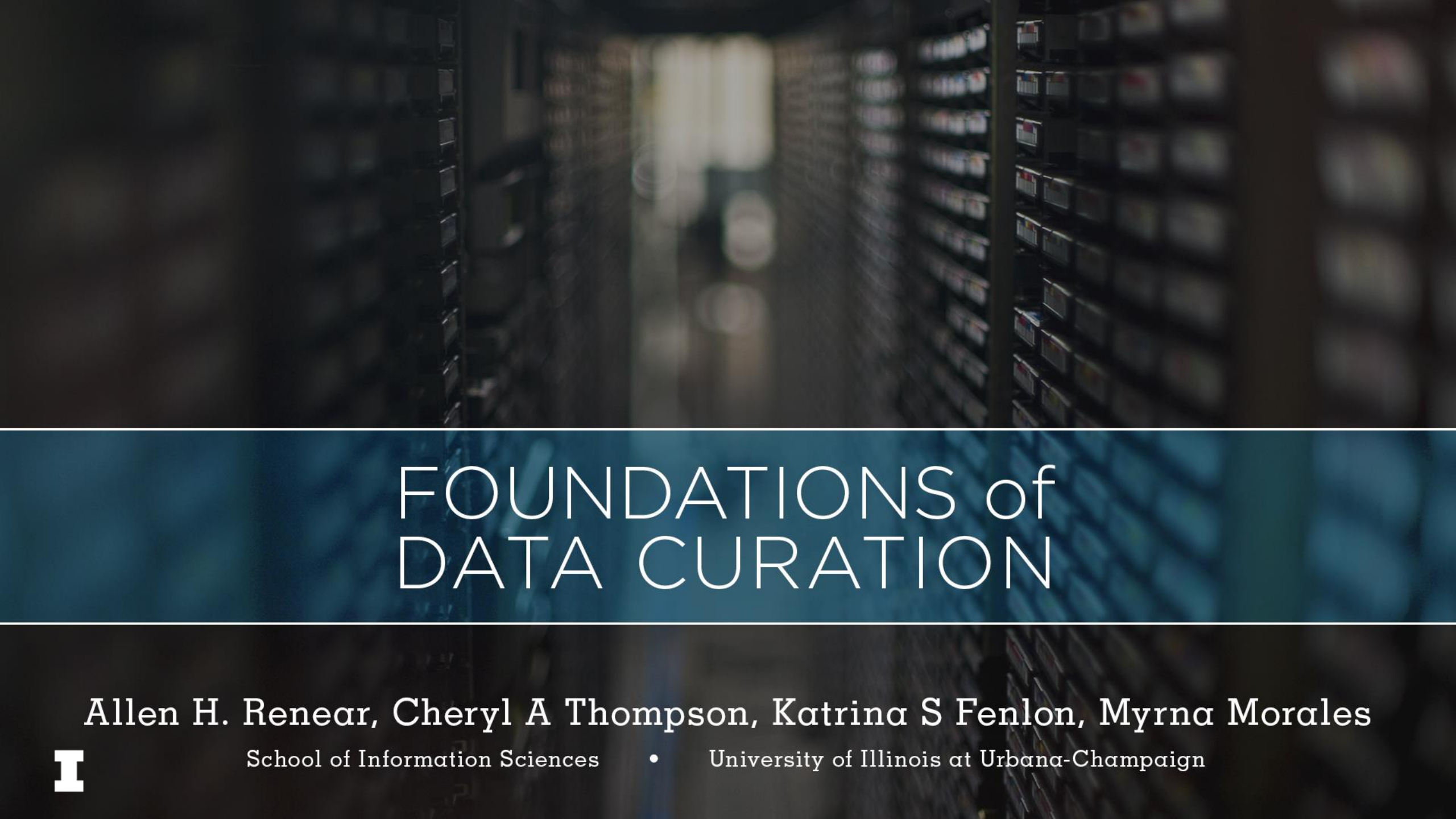


FOUNDATIONS of DATA CURATION

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

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


DATA
CONCEPTS

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 . . .

