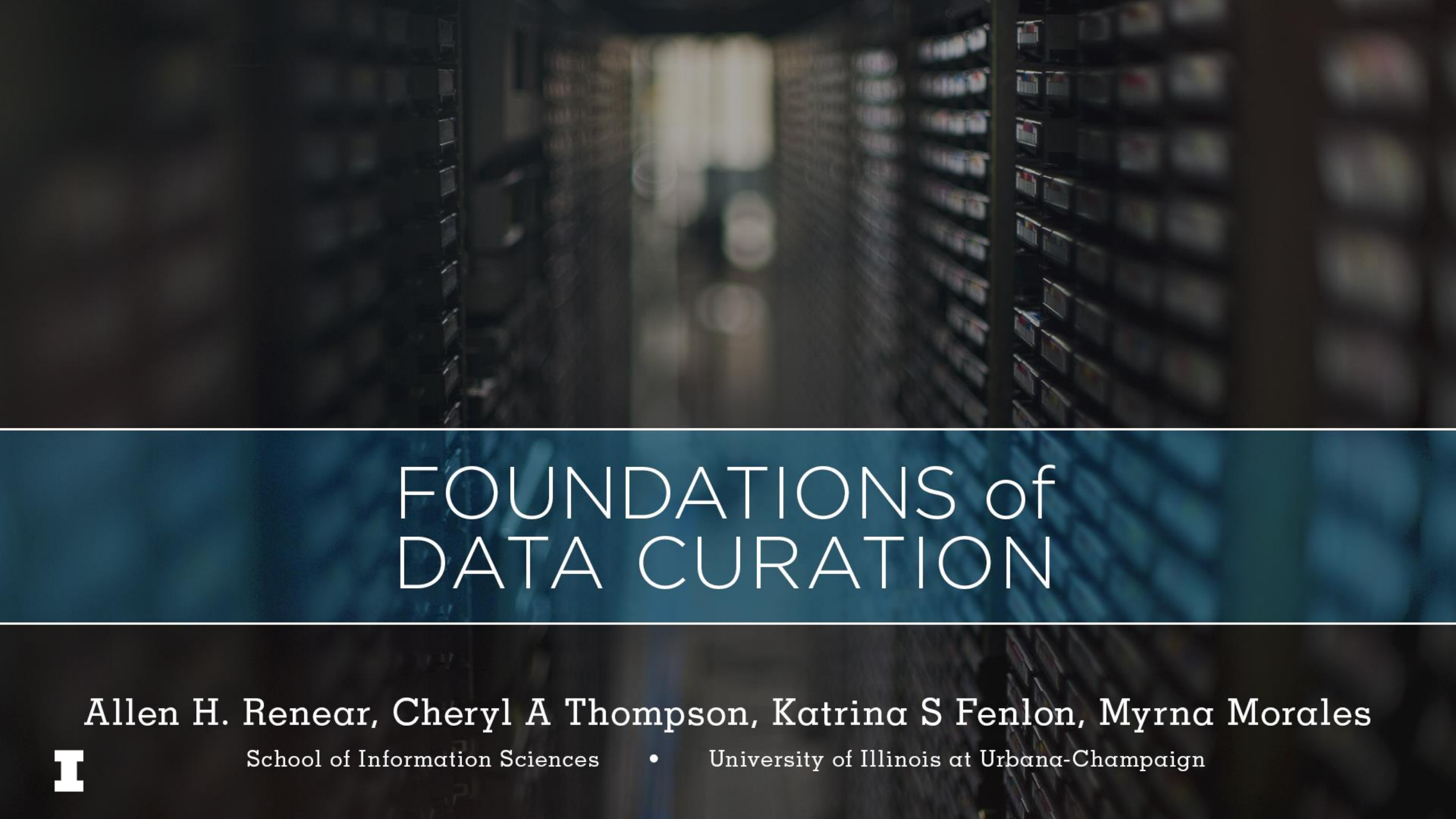


FOUNDATIONS of DATA CURATION

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

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


DATA
CONCEPTS


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

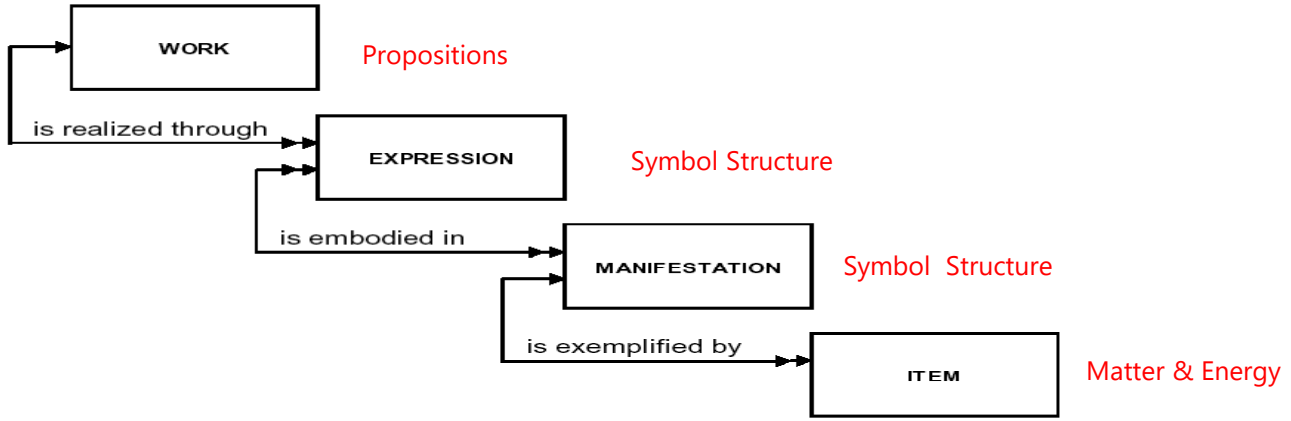
-- multiplying middle entity types

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]

