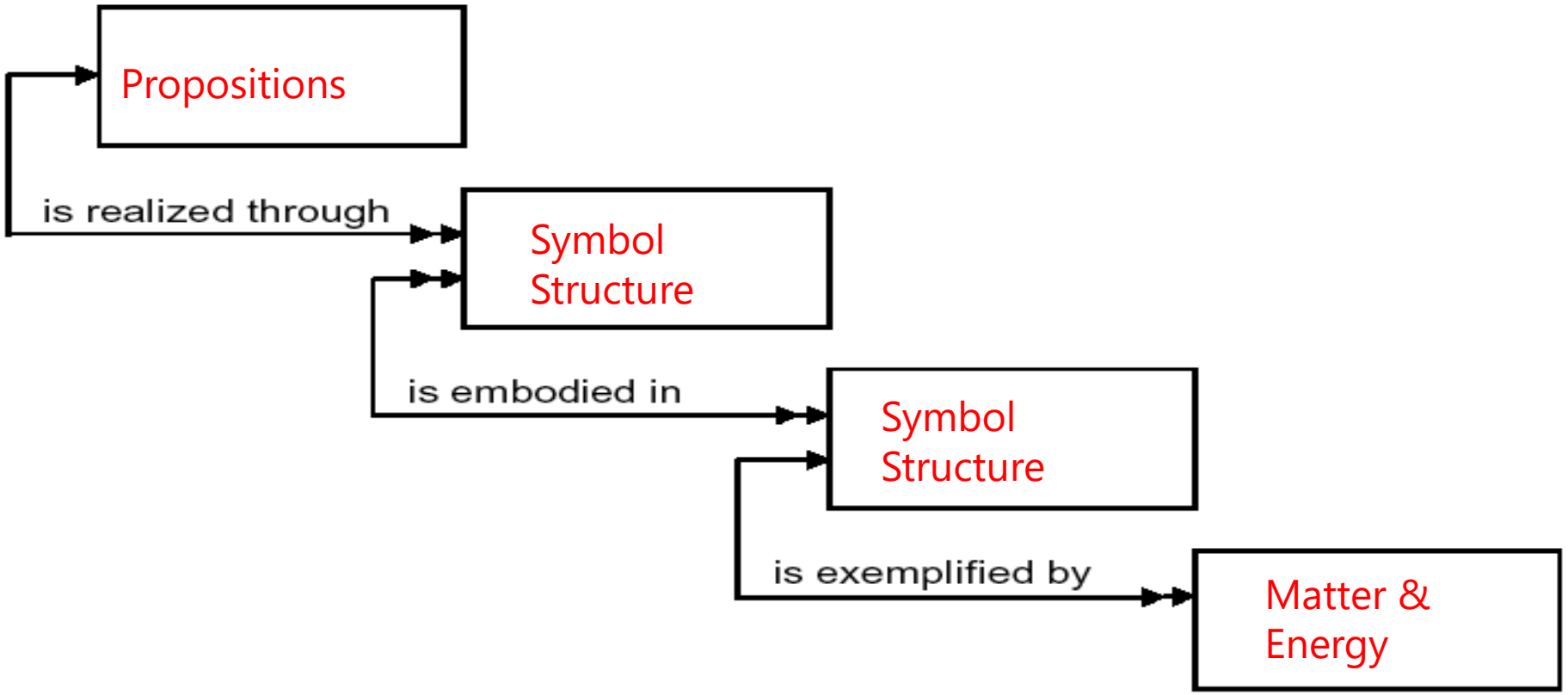
FRBR	**Linguistic Representation**	**Entity Type**
Work	proposition	proposition
Expression	sentence	?
Manifestation	encoding	?
Item	inscription	Patterned Matter & Energy

Hmm . . .

FRBR	Linguistic Representation	Entity Type
Work	proposition	proposition
Expression	sentence	?
Manifestation	encoding	?
Item	inscription	Patterned Matter & Energy



Oh Oh



*Renear & Dubin (2007)

Encodings everywhere??!!!

	Snow is white	*a proposition*
can be expressed by	"Snow is white"	*a sentence*
can be encoded by:	S,n,o,w, ,i,s, ,w,h,i,t,e	*characters*
which can be encoded by:	*Snow is white.*	*glyphs*
which can be encoded by:	**83, 110, 111**…	*integers+**
which can be encoded by:	53, 6E, 6F…	*numerals*
which can be encoded by	01010011 01101110 01101111. . .	*binary octets*

[But how many levels are there here, really? *There can be any number*!]

The situation :

1) We have an indefinite number of symbolic encodings, not just one [or two]
2) the first level seems to be similar to a FRBR expression
3) the rest seem to be either encoding an expression, or encoding an encoding (!)

*taking liberties here to make a point.

Where are we now?

We have two problems, paradoxically inconsistent.

- We need to replace the middle two entity types with one entity type, as both appear to be the same sort of thing (*symbol structures*).
- We need to replace the middle two entity types with many entity types, to accommodate the many levels of encoding.

The solution to these problems, which emerges in the next two videos, provides a powerful conceptual insights into the nature of digital representation

4

THE WAY FORWARD: ROLES AND TYPES

Types and Roles

In the last video we saw how our FRBR-inspired model for representation encountered two problems

- First, it appeared that the two middle entity types should be collapsed into a single entity type
- Second, it appeared that the two middle entity types should be expanded into an indefinite number of entity types.

