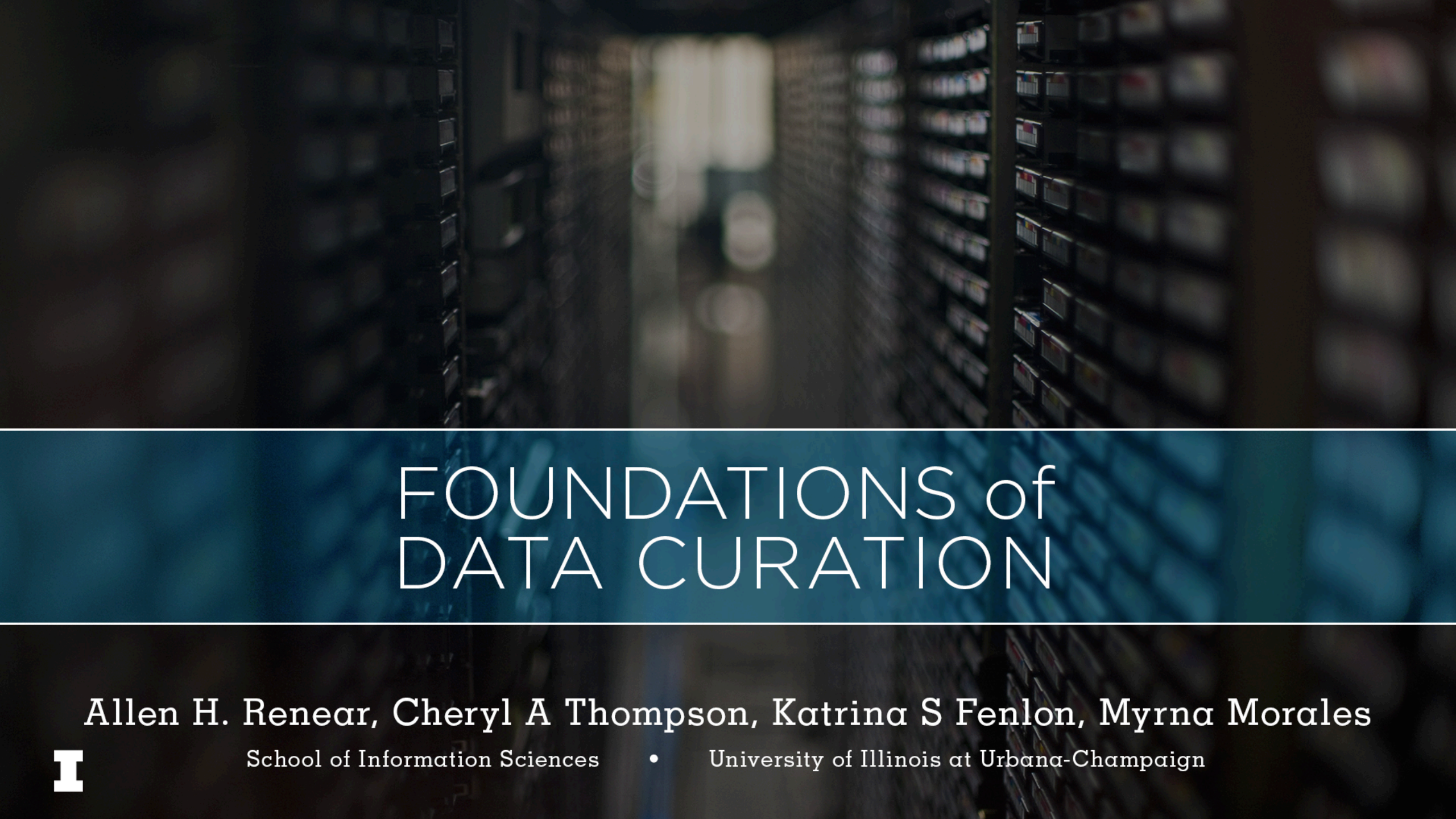
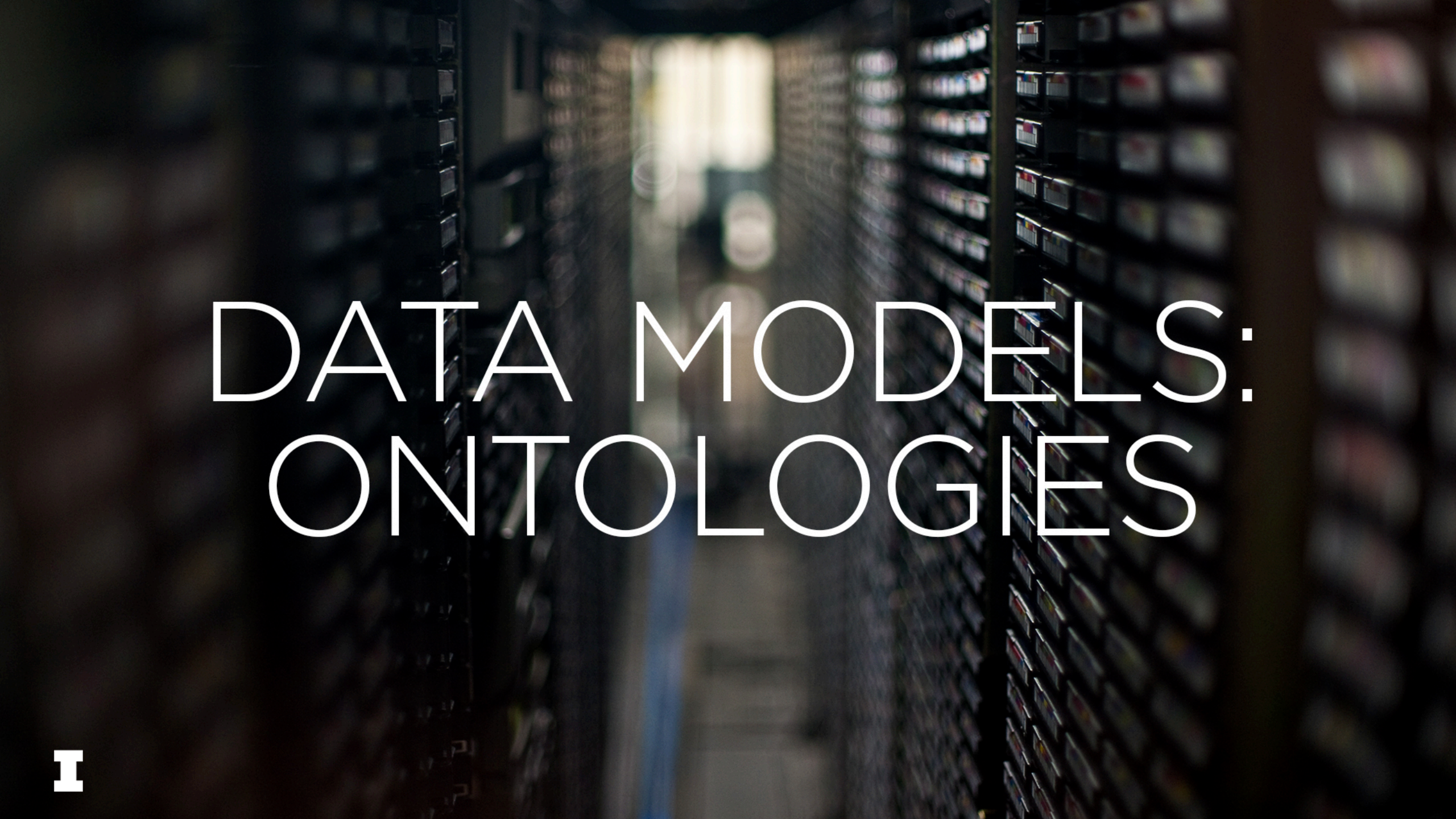
