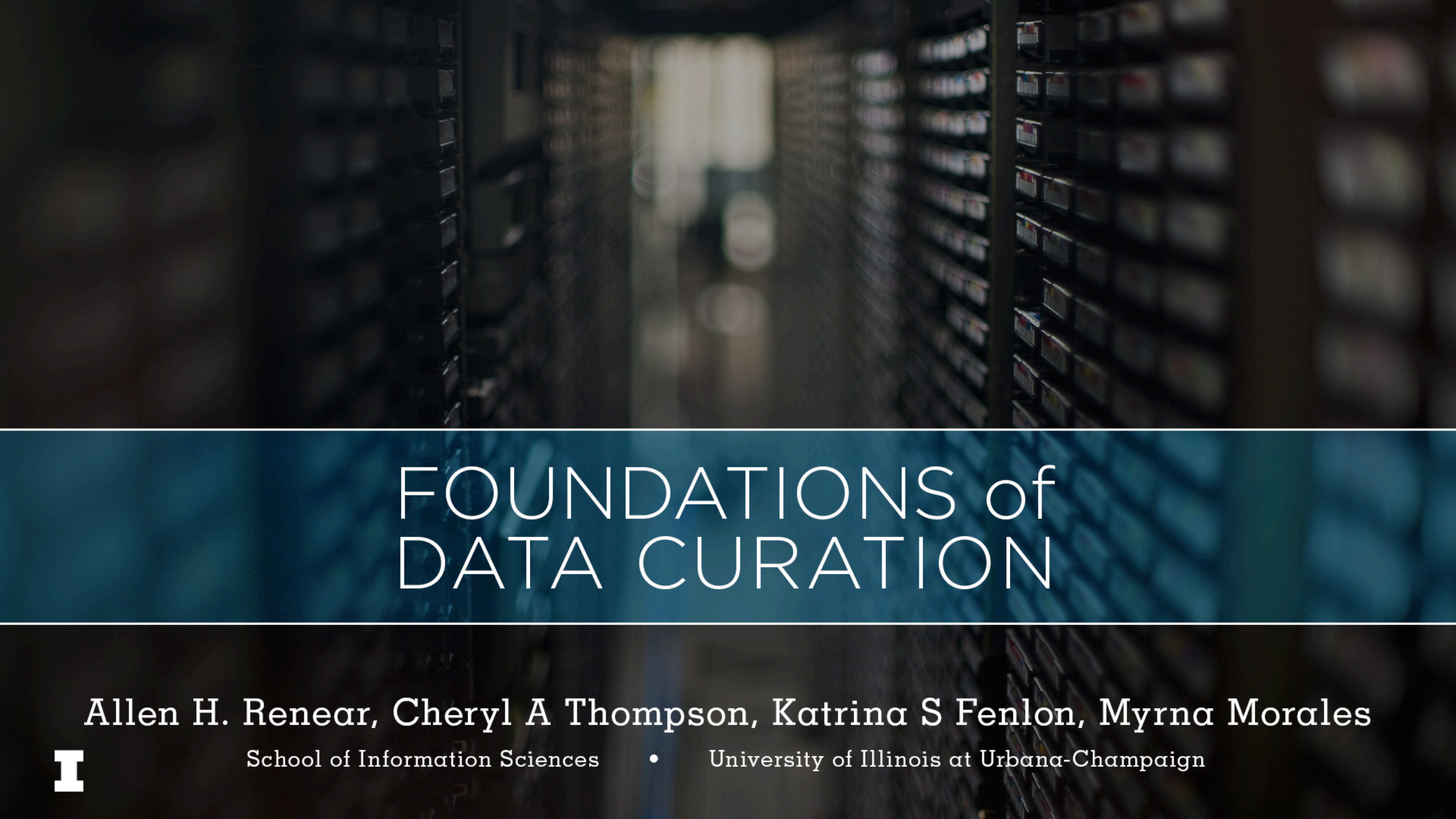