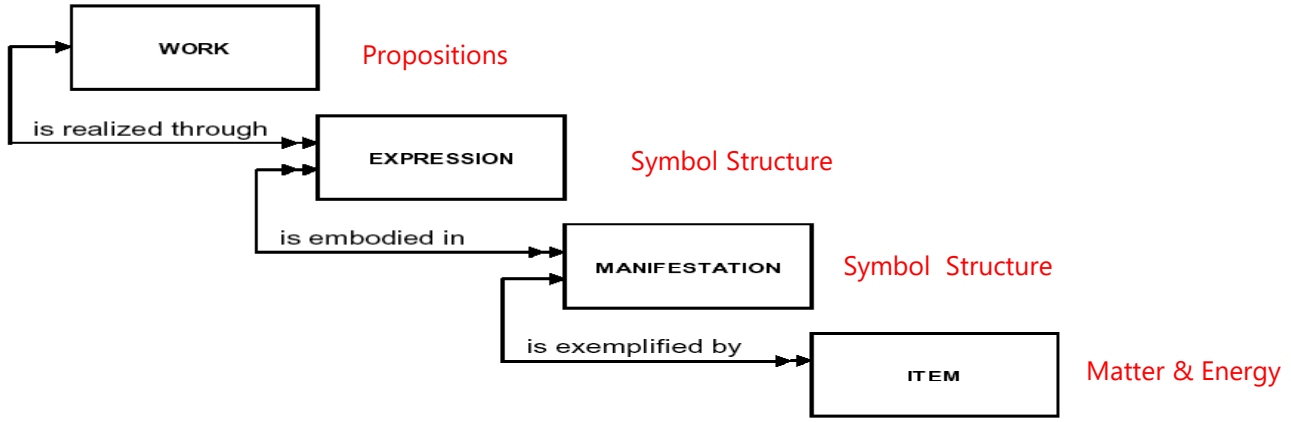
So apparently some considerable *refactoring* is required.

In this video we present the foundation for this refactoring, describing a distinction (*types* vs *roles*) that has become fundamental in ontological analysis.

The first problem: the collapse (again):

How about:

FRBR	**Linguistic Representation**	**Entity Type**
Work	proposition	Proposition
Expression	sentence	Symbol Structure
Manifestation	encoding	Symbol Structure
Item	inscription	Patterned Matter & Energy



The second problem: the proliferation (again):

	Snow is white	*a proposition*
can be expressed by	"Snow is white"	*a sentence*
can be encoded by:	S,n,o,w, ,i,s, ,w,h,i,t,e	*characters*
which can be encoded by:	*Snow is white.*	*glyphs*
which can be encoded by:	**83, 110, 111**...	*integers+**
which can be encoded by:	53, 6E, 6F...	*numerals*
which can be encoded by	01010011 01101110 01101111. . .	*binary octets*

[But how many levels are there here, really? *There can be any number*!]

The situation :

1) We have an indefinite number of symbolic encodings, not just one [or two]
2) the first level seems to be similar to a FRBR expression
3) the rest seem to be either encoding an expression, or encoding an encoding (!)

*taking liberties here to make a point.

Role and types, intuitively

We can distinguish two sorts of properties,
those that are roles and those that are types

Very roughly*:

The property being a student is a **role**
because persons are students only in virtue of particular contingent circumstances,
(namely: being enrolled in a school).

The property being a person is a **type**
because persons are persons regardless of contingent circumstances

Other properties that are roles: parent, president, planet nearest the sun, etc.

Other properties that are types: number, color, physical object, copper, water, etc.

There's much more to this; see: Guarino & Welty (2000) A Formal Ontology of Properties.

Roles vs types

Rigidity can help distinguish roles and types.

Because types are rigid and roles are not.

Defining rigidity (in natural language)

A property is rigid if and only if
Nothing that has it could exist and fail to have it
(in the past, present, or the future or in any alternative circumstances)

examples: person, number, color, physical object, copper. .

A property is a role if and only if
Anything that has it could exist and fail to have it
(in the past, present, or future or in any alternative circumstances)

examples: student, president, parent, planet nearest the sun

Adapted from Guarino & Welty (2000) A Formal Ontology of Properties.

A little more elucidation

The property of being a person is rigid

Nothing that is a person could exist and fail to be a person

so, a physical object could not have existed and not been a physical object (e.g., have been a color or a number instead)

The property of being a student is not rigid

Things that are students could exist and not be a student

so, a student may be a student now, but if their life had gone otherwise they might not have been a student.

Or:

Once you were not a student, now you are a student, soon you will not be a student.

But everything that is a person has been a person since it existed, and will always be a person as long as it exists. (same for numbers, colors, physical objects, etc)

The distinction in modal logic

In modal logic rigidity is defined:

A property ϕ is *rigid* **=df** $\Box(\forall x)(\phi x \rightarrow \Box\phi x)$

Or, in the model theoretic semantics for modal logic:

A property ϕ is *rigid* =df

if ϕ is had by some *x* in some possible world,

then *x* has ϕ in every possible world in which *x* exists

Guarino & Welty (2000) A Formal Ontology of Properties.

Are the middle entity types roles or types -- FRBR

With respect to the two models are are trying to align we can ask

"Are the middle entity types roles or types?

Let's start with FRBR.

A **expression** is a symbol structure that realizes a work

A **manifestation** is a symbol structure that embodies an expression

But:

Symbols have their meanings only as a result of contingent social convention

In different circumstances symbol structures mean different things

➔ So both expressions and manifestations would seem to be *roles*, not *types*.