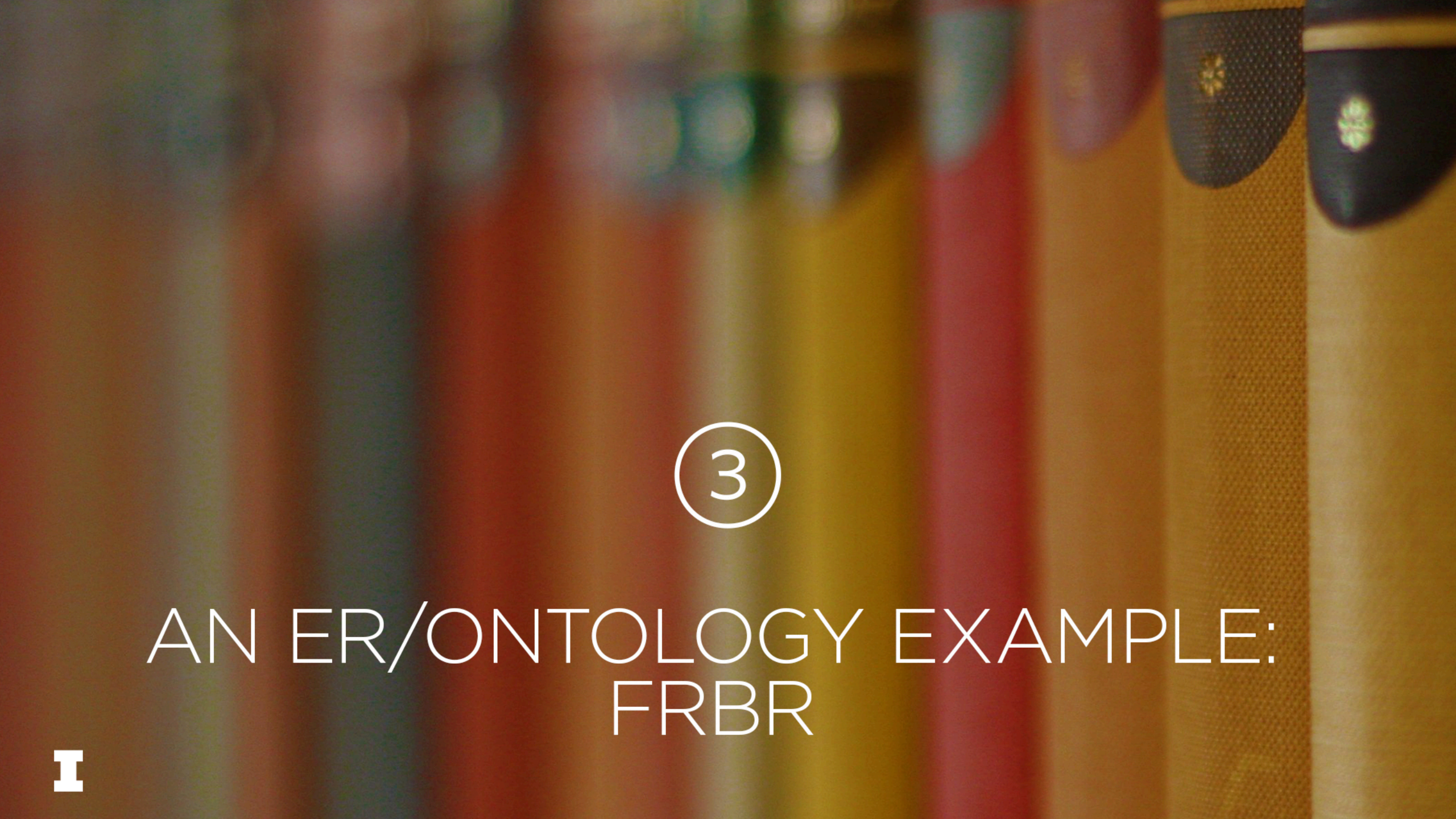