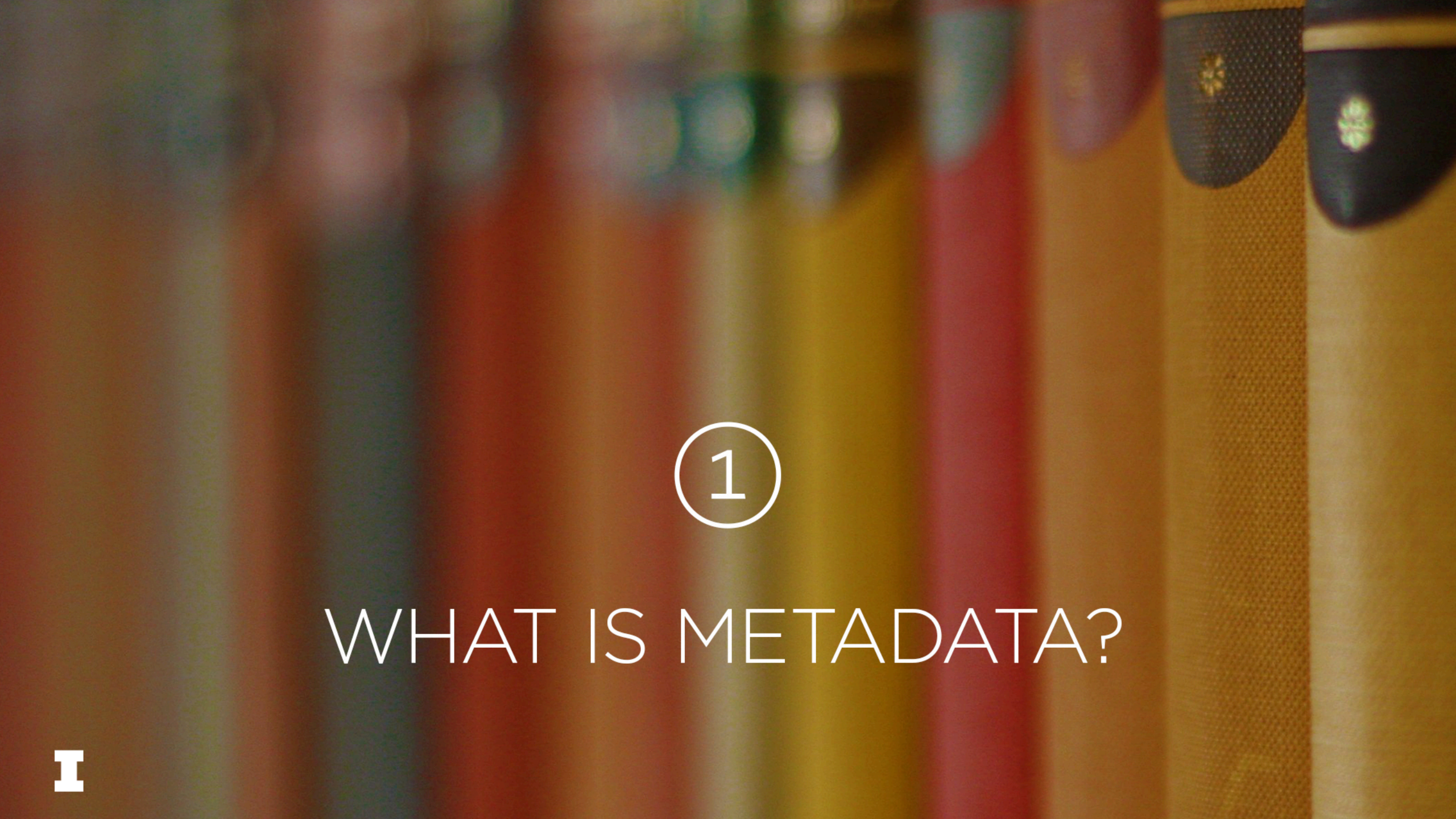