Are the middle entity types roles or types -- language

Now lets consider our simple model of linguistic representation:

A **sentence** is a symbol structure that expresses a proposition

An **encoding** is a symbol structure that encodes a sentence

But, again:

Symbols have their meanings only as a result of contingent social convention

In different circumstances symbol structures mean different things

➔ So both sentences and encodings would seem to be *roles*, not *types*.

The middle entity types are *roles*, not *types!!*

Our model of a type/role relationship:

A **student** is a *person* enrolled in a school

So *being student* is a role that things of a particular type (persons) have in particular contingent circumstances.

This is parallel to:

A **sentence** is a *symbol structure* that expresses a proposition

An **encoding** is a *symbol structure* that encodes a sentence

red=types
green=roles

So we can collapse the two middle entity types
Both are symbol structures, the difference is a difference of role, not type.

But what about the proliferation problem?

	Snow is white	*a proposition*
can be expressed by	"Snow is white"	*a sentence*
can be encoded by:	S,n,o,w, ,i,s, ,w,h,i,t,e	*characters*
which can be encoded by:	*Snow is white.*	*glyphs*
which can be encoded by:	**83, 110, 111**…	*integers+**
which can be encoded by:	53, 6E, 6F…	*numerals*
which can be encoded by	01010011 01101110 01101111. . .	*binary octets*

[But how many levels are there here, really? *There can be any number*!]

The situation :

1) We have an indefinite number of symbolic encodings, not just one [or two]
2) the first level seems to be similar to a FRBR expression
3) the rest seem to be either encoding an expression, or encoding an encoding (!)

*taking liberties here to make a point.

Now we have a simple solution

Some new definitions:

A sentence is a symbol structure that expresses a proposition

An encoding is a symbol structure that encodes a [sentence or encoding]

So a single entity type can, with a recursive relationships,
represent an indefinite number of encoding levels

How cool is that?

You'll see in the next video.


5
AN ONTOLOGY
FOR DATA CONCEPTS

An ontology for data concepts

The preliminaries over; we now present our ontology of data concepts

This ontology generalizes and refactors FRBR

Accommodating both

-- collapsing middle entity types

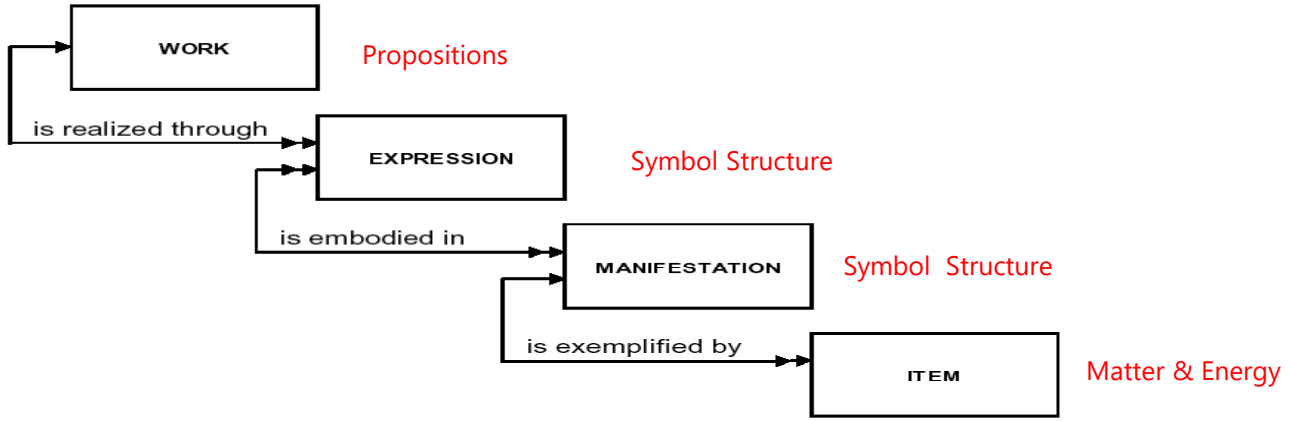
-- multiplying middle entity types

And illuminating the fundamental role of both standards and human intentionality

Recall why the middle two entity types seemed to collapse

How about:

FRBR	**Linguistic Representation**	**Entity Type**
Work	proposition	Proposition
Expression	sentence	Symbol Structure
Manifestation	encoding	Symbol Structure
Item	inscription	Patterned Matter & Energy



Recall why the two middle entity types seem to multiply

	Snow is white	*a proposition*
can be expressed by	"Snow is white"	*a sentence*
can be encoded by:	S,n,o,w, ,i,s, ,w,h,i,t,e	*characters*
which can be encoded by:	*Snow is white.*	*glyphs*
which can be encoded by:	**83, 110, 111**…	*integers+**
which can be encoded by:	53, 6E, 6F…	*numerals*
which can be encoded by	01010011 01101110 01101111. . .	*binary octets*

[But how many levels are there here, really? *There can be any number*!]

The situation :

1) We have an indefinite number of symbolic encodings, not just one [or two]
2) the first level seems to be similar to a FRBR expression
3) the rest seem to be either encoding an expression, or encoding an encoding (!)

*taking liberties here to make a point.

