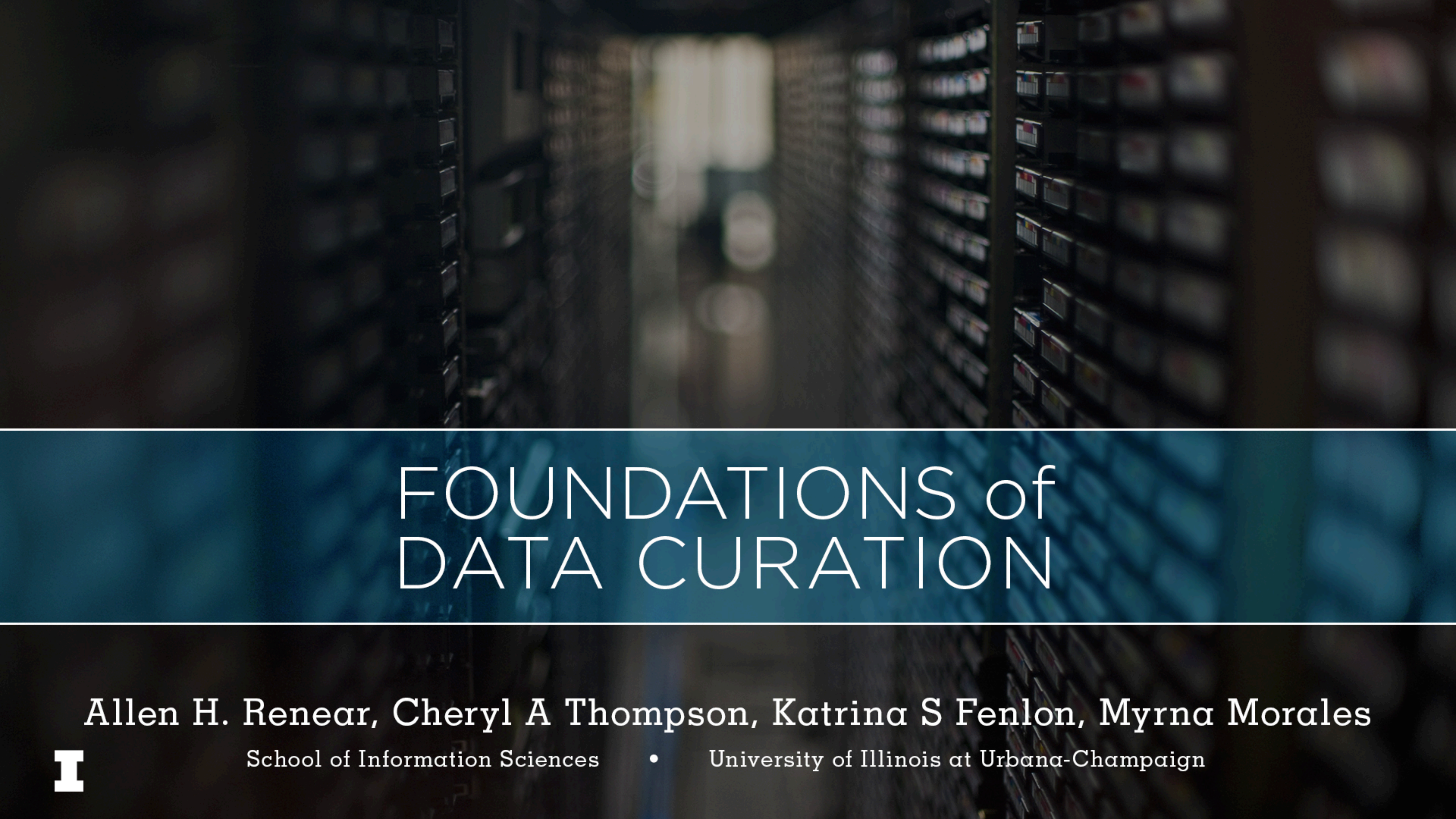
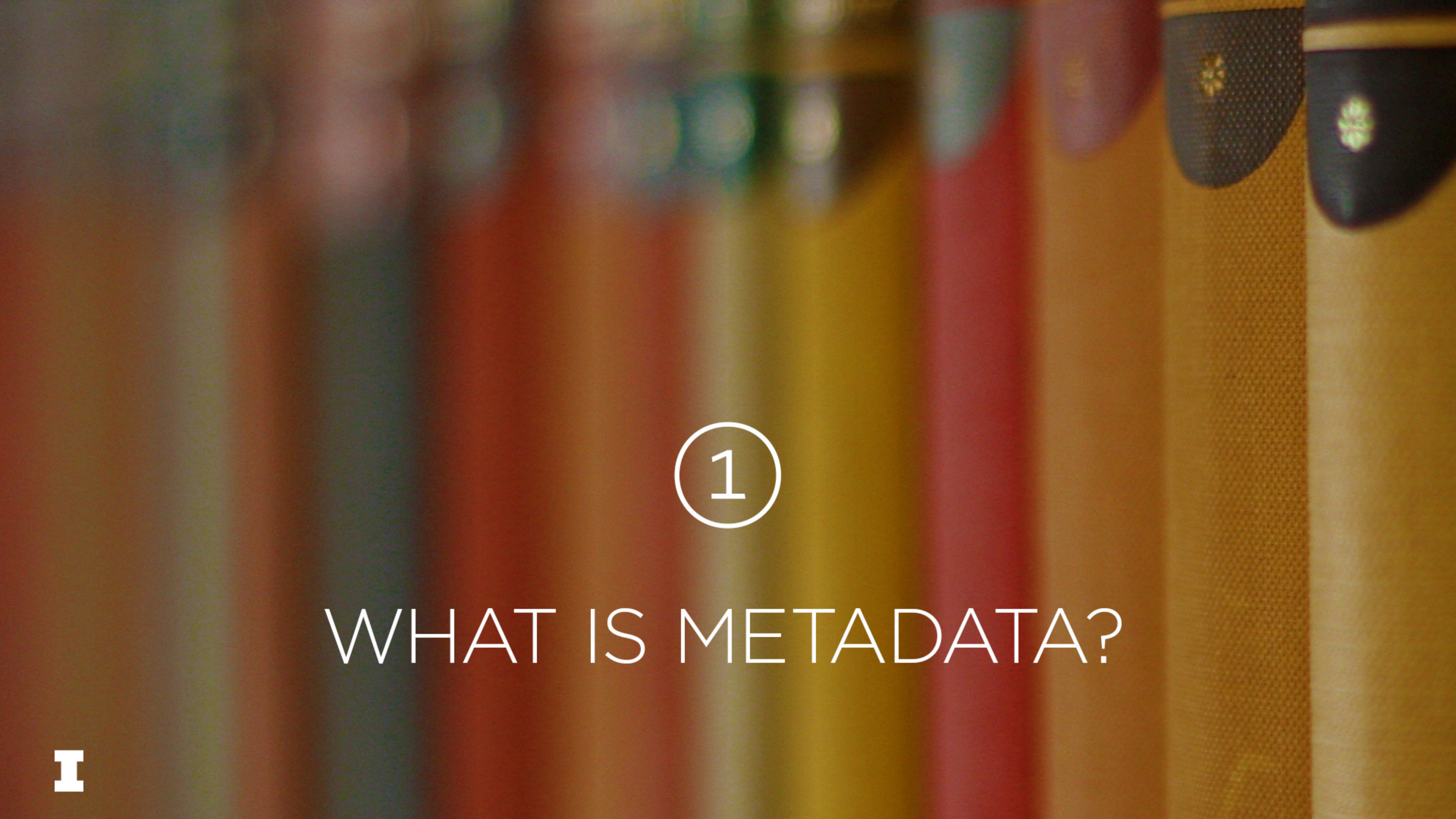