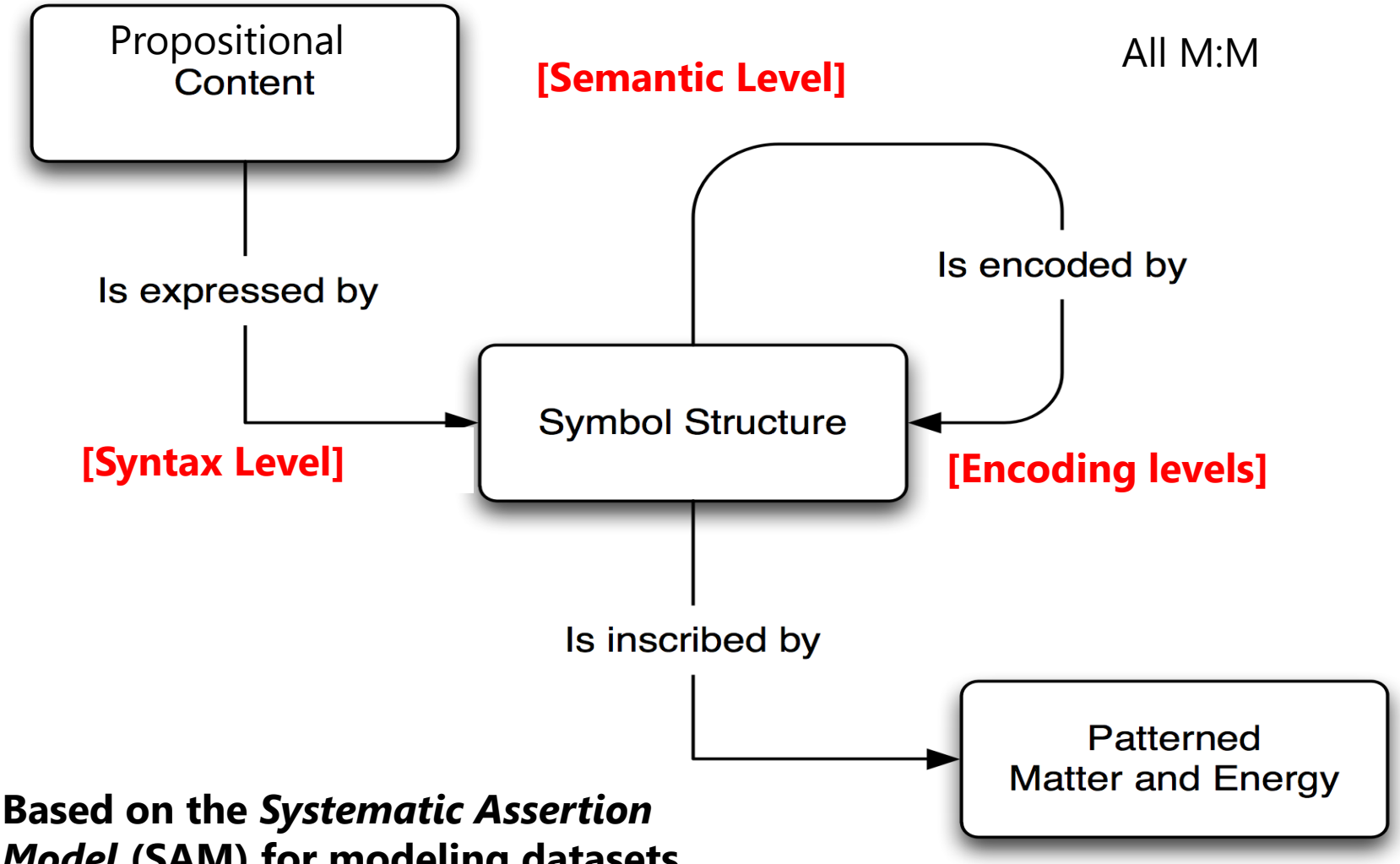
Ok, here we go

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)

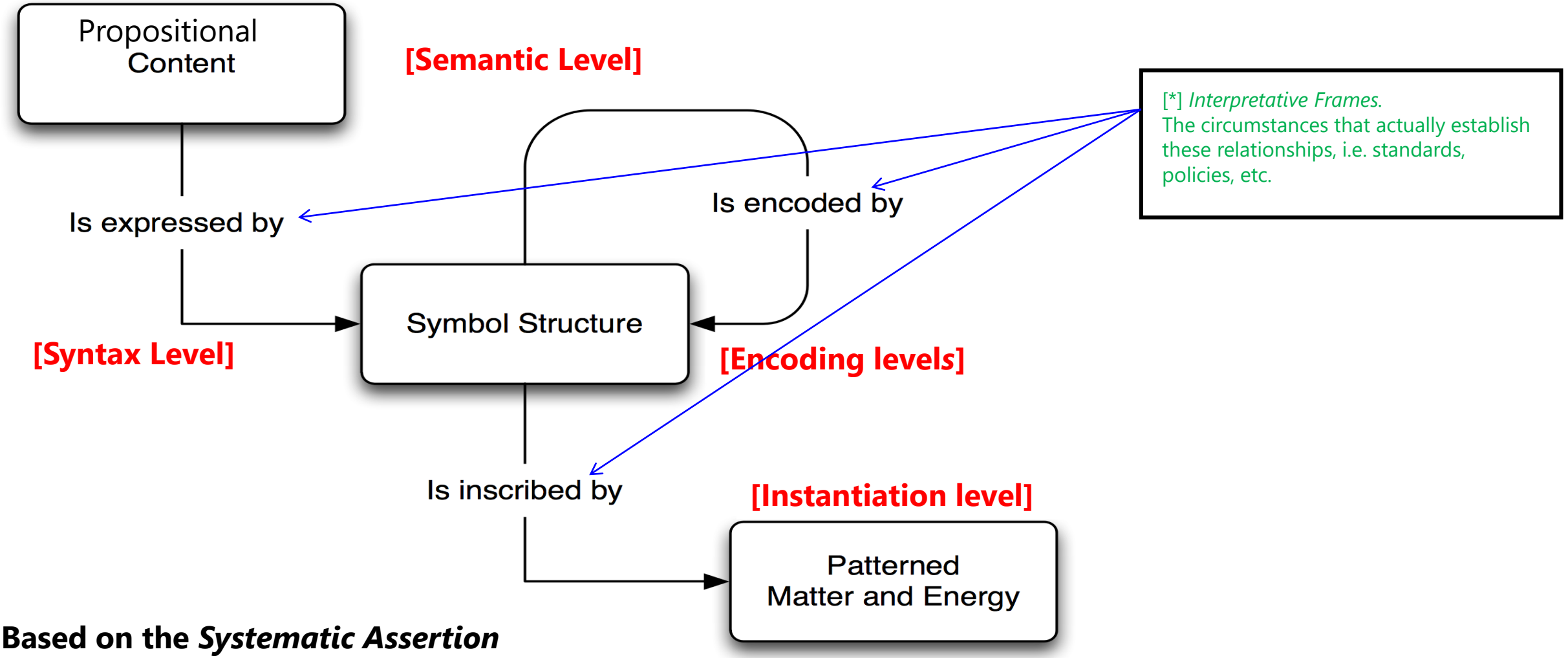
These include simple mappings as above, but also specifications of syntax and semantics for data representation languages.