Recall our recursive solution

A sentence is a symbol structure that expresses a proposition

An encoding is a symbol structure that encodes a [sentence or encoding]

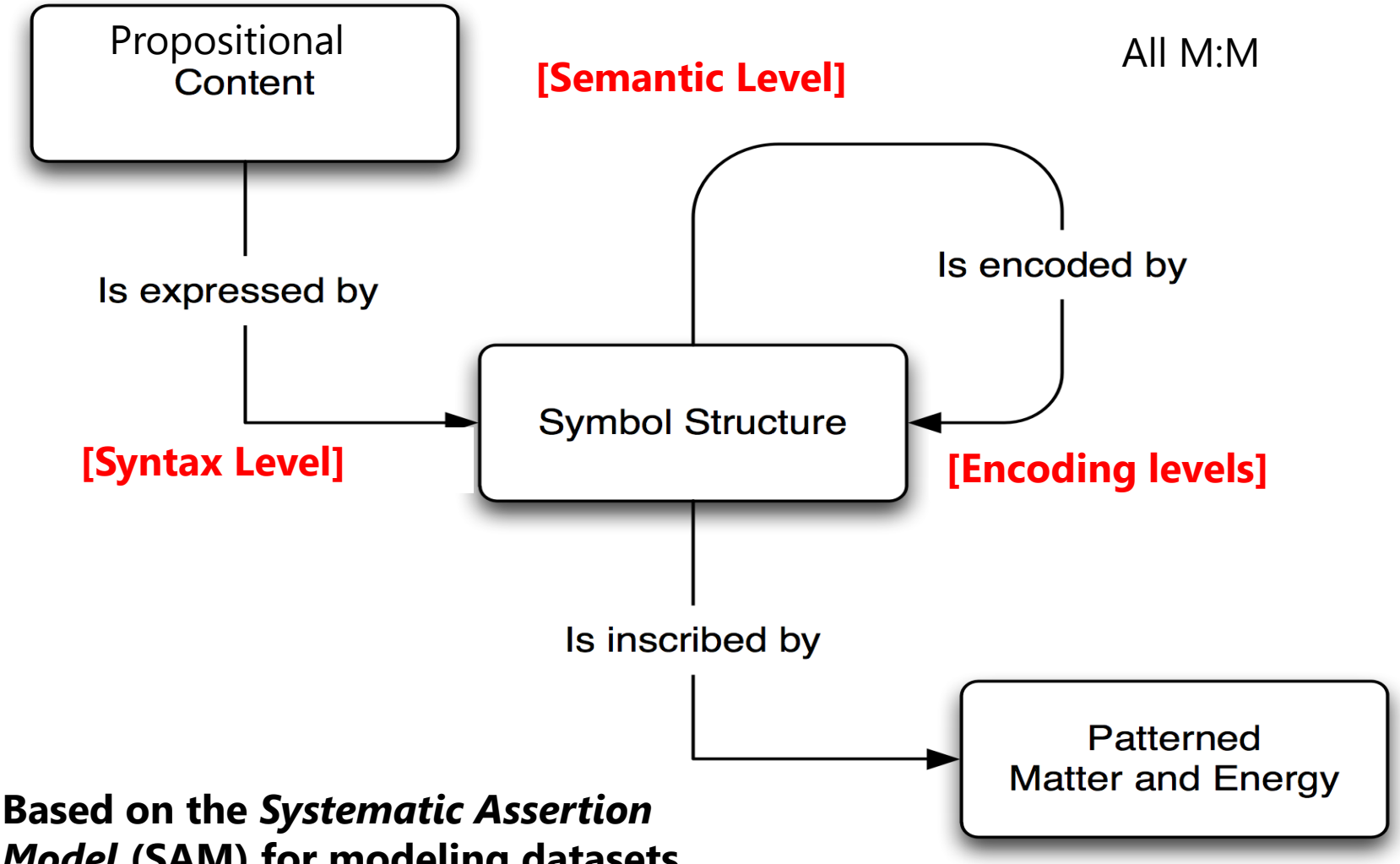
Ok, here we go

An ontology for data. . .

<drama type="drumroll" kind="imaginary" alttext="Drum roll please">

[Drum roll please]

The basic representation model (or FRBR refactored)



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

For example:

C1: propositions
expressed by...

S1: RDF triples
encoded by...

S2: RDF/XML
encoded by ...

S3: Unicode characters
encoded by...

S4: UTF-8 bit streams
inscribed in...

M1: actual RAID array state

Interpretive frames

How do *expressing, encoding,* and *inscribing* actually happen?

Part of the answer: information processing standards (e.g., those from ISO, IETF, W3C, NISO, etc.).

Within data representation and processing realms these codify things like:

"<p>" indicates a paragraph	(HTML).
integer 101 encodes a latin "e"	(Unicode/ASCII)
octet 01100101 encodes integer 101	(Unicode/ASCII)