FOUNDATIONS of DATA CURATION

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

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


1
WHAT IS METADATA?

What is metadata?

Defining metadata (a simple definition)
Some examples of metadata data
Defining metadata (a better definition)
What is metadata *for*?
The standard classification of metadata: descriptive, administrative, structural
Metadata vs data

What Is Metadata? [First Definition]

The simple, and most common, colloquial definition is:

data about data

What is metadata?
[Information that might be metadata]

Metadata for a data set of temperatures on the surface of the earth at some time might include:

- nature of data (here: temperatures on surface of the earth)
- location relevant to data application (e.g. a 3D latitude, longitude, altitude box)
- when the data was recorded and where the recording equipment was located (maybe in orbit)
- what equipment was used, along with what settings and calibrations
- the data format and schemas (semantics, syntax, encoding); any standards being used
- version history (with who, when, what, why, for changes)
- input data sets and algorithms involved in deriving this data set (if not raw data)
- checksum or other fixity signature
- identifier (located in system reflecting format and content change history).
- organization responsible, and perhaps owns the data or copyright
- restrictions on use (legal or local policy)

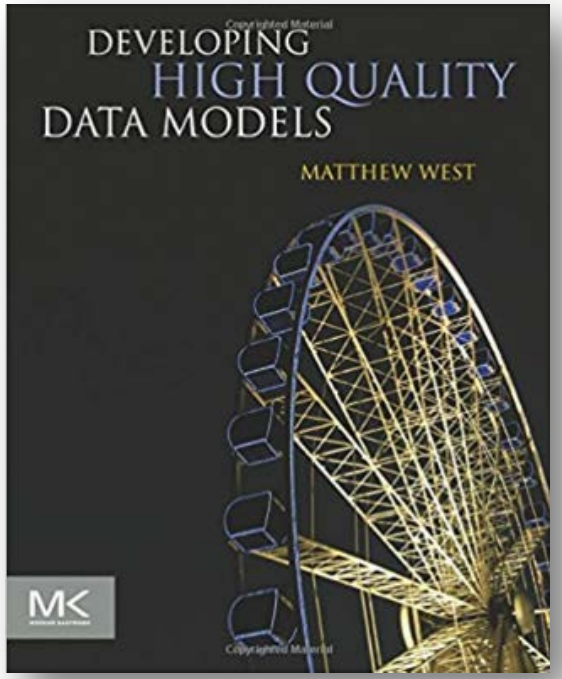
and so on

The first thing to notice here is the extremely varied nature of this information,
and the similar variation in the implied purposes.

Example – Dataset Metadata

- <origin>USGS Alaska Science Center, 4210 University Drive, Anchorage, Alaska 99508</origin>
  <pubdate>20101231</pubdate>
  <title>Catalogue of Polar Bear (Ursus maritimus) Maternal Den Locations in the Beaufort Sea and Neighboring Regions, Alaska, 1910 – 2010</title>
  <geoform>Tabular Digital Data</geoform>
  … <abstract>This report presents data on the approximate locations and methods of discovery of 392 polar bear (Ursus maritimus) maternal dens found in the Beaufort Sea and neighboring regions between 1910 and 2010 that are archived by the U.S. Geological Survey, Alaska Science Center, Anchorage, Alaska. ….</abstract>
  …
- <begdate>1910</begdate>
  <enddate>2010</enddate>
  …
- <descgeog>Beaufort Sea and Chukchi Sea of northern Alaska, Canada, and Russia</descgeog>
  <bounding>
  <westbc>178.2167</westbc>
  <eastbc>-178.9167</eastbc>
  <northbc>83.921</northbc>
  <southbc>63.3667</southbc>
  </bounding>
- https://www2.usgs.gov/datamanagement/documents/USGS_ASC_PolarBears_FGDC.xml

Example — Bibliographic Metadata



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

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*}

What is metadata for?

"structured data … that **supports functions …**"

Mostly human-oriented functions	Mostly machine-oriented functions
Find potentially relevant data	Read data with appropriate software
Determine relevance [e.g., understand exactly what the data includes and excludes]	Visualize and display data
Understand and interpret data	Analyze data
Assess data quality and integrity	Integrate data from different sources
Authenticate data	Convert or migrate data
Avoid inappropriate use	Organize data
Etc.	Etc.