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

We call all these things: *interpretative frames*.

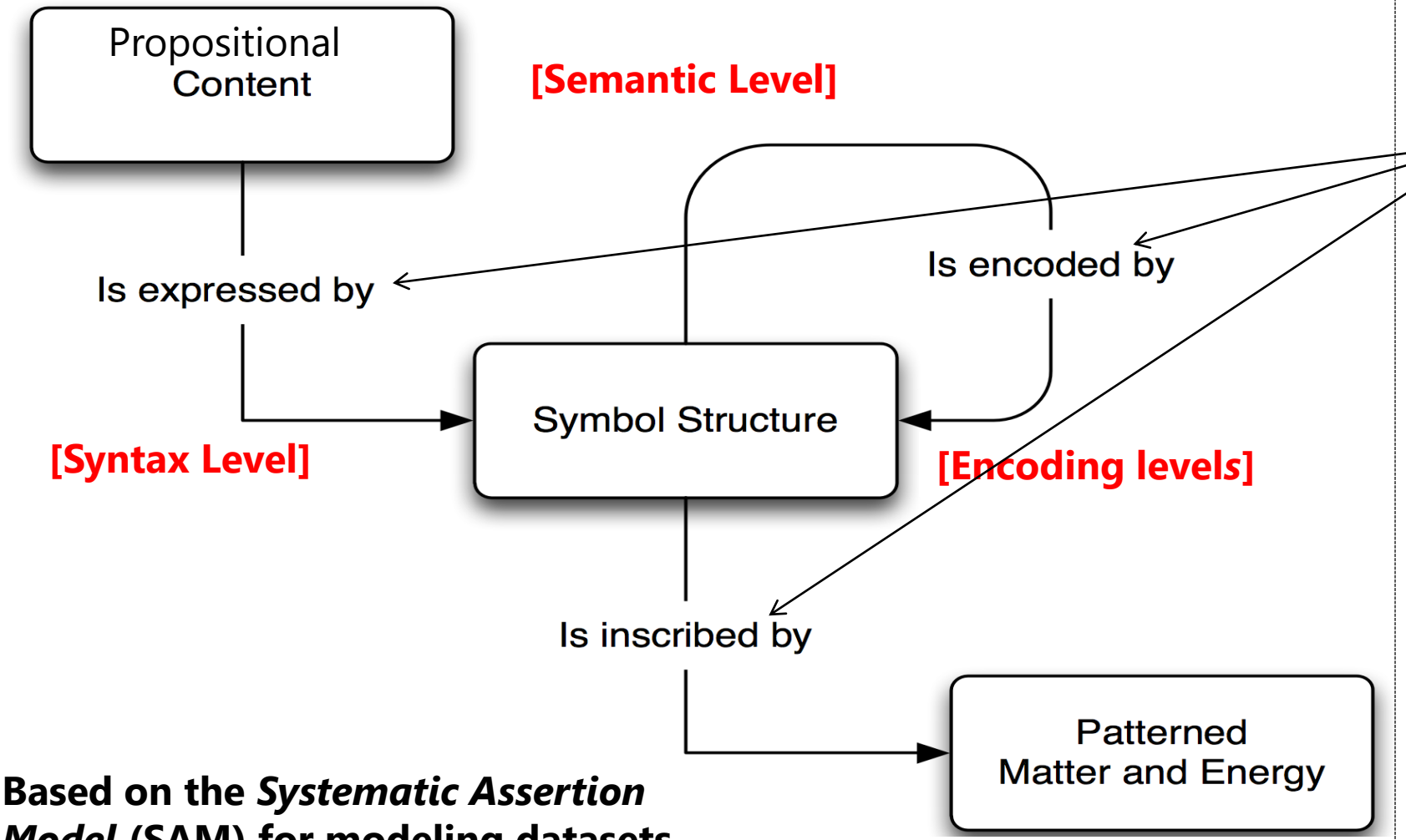
[Dubin, Wickett, & Sacchi, (2011)]