FOUNDATIONS of DATA CURATION

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

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

METADATA



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)

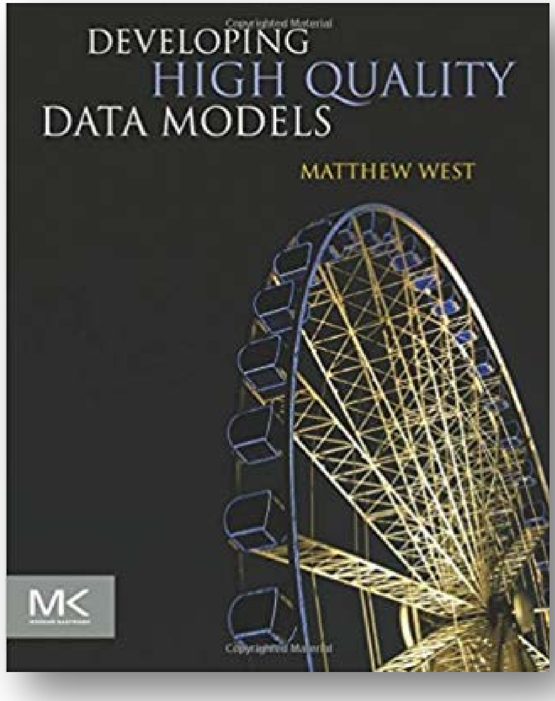
  and so on

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	For decoding and rendering
Preservation	For long-term management
Rights	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

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.