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

Metadata vs Data

First, metadata *is* data

Here's some metadata, obviously data

Data Set	Ontology	File Format	Status	Lead
DS4501	PROT 42.0	PROT-RDF 42.0	On deck	Kristof
DS4502	PROT 42.0	PROT-RDF 42.0	Released	Tzikas
DS4503	PROT 42.3	PROT-JSN 42.3	Embargoed	Kipper

Data Set	Location
DS4501	ox.ac.uk/files/4521
DS4502	ox.ac.uk/files/9883
DS4503	ox.ac.uk/files/8664

A relational data base that combine metadata and data set locations.

[hmm . . . metadata is data . . .
. . . and so will have its own metadata . . .
. . . which will have its own metadata . . .
. . . which . . . (etc).

In a slogan. . .

One person's metadata is another person's data

So all metadata is data,
but what makes some, and only some, data metadata?

What data is metadata and what data is not?

Data point: Temperature is 31.5.

Information that might support the use of this data point:

Temperature of what?
What is the unit?
Collected when?
For what purpose?
Etc.

Temperature	Unit	Instrument
31.5	celsius	ACMEtherm

<ex:temp @unit="celsius" @Instrument="ACMEtherm">31.5</ex:temp>

But the instrument identification might have been metadata on the entire dataset;
and the unit designation might have been an metadata on a schema

Often the distinction is pragmatic

Suppose that a process was generating data sets by making 10,000 observations all at the same place but over an interval of time.

> We would probably treat *time as data* (e.g., a column if the data set is relational), and treat *place as metadata* attached to the dataset.
>
> > But if the example is reversed for time and space (a single point in time but varying locations in space) we would probably treat *time as metadata* and *space as data.*
> >
> > > (Here the motivation is at least in part reducing complexity and avoiding update anomalies)
>
> And if we anticipated integration with records where both time and space information varies we would probably represent *both* time and space data *as data*, i.e., with two separate columns in the table.

But is the distinction always pragmatic?

Perhaps some features are essentially data about data:

For instance:

- Value related features
  - Accuracy specifications (±)
  - Datatypes
  - Value constraints
  - Notation system
  - *etc.*
- Data set features
  - Size of data set
  - Coverage of data set (time or space intervals)
  - Schemas
  - *etc.*

More differentiation problems

Are these things metadata (in red)?

<!ELEMENT anthology (poem+)>
<!ELEMENT poem (title?, stanza+)>
<!ELEMENT title (#PCDATA)
<!ELEMENT stanza (line+) >
<!ELEMENT line (#PCDATA) >

<geoform>Tabular Digital Data</geoform>
<title> Catalogue of Ursus maritimus Maternal Den Locations</title>
<begdate> 1910</begdate>
<enddate>2010</enddate>
<descgeog> Beaufort Sea and Chukchi Sea of northern Alaska, Canada, and Russia</descgeog>
<bounding>
<westbc>
178.2167</westbc>
<eastbc>-178.9167</eastbc>
<northbc>83.921</northbc>
. . .

Data Set	Location
DS4501	ox.ac.uk/files/4521
DS4502	ox.ac.uk/files/9883
DS4503	ox.ac.uk/files/8664

Again. . .

Some sorts of information are considered to be classic metadata
but when you look closely it appears that the data/metadata distinction
is typically based on practical considerations
and not a clear hard distinction

Nevertheless:

Some metadata is seems clearly about data
(e.g. accuracy, datatype, notation, etc.)

And *some* clearly about data sets
(e.g. size, coverage, owner, model & encoding features etc.

And so some metadata appears to be metadata in a strict sense


2
METADATA SCHEMAS

Metadata Schemas

Metadata schemas
Vocabulary independence
Syntax/serialization independence
Mixing and matching metadata schemas
Examples of schemas

Metadata Schemas

A set of data elements, with specified meanings

for supporting metadata statements in particular contexts

Sometimes vocabulary independent*

Often syntax/serialization independent**

Pure metadata schemas are *conceptual*

*Sometimes vocabulary independent

Metadata elements in a metadata schema can be purely conceptual, without a controlled vocabulary.