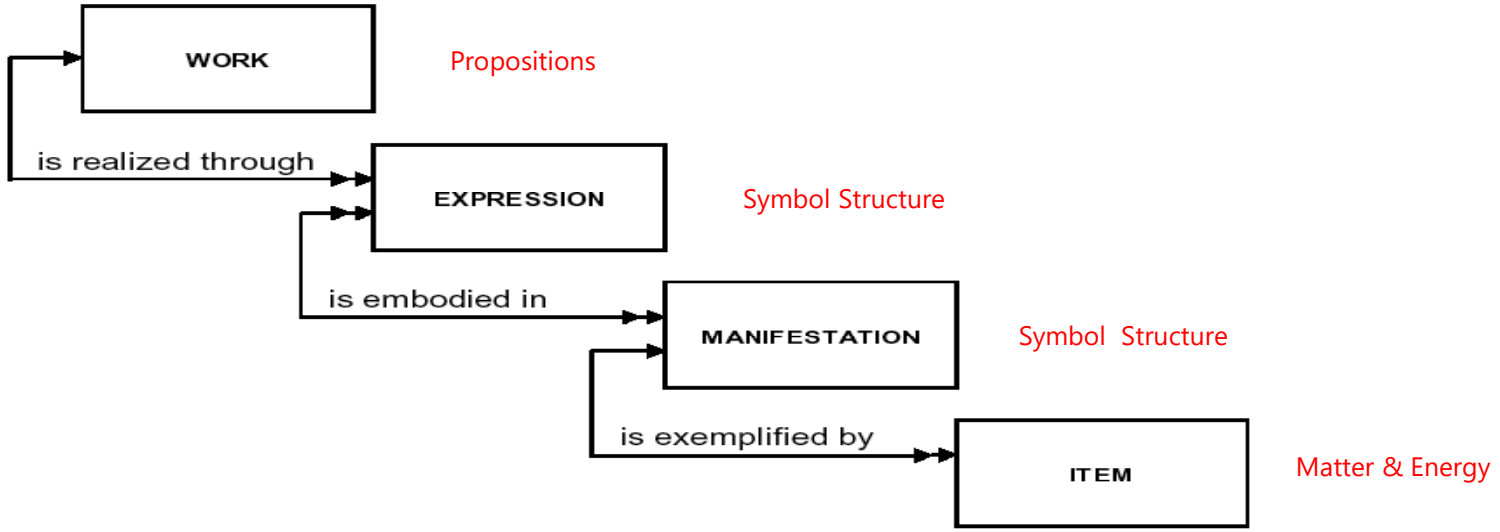
Ok, where have we got to?

How about:

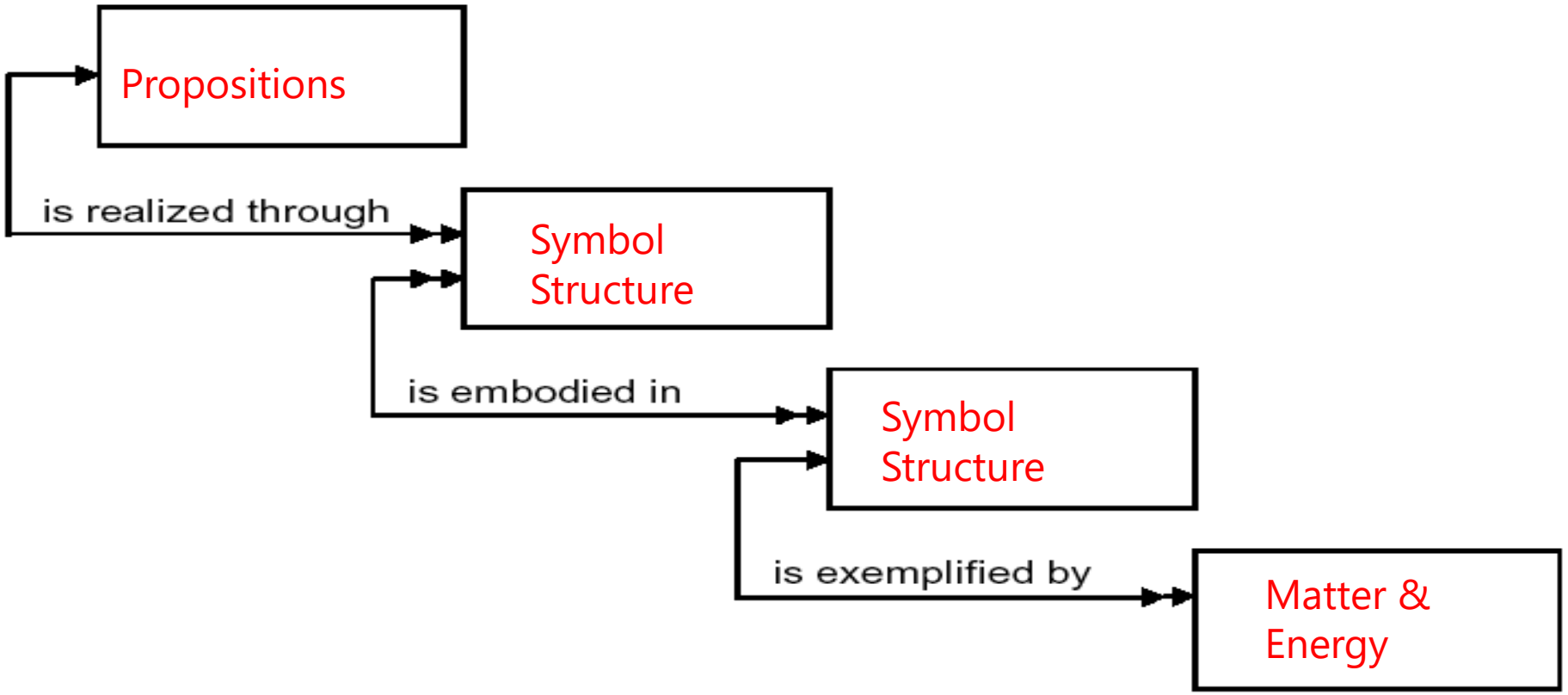
FRBR	**Linguistic Representation**	**Entity Type**
Work	proposition	proposition
Expression	sentence	?
Manifestation	encoding	?
Item	inscription	Patterned Matter & Energy

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

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.