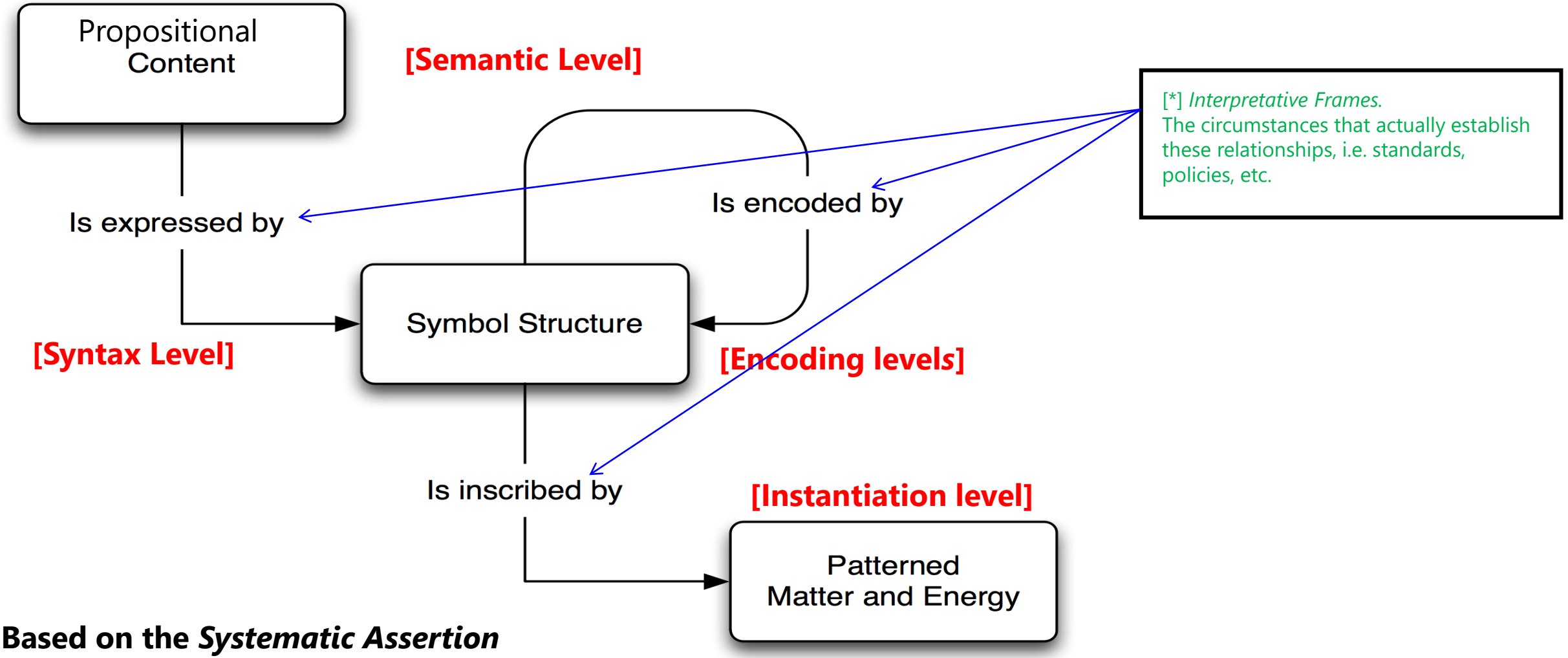
These include simple mappings as above, but also specifications of syntax and semantics for data representation languages.

In addition natural language prose descriptions are also important – and difficult to interpret precisely

We call all these things: *interpretative frames*.

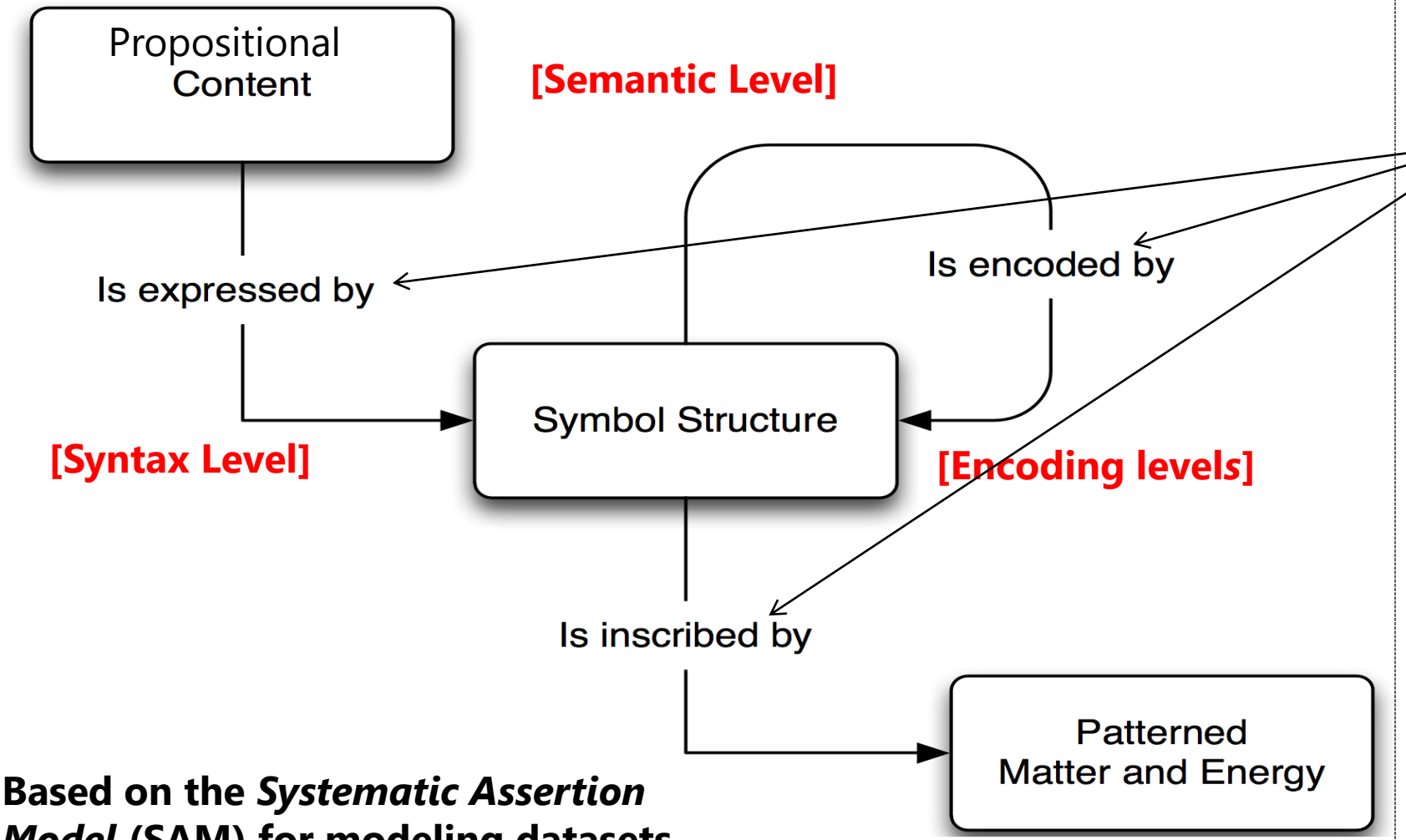
[Dubin, Wickett, & Sacchi, (2011)]