By allowing different vocabulary terms for the same concept the schema can more gracefully support different languages, as well as avoid, or take advantage of, common meanings for common words.

However advantages of having a level of indirection between concept and term may be outweighed by the advantages of providing a controlled term vocabulary, and so avoiding unnecessary variation.

So many metadata schemas specify a controlled vocabulary as well as conceptual elements; users who wish to use an alternative vocabulary for the same concepts can provide a mapping.

Pure metadata schemas are *conceptual*

**Often syntax/serialization independent

Similarly metadata schemas need not specify any particular syntax for applying specified concepts.

In this case most metadata schemas do strictly separate the conceptual schema from the variety of options for applying metadata to objects.

Given the variety of contexts (data models, file formats etc.) in which metadata will be applied allowing metadata statements to implemented in different serialization syntaxes is profoundly useful, and essential for wide adoption.

Recommended serialization syntaxes can be developed and standardized independently.

e.g., Dublin Core

http://www.dublincore.org/documents/dcmi-terms/

- 15 elements for describing resources on the web
- With defined semantics and recommended vocabularies for elements

Contributor	Format	Rights
Coverage	Identifier	Source
Creator	Language	Subject
Date	Publisher	Title
Description	Relation	Type

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">

Combining, specializing, and extending metadata schemas

Metadata schemas can be combined, specialized, and extended in various ways.

For example, a defined application of Dubin Core,

- may limit the values of the Dublin Core *dc:format* element to the controlled vocabulary for IANA media types (MIME types).

or

- specify the refinements *spatial* and *temporal* for dc:coverage

To ensure graceful processing of adaptations techniques for schema modification or other notifications may also be specified.

Take a look at these standard metadata schemas

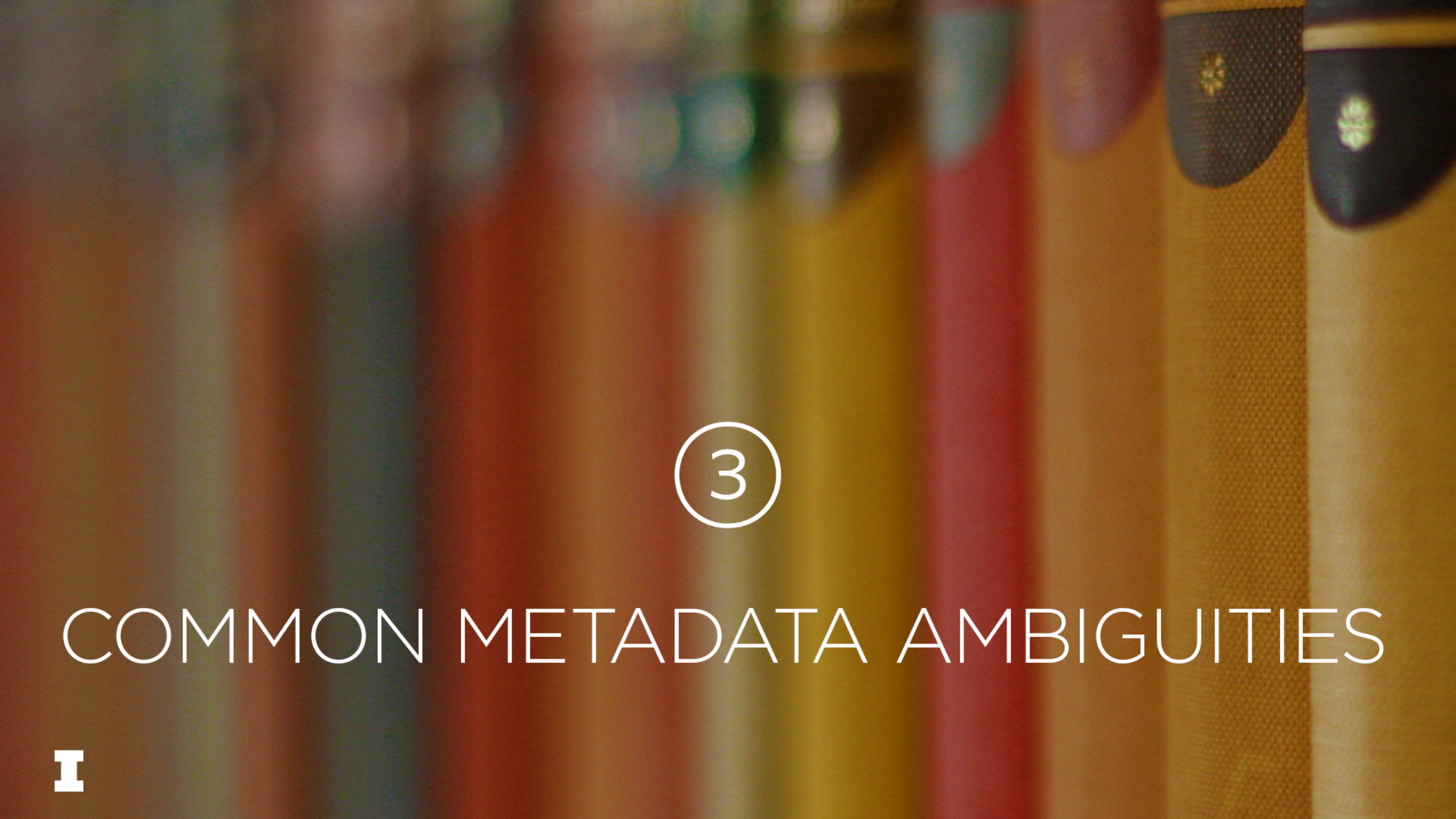
Lists of Metadata Schemas:

Biodiversity: http://www.dcc.ac.uk/resources/metadata-standards/list

Libraries and museums: http://jennriley.com/metadatamap/seeingstandards_glossary_pamphlet.pdf

Other areas: https://en.wikipedia.org/wiki/Metadata_standard

Individual research groups, businesses, and other organizations may develop custom metadata schemas for private or limited use; usually these are specializations of existing schemas.



3

COMMON METADATA AMBIGUITIES

Common Metadata Ambiguities

Metadata is often casually used, with ambiguous semantics

And so many problems ensue, especially for robots.

It may not be clear (to a robot) what an attribute/value pair is *about*

Though it is obvious enough to humans

Attributes themselves can be, to robots, problematically ambiguous

Even when perfectly clear to a human

Yes, humans can figure things out

Ok, but that's not safe, and *it is not fair to robots!*

To get your intuitions going . . .

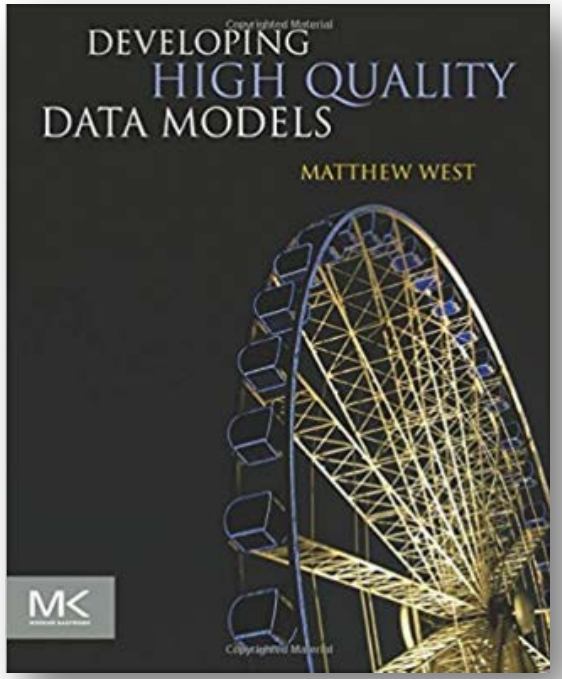
In a collection of digital images of paintings; it is not uncommon to see, attached to the same image:

. . .
ID: BNF4242a
. . .
Artist: Leonardo da Vinci
. . .
Format: JPEG
. . .
Format: oil on panel

What's wrong with this?

