Metadata

What is metadata? [first definition]

The simple, and most common, colloquial definition is:

data about data

What is metadata? [a better definition]

"structured data about an object
that supports functions associated with the designated object"

(Greenberg, 2003)

[here the concept of *object* includes *data set*}

The standard classification of metadata by function

Descriptive	For describing a resource to support things like finding, understanding, evaluating, choosing among digital objects or data
Administrative Technical Preservation Rights	 For decoding and rendering For long-term management For describing intellectual property rights
Structural	For relating parts of resources to one another

Adapted from:
http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

Conceptual Metadata schemas vs. their serialization

Dublin Core metadata element set

(select terms)

- **Creator:** William Blake
- **Title:** “A Sick Rose”
- **Date:** 1794

Serialized as RDF/XML

<xml> <?namespace href = "http://www.w3.org/schemas/rdf-schema" as = "RDF"> <?namespace href = "http://www.purl.org/RDF/DC/" as = "DC">
<RDF:RDF>
<RDF:Description RDF:HREF="http://purl.org/metadata/dublin_core_elements"

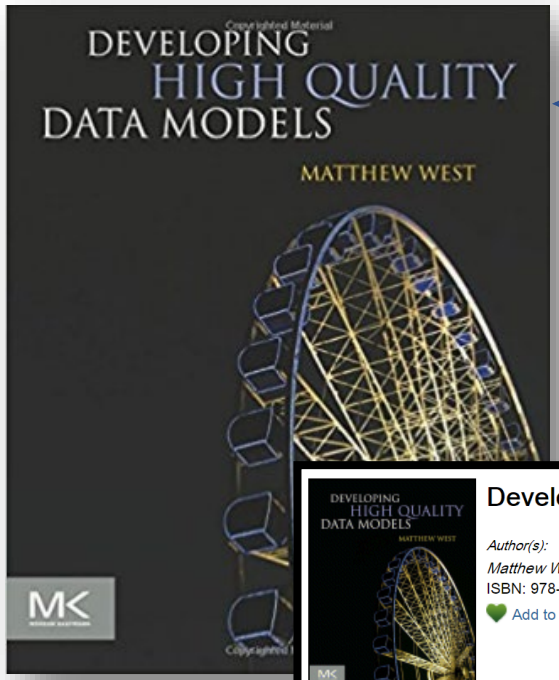
DC:Title = "The Sick Rose" DC:Creator = “William Blake" DC:Date = “1794" />

</RDF:RDF> </xml>

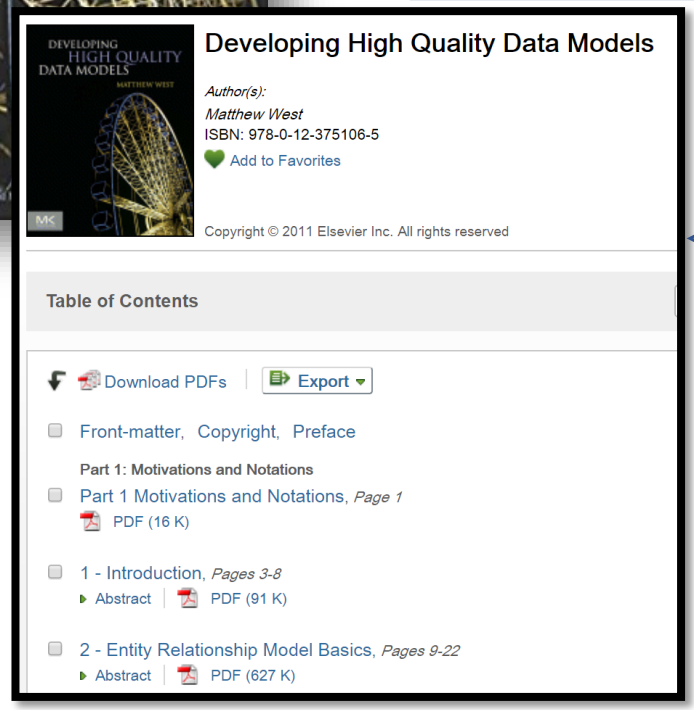
Serialized with HTML meta elements.

```
<meta name="DC.Title"      content=”The Sick Rose">
<meta name="DC.Creator"    content=“William Blake">
<meta name="DC.Date"       content=”1794">
```