FRBR refactored and extended. What's still missing? [*]



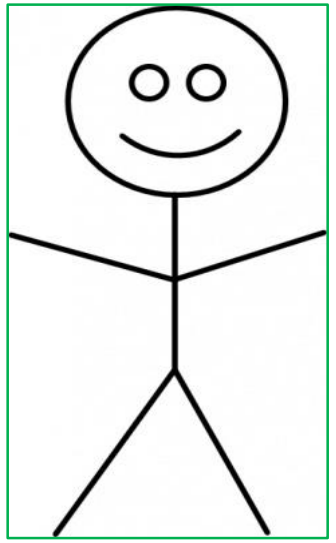
Based on the *Systematic Assertion Model* (SAM) for modeling datasets, developed by David Dubin et al.

Also needed: human intentionality [*]



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

Interpretative Frames



[*] standards, policies, etc. are inert unless they are intended, by people, to be applied; so we also need beliefs, intentions, practices, etc.

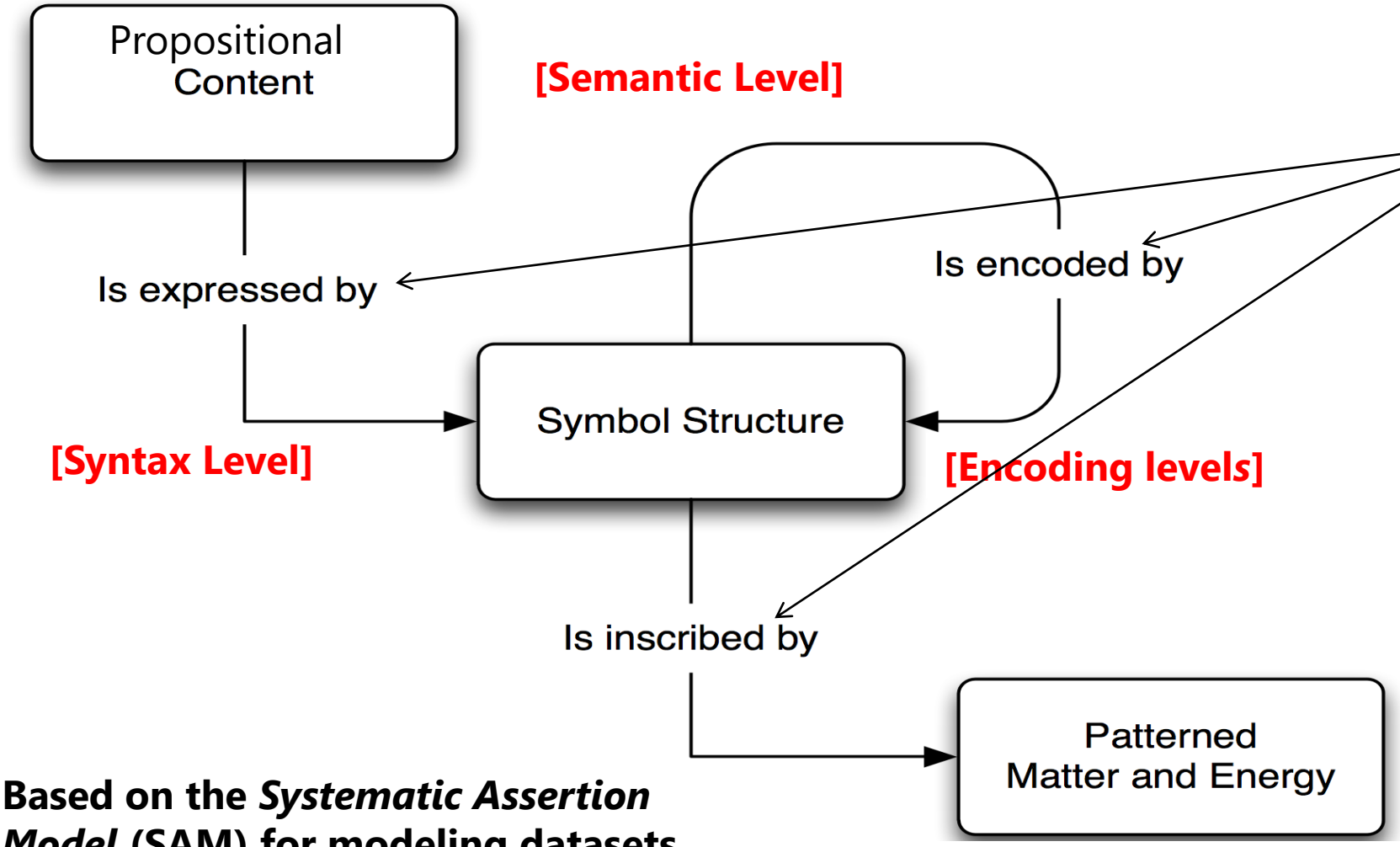
The importance of human intentionality

Human agreement and intentionality is particularly prominent in the digital world in the form of *standards* and *policies*.