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



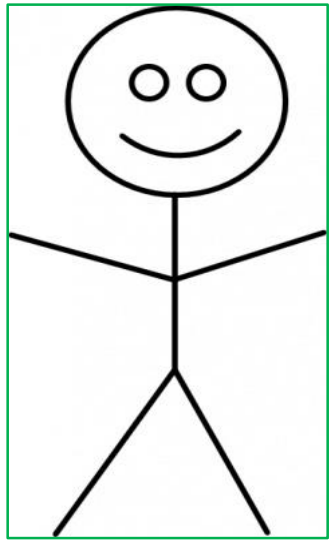
Based on the *Systematic Assertion Model* (SAM) for modeling datasets, developed by David Dubin et al.

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.

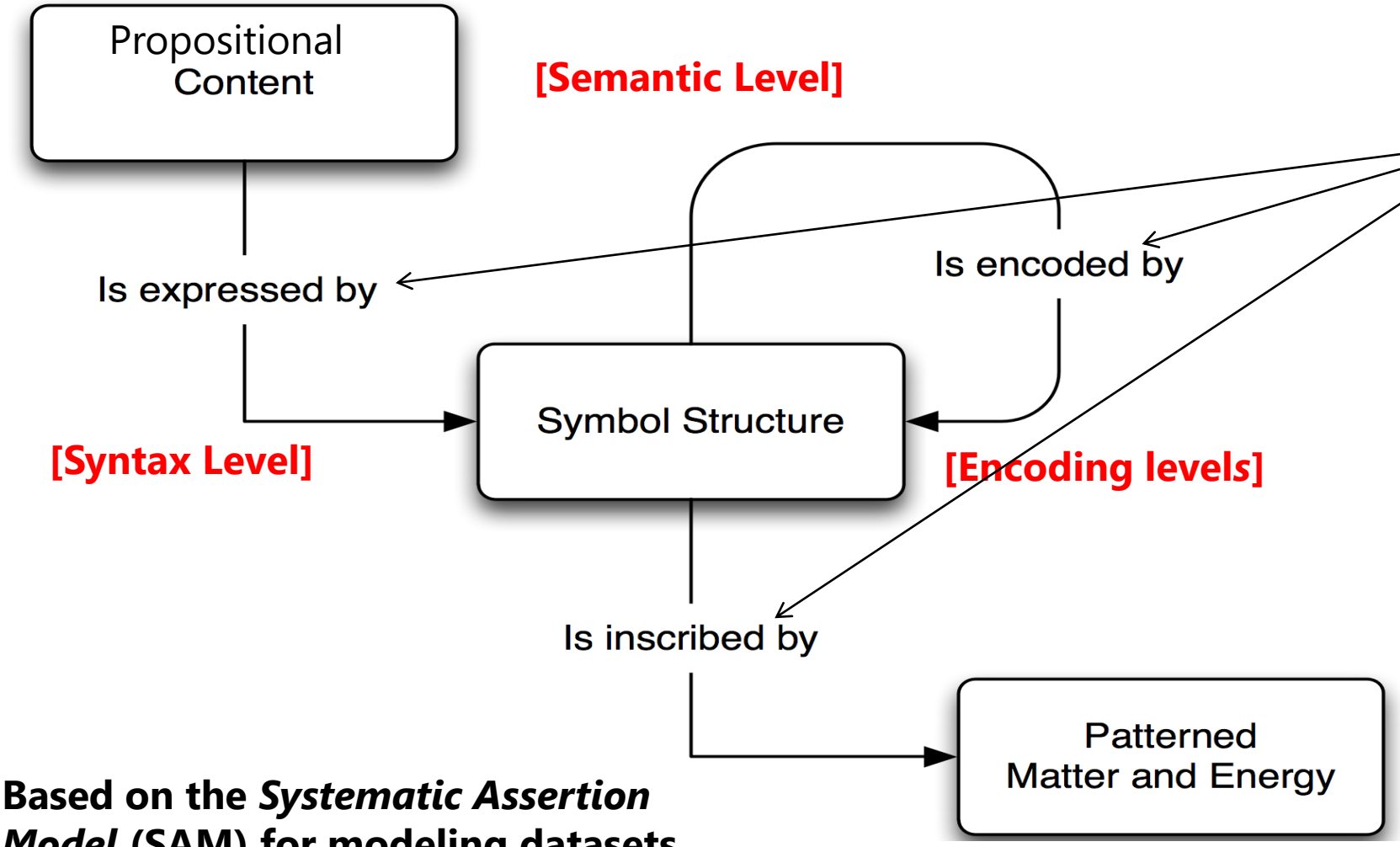