"*What* is being described?" problems (think FRBR)



identifier (ISBN):	978-0123751065
author:	Matthew West
title:	Developing High Quality Data Models
date:	2011
publisher:	Morgan Kaufmann
subject:	database design
subject:	data structures (computer science)
language	English
pages:	**408**



identifier (ISBN):	978-0123751065
author:	Matthew West
title:	Developing High Quality Data Models
date:	2011
publisher:	Morgan Kaufmann
subject:	database design
subject:	data structures (computer science)
language:	English
pages:	**389**
format:	**PDF**
identifier:	http://www.sciencedirect.com/science/book/9780123751065

Metadata is fundamental to data curation

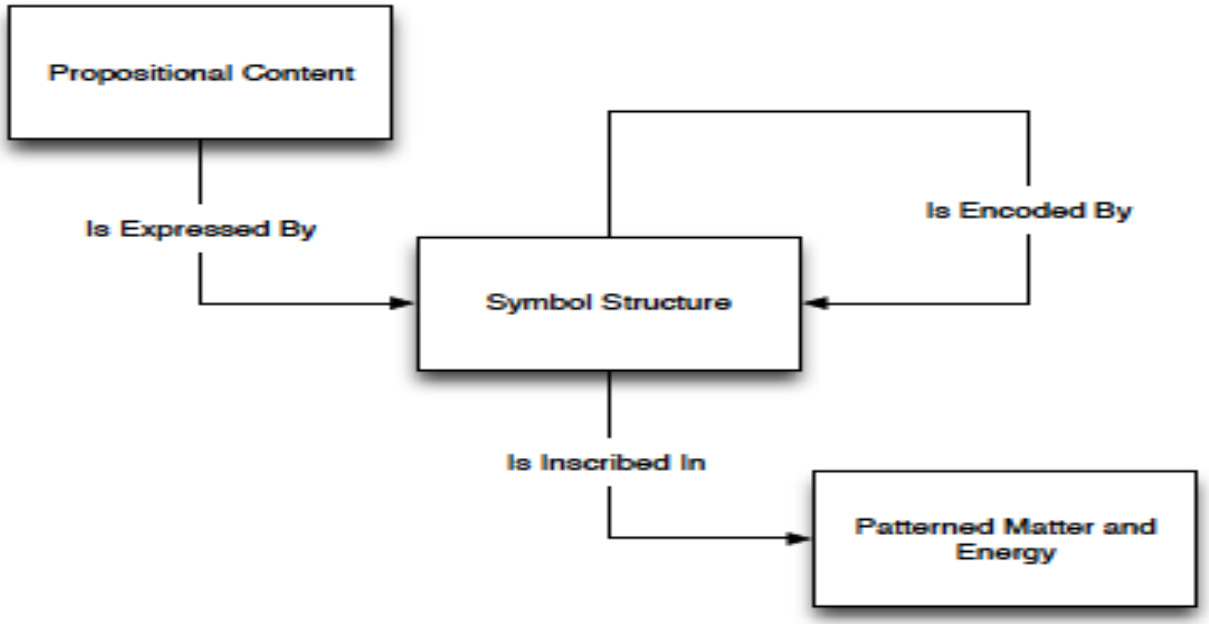
Area	Description	Supported by metadata that documents. . .
Collection	Support the collection and acquisition of data	Method, location, time, instruments, settings, calibration. . .
Organization	Employ an appropriate data model and use appropriate standards	Schemas and schema documentation for semantics, syntax, and encoding.
Storage	Support reliable and effective storage	Authoritative and alternative copies, physical locations, redundancy, compression, reduction, backups
Preservation	Ensure that data will be understandable and useable in the future	[see Organization, Identification, Storage]
Discoverability	Support the ability to search for and locate relevant data	Topic, coverage, formats, availability, currency.
Access	Support the ability to retrieve and distribute data	[see Organization, Security], licensing, owner, location.
Workflow	Support the ability to systematize data workflows	[See Organization], scripts, processes, transformations, inputs.
Identification	Support the ability to identify, authenticate, and validate data	[See Organization], identifiers, version data, integrity checks, authentication.
Integration	Support integration of data from different sources using different data models	[See Organization, Identification, Access, Discoverability]

