Identity and Identifiers

Contents

V2: *What* are we identifying?

What exactly needs to be identified?

Lots of things to be identified

We need identifiers for lots of things:

Because without shared identifiers we cannot reliably communicate what we are taking about!

So: persons, properties, values, counties, automobiles, nations, proteins, events, etc. . .

(EVERYTHING!)

There are some shared standards in specialized domains,
and some good systems for developing identifiers, but much remains to done

Identifiers are at the heart of the semantic web:

Identifiers in RDF/OWL are URIs:

<http://example.org/#spiderman>
<http://www.perceive.net/schemas/relationship/enemyOf>
<http://example.org/#green-goblin>

Or literals:

<http://example.org/#spiderman>
<http://xmlns.com/foaf/0.1/name>
"Spiderman”

All these are important in data curation.

But in this video we will focus on identifiers for datasets.

Identity and Representation Levels (again)

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*

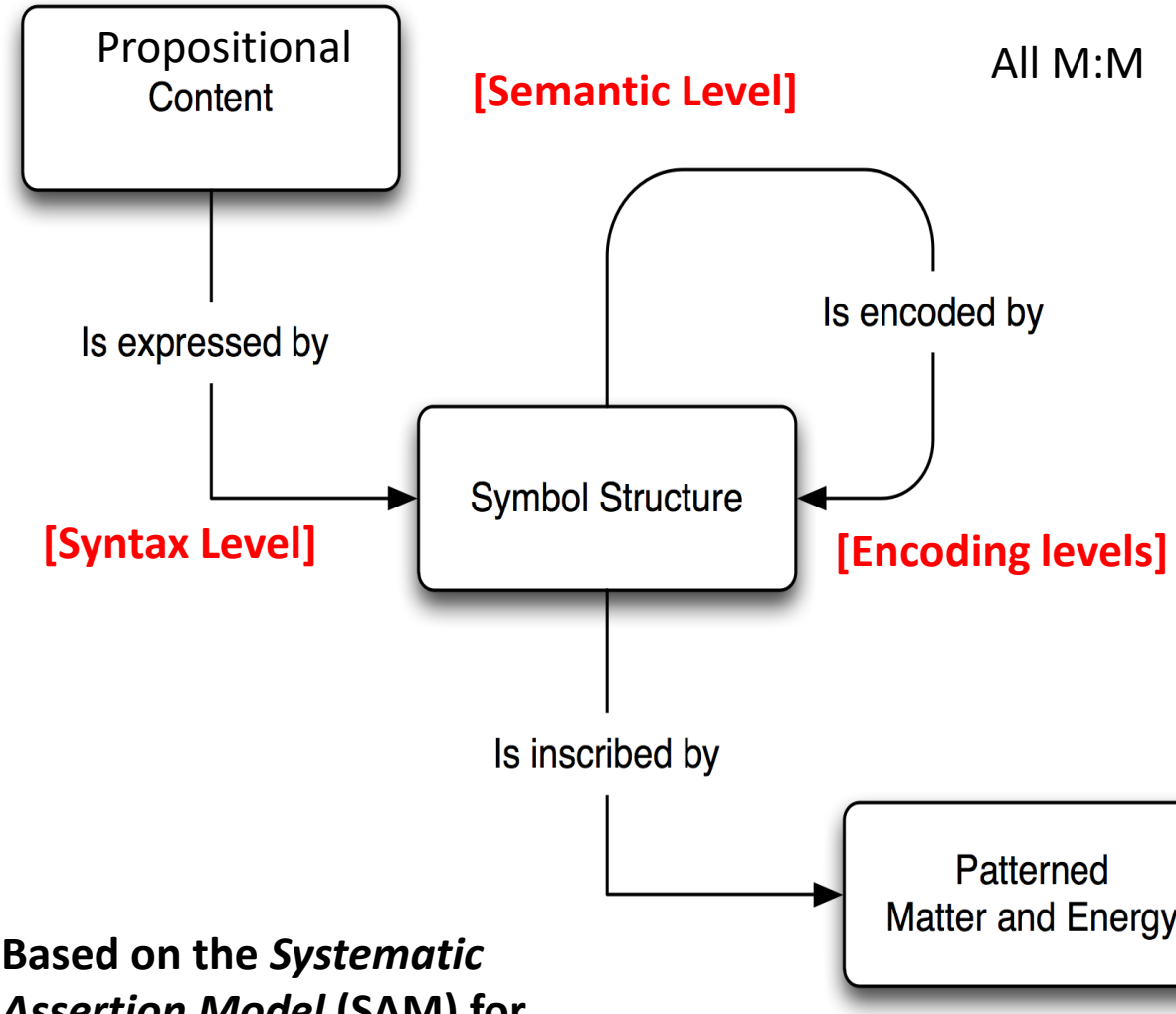
Identity and Representation Levels

How many of these levels are there ?