The importance of human intentionality

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

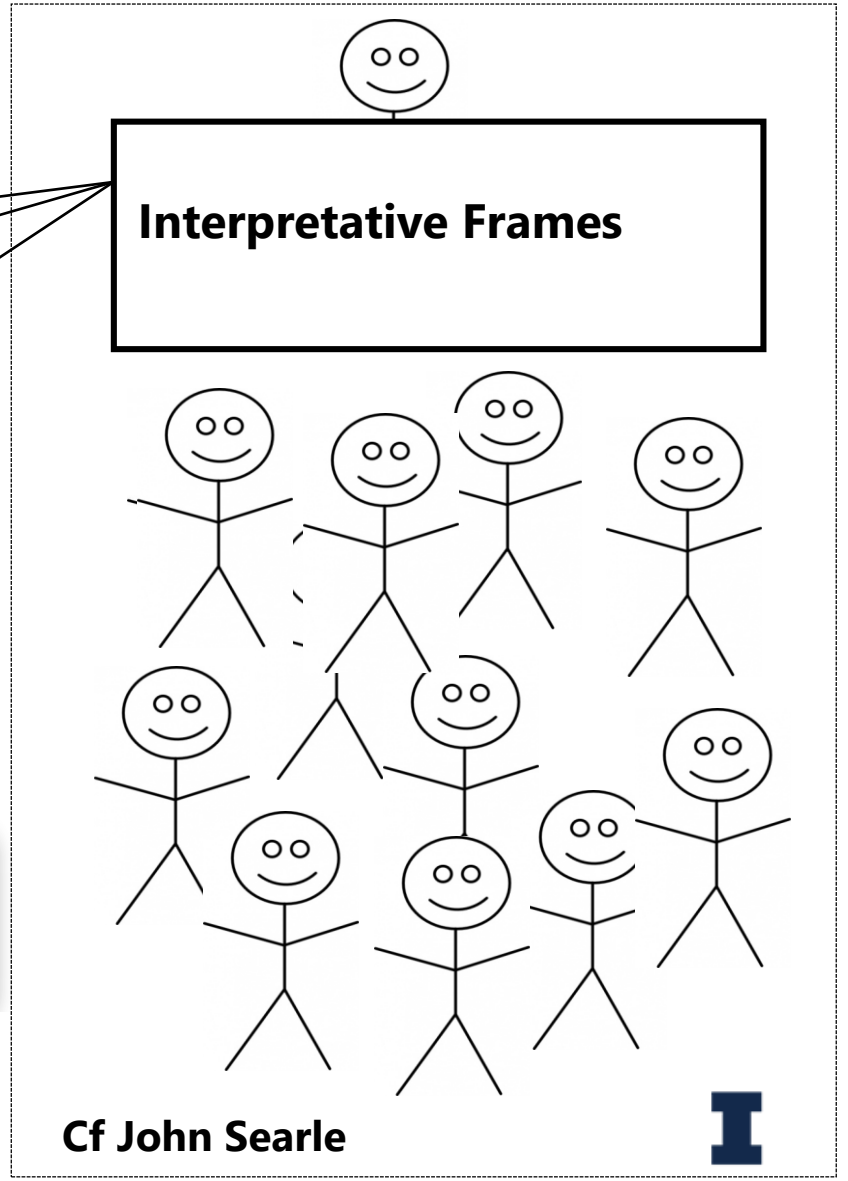
But standards and policies alone are not enough.

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.



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.