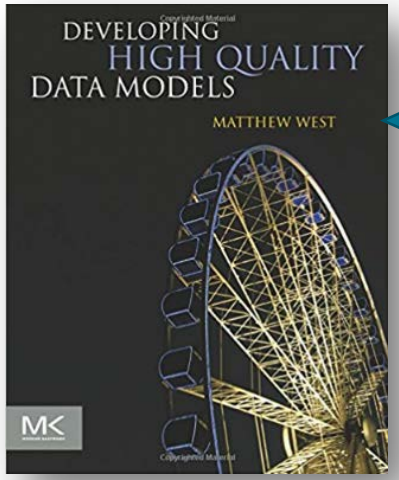
Adapted from work by Richard Urban

Recall: Bibliographic metadata



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

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



title: Developing High Quality Data Models
date: 2011
publisher: Morgan Kaufmann
subject: database design
subject: data structures (computer science)

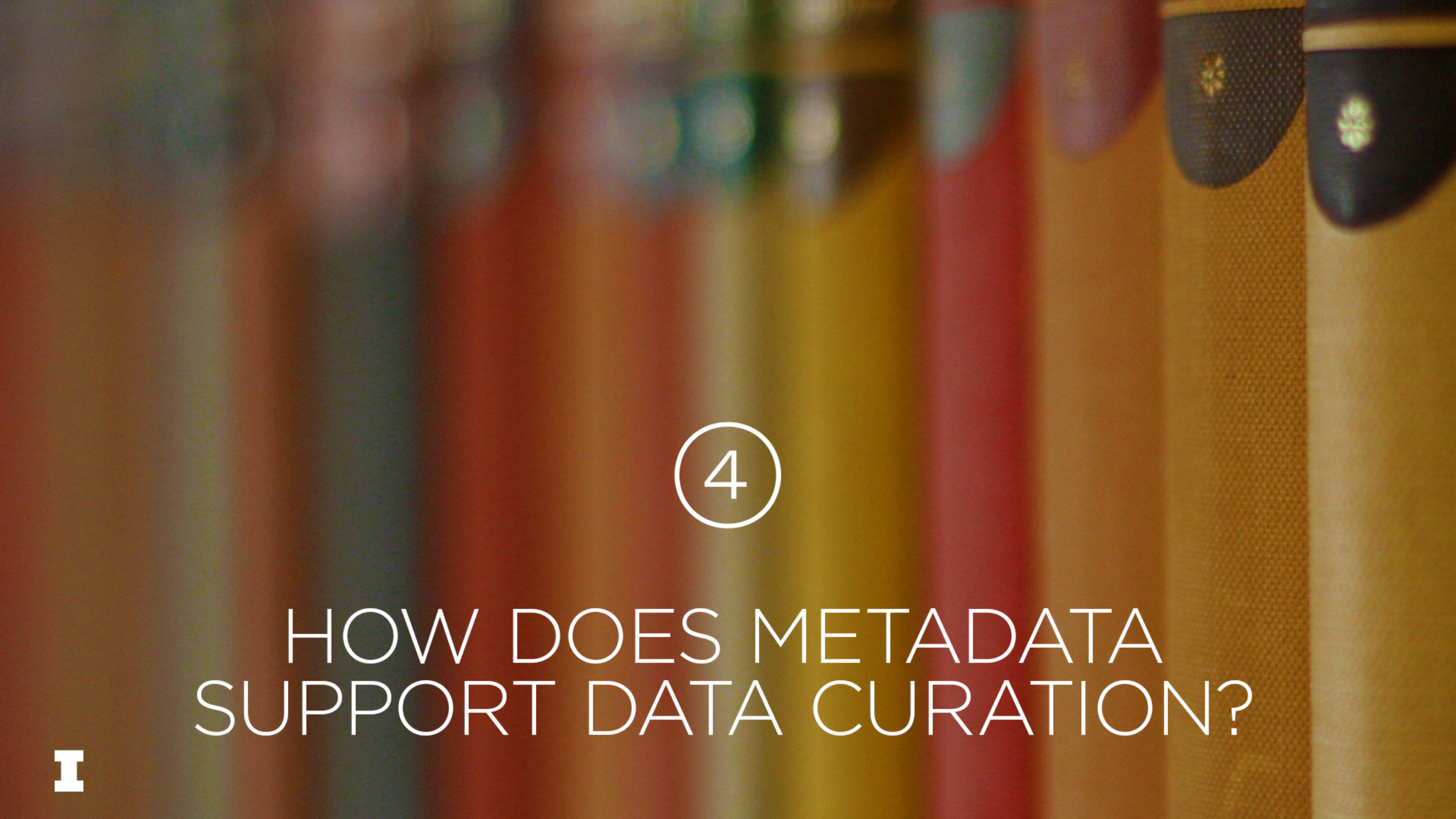
date: 2011
publisher: Morgan Kaufmann
subject: database design
subject: data structures (computer science)
language: English
pages: 389