How do we name, define, and manage them?

How can they be identified and re-identified?

The Basic Representation Model (or FRBR refactored)



For example:

C1: propositions
expressed by…

S1: RDF triples
encoded by…

S2: RDF/XML
encoded by …

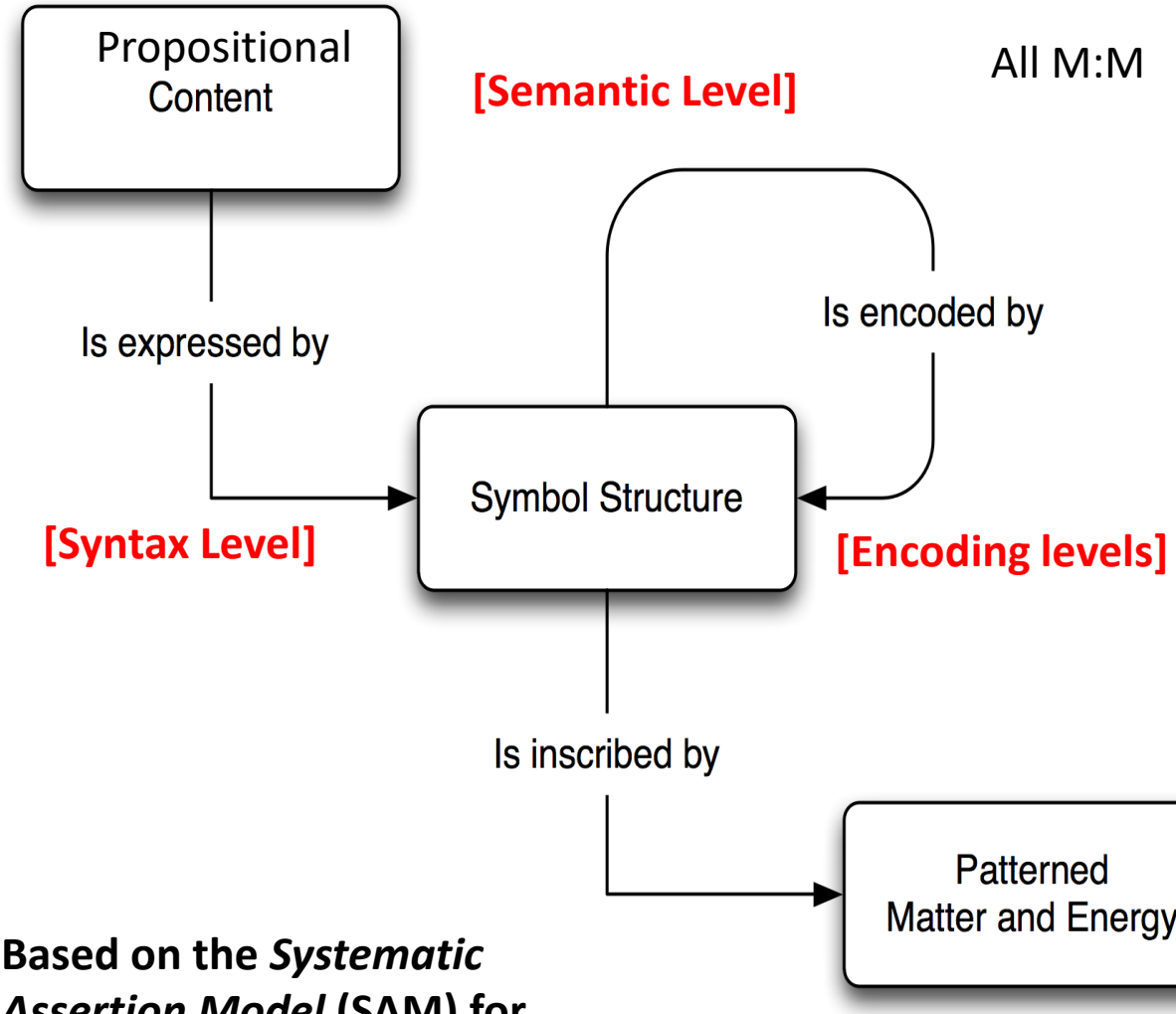
S3: Unicode characters
encoded by…

S4: UTF-8 bit streams
inscribed in…

M1: actual RAID array state

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

So what are we identifying?



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**

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