Metadata is fundamental to data curation

Activity	Description	Metadata documenting . . .
Modification	Support management of corrections and updates	[See Organization, Workflow, Provenance, Identification]
Reformatting	Support reformatting for use by different tools or to match new format standards	[See Organization, Identification, Workflow]
Provenance	Support identifying what inputs and calculations are responsible for data values	[See Workflow, Identification, Reproducibility]
Reproducibility	Support ability to reproduce results, ensuring scientific validity	[See Organization, Workflow, Provenance, Identification]
Preservation	Ensure that data will be understandable and useable in the future	[see Organization, Identification, Storage]
Compliance	Ensure compliance to legal, regulatory, and local policy requirements	[see Organization, Provenance, Workflow, Discoverability]. Certification.
Security	Ensure that data is secure from tampering or inappropriate access and distribution	[see Organization, Provenance, Workflow, Discoverability]. Encryption, Certification.
Communication	Support representation, publishing, and visualizations that provide insight	[See Organization, Identification, Reproducibility, Compliance]
Sharing	Support sharing data between researchers, teams, and institutions.	[See Discoverability, Organization, Workflow, Provenance, Identification]

Preservation

Recall our data ontology



Now ask yourself:

In a successful preservation scenario, *what exactly is preserved?*

What . . . exactly . . . is . . . preserved?

No, really, ***what?***

Preservation is not about preserving any *thing*

Data preservation is not about preserving the existence of objects

It is about *communication with the future*. [1]

The best simple definition: preservation is. . .

Ensuring reliable communication with the future [2]

- [1] Reagan Moore, Towards a theory of digital preservation, *The International Journal of Digital Curation*, 2008
- [2] Simone Sacchi, *What do we mean by preserving digital information'? Towards sound conceptual foundations for digital stewardship* (Doctoral dissertation, University of Illinois at Urbana Champaign), 2015.

The definition expanded. . .

Preservation is not about preservation.

Preservation is *ensuring reliable communication with the future*

More exactly: preservation actions are intended to ensure that future researchers. . .

1) will come into possession of physical media and encodings
2) from which they will correctly recognize the originally intended propositional content
3) and from which they will be justified in believing that this propositional content is in fact the intended propositional content

* this can be adapted to more explicitly accommodate software agents and automatic processing. The key thing is that the process is reliable: all interpretations are (1) correct and (2) justified.

And a longer one*

Viable	can be read (correctly) from media
Renderable	can be (correctly) viewed, processed, executed
Understandable	can be (correctly) understood
Authenticatable	can be (correctly) determined to be what it purports to be
Identifiable	can be (correctly) identified and re-identified

And more can be added of course: findable, conformant . . . etc..

Following our own definition of preservation we would emphasize that

each of these five not only must be achievable,

but the user must have *justified confidence* in the result.

*See: PREMIS https://www.loc.gov/standards/premis/

Four common strategies

Replication

Make lots of copies, distribute them widely

Migration

Keep updating your data to new formats, as needed

Emulation

Maintain software that emulates the original processing

Normalization

Convert data sets to a standard format optimized for preservation

Transformations

Both migration and normalization strategies involve transforming a data set in one format to a data set in another format,

both presumably with the same information.

Ideally the transformation, as well as the resulting data set, should be documented in a standard computer-processable metadata languages.

Regardless of whether this is a migration scenario or normalization scenario

The next slide indicates one way this could happen.

[This is also an example of *workflow* and *provenance* documentation.]