A converse problem: two meanings for same attribute

For a while this was common practice

DC:Identifier=Book2424a	DC:lang=EN	to indicate a book in English
and		
DC:Identifier/Coll99LB	DC:lang/EN	to the books in a collection are in English

What's wrong with this?


4
HOW DOES METADATA
SUPPORT DATA CURATION?

How Does Metadata Support Data Curation?

Relating metadata to curation *objectives*, *activities*, *actions*

Recall the data curation objective

Data curation is concerned with all aspects of the management of data

in order to efficiently and reliably support the analysis of data,
and enable reuse over time

Recall: Areas of curatorial activities

Collection:	Support the collection and acquisition of data
Organization:	Employ an appropriate data model and use appropriate standards
Storage:	Support reliable and effective storage
Preservation:	Ensure that data will be understandable and useable in the future
Discoverability:	Support the ability to search for and locate relevant data
Access:	Support the ability to retrieve and distribute data
Workflow:	Support the ability to systematize data workflows
Identification:	Support the ability to identify, authenticate, and validate data
Integration:	Support integration of data from different sources using different data models
Reformatting:	Support reformatting for use by different tools or to match new format standards
Reproducibility:	Support ability to reproduce results, ensuring scientific validity
Sharing:	Support sharing data between researchers, teams, and institutions.
Communication:	Support representation, publishing, and visualizations that provide insight
Provenance:	Support identifying what inputs and calculations are responsible for data values
Modification:	Support management of corrections and updates
Compliance:	Ensure compliance to legal, regulatory, and local policy requirements
Security:	Ensure that data is secure from tampering or inappropriate access and distribution

Recall: Methods of curatorial action

Analysis
To determine needs, and develop relevant data models and *metadata*, and reformat, correct, or update data.

➔ **Documentation**
To record essential information (typically via *metadata*)

System design and implementation
To support all data curatorial activities
To support the generation and use of data documentation and processing documentation

Policy
To specify objectives, procedures, practices, and formats.

Process
To ensure success and efficiency by managing the development of appropriate organizational units and roles, providing training, advocating for change, and managing curatorial activities.

➔ Metadata is rigorously defined machine-processable documentation

Areas of curatorial activities (Part I: the core)

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]
Discoverabilty	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]

Areas Of Curatorial Activities (Part II: Partial Dependencies)

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]

Frictions In Creating And Managing Metadata

Metadata Friction Categories	Specific frictions
Standardization	• Unfinished and multiple metadata versions • Proliferation of metadata standards
Temporal	• Metadata in the life cycle • Timeframe for metadata use
Data Sharing	• Audience for metadata • Individual vs. group knowledge
Human Support	• Metadata is nobody's job • Proliferation of metadata tools

Metadata frictions specific to the "problems and conflicts that must be addressed to make metadata useful"

Adapted from Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). Retrieved from http://dl.acm.org/citation.cfm?id=1940818