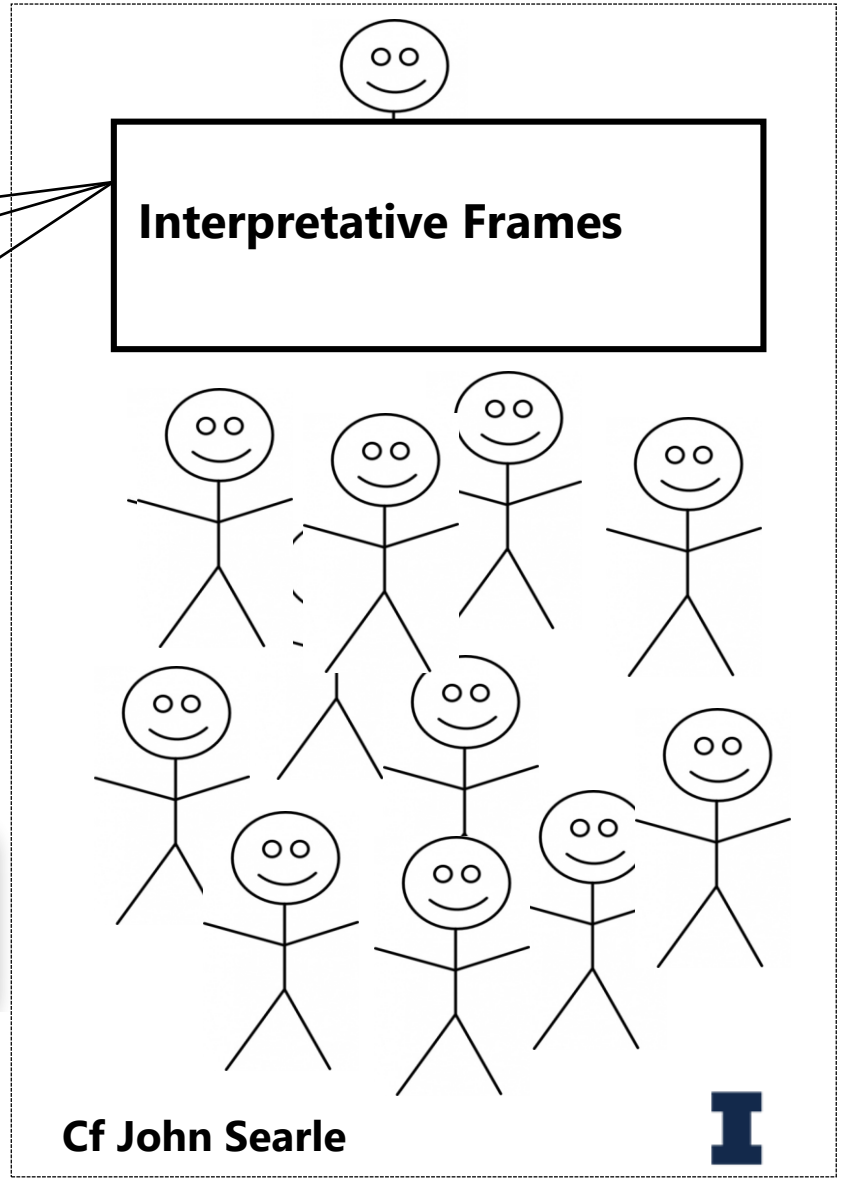
But standards and policies alone are not enough.

The circumstances that establish and sustain the contingent relationships indicated in the model also involve, and essentially involve, the actual collective beliefs, intentions, and expectations of engineers, programmers, and end users.

Actually "it takes a village" (i.e. *collective* intentionality)



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




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:

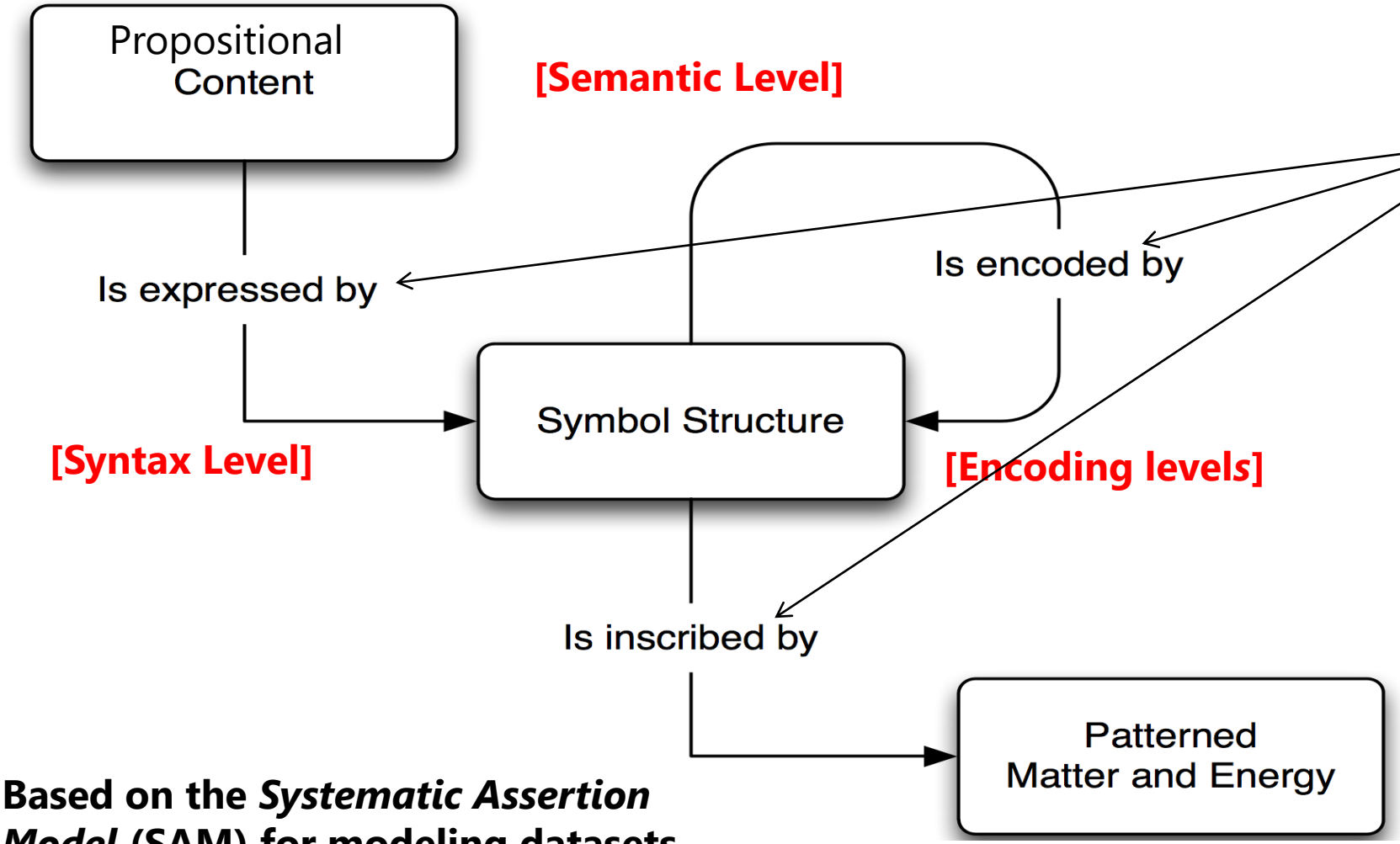
-- nothing is data intrinsically, only with respect to use

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

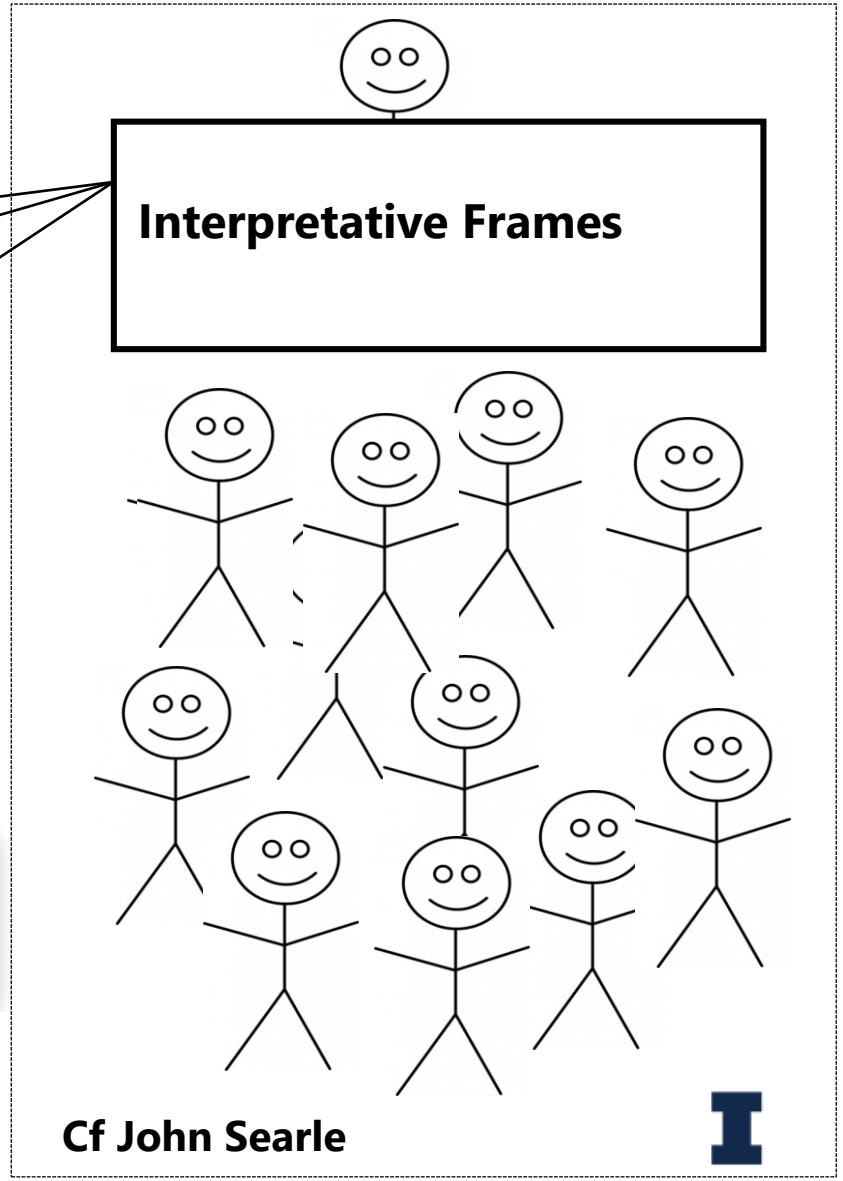
And it is routinely the case in science that

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 role 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.