Example of transformation documentation (very liberally modified from UIUC Medusa record by T. Habing)

<event version="2.1">
<eventIdentifier>
<eventIdentifierType>LOCAL</eventIdentifierType>
<eventIdentifierValue>MEDUSA:b21248fa-75ac-4c45-aae3</eventIdentifierValue></eventIdentifier>
<eventType>MIGRATION</eventType> <eventDateTime>2011-05-03T10:15:32</eventDateTime>
<eventDetail> **The contentdm_record_1.xml file was transformed into the mods_1.xml file using XSLT.** </eventDetail>
<linkingAgentIdentifier>
<linkingAgentIdentifierType>UIUC_NETID</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>UIUC\gnibaht</linkingAgentIdentifierValue></linkingAgentIdentifier>
<linkingAgentIdentifier>
<linkingAgentIdentifierType>FILENAME</linkingAgentIdentifierType>
<agentIdentifierValue> **contentDM_to_MODSv32.xsl** </agentIdentifierValue></linkingAgentIdentifier>
<linkingObjectIdentifier>
<linkingObjectIdentifierType>FILENAME</linkingObjectIdentifierType>
<linkingObjectIdentifierValue> **mods_1.xml** </linkingObjectIdentifierValue></linkingObjectIdentifier>
<linkingObjectIdentifier>
<linkingObjectIdentifierType>FILENAME</linkingObjectIdentifierType>
<linkingObjectIdentifierValue> **contentdm_record_1.xml** </linkingObjectIdentifierValue></linkingObjectIdentifier>
</event>...
<agent version="2.1"> . .

A reformatting recorded in computer processable documentation; here PREMIS XML.

The input and output files are in bold black, the XSL file that specifies the transformation is in red.
Documented transformation can support both migration strategy and normalization.

To see how agent is used to represent software associated with a preservation event in this example (search "FIDO") in
https://www.loc.gov/standards/premis/v3/sample-records/PREMIS-3-example-1.xml

(The example is liberally modified from UIUC Medusa record)

Identity and Identifiers

Identification Problems in Data Curation (again)

Identification problems

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?

Lots of things to be identified

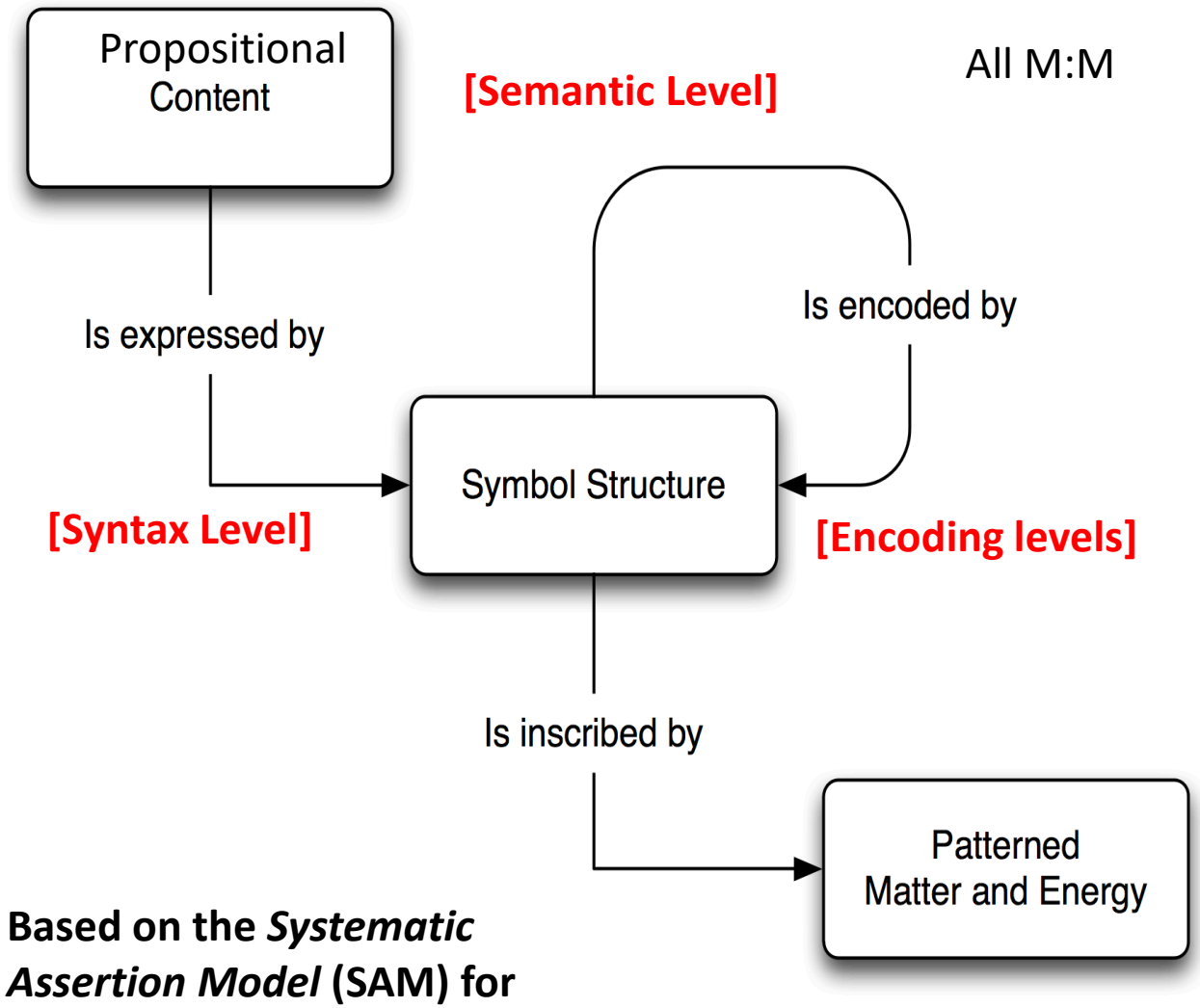
We need identifiers for lots of things:

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

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

(EVERYTHING!)

So what are we identifying?



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*…

W*hat* are we identifying?

Propositions	Semantic level	
Symbols	Syntax level	
Symbols	Encoding level (1)	
Symbols	Encoding level (2)	
Symbols	Encoding level (3)	
. . .		
Symbols	Encoding level (n)	(e.g. 1s&0s, for inscription)

But operationally identification works from the bottom up.

We identify the bitstream or character sequence in a normal form in order to *indirectly* identify the higher level encodings, syntax, or propositional content.

Identifiers for abstraction levels

The general problem:

The things we want to identify can be expressed or encoded in different ways,

We are typically looking at a 1:M* relationship between

- Propositions and representations
- Representations and encodings
- Encodings and encodings and encodings . . . [lather, rinse, repeat]

and so we can't easily use the lower level instantiations
to determine identity of the upper levels entities

[in other words: variation at the lower level
does not necessarily entail variation at the upper levels]

But, there is a solution to this problem.

(in the next video)

*Actually the relationships are M:M because of the arbitrary nature of representation described in the Data Concepts videos, that is: depending on conventions (like standards) and other social circumstances (agreements, decisions, intentions, expectations) the same (e.g.) encoding can encode multiple representations. But most of the time in identity problems we can ignore this and rely on a shared understanding of the relevant concepts.

Canonicalization

Canonicalization is a technique for determining representational identity

and is a reasonable proxy for propositional identity.

In short:

1) If necessary convert to the same representation language and encoding
2) Normalize incidental variations
3) Test resulting files for byte level identity.

Standards. . .

Why standards?

Standards are fundamental in data curation

standards promote reliable efficient communication,
with others now, and with ourselves in the future.

Supporting integration, interoperable tools, validation, authentication, preservation, regulatory compliance, . . . etc. etc. etc.

Data format conformance vs processor conformance

Data standards can define conformance for both (i) data and (ii) processing

Data set conformance might require such things as

- a particular character encoding
- particular delimiters
- serialization matching a particular formal grammar
- constraints such as referential integrity, data types
- inclusion of relevant metadata

Processor conformance might require the processing software to

- correctly tokenize
- verify that statements match a particular grammar
- perform additional validation (referential integrity, data types, etc)
- confirm required metadata
- process data sets correctly,
  - e.g., performing particular actions, such as generating a normalized parse tree,
  - or performing calculations, rendering visualizations, etc.
  - displaying an error when processing non-conformant datasets
    - [the processor may also be required to halt, or it may be allowed to continue

Processing conformance may be tested by a specified “test suite” of data sets
Data set conformance is tested by validating softward.