FOUNDATIONS of DATA CURATION

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

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


DATA MODELS:
ONTOLOGIES



3

AN ER/ONTOLOGY EXAMPLE: FRBR

An ER/Ontology Example: FRBR

- We will show how a fragment of the FRBR conceptual model was developed in order to support the complexities of library cataloguing.
- Although the example is from library cataloguing the solution is directly related to problems in data curation and digital asset management, as we will see.

What is a *book*?

Did you read the same book I did?

This can mean different things:

the same novel I read?	Yes, but the French translation (text)
the same text I read?	Yes, but but the large type edition
the same edition I read?	Yes, but the copy in the summer house
the same copy I read?	The one with the mustard stain

FRBR: Functional Requirements for Bibliographic Records

The ordinary “book” can be simultaneously

about Manet,
in French,
typeset in neo-Bauhaus,
mustard-stained

Which is, strictly speaking, impossible.

so FRBR replaces the book with four kinds of entities:

the work	is about Manet,
the expression	is in French,
the manifestation	is typeset in neo-Bauhaus,
the item	is mustard-stained

The FRBR Group 1 Entity Types

There are four Group 1 entity types

Work: “a distinct intellectual or artistic creation”

Expression: “the intellectual or artistic realization of a work in the form of alphanumeric, musical, or choreographic notation, sound, image … etc.”

Manifestation: “the physical embodiment of an expression of a work”.

Item: “a single exemplar of a manifestation”

Or, colloquially (for book-like objects): work, text, edition, copy

Illustrative attribute assignments

Each entity type is assigned a *distinctive* set of attributes…

- works have attributes such as *subject* and *genre*
- expressions have attributes such as *language*
- manifestations have attributes such as *typeface*
- items have attributes such as *condition* and *location*.

NB: these attribute assignments are *disjoint*

A work may have a *subject*.
It does not have a *language*, *typeface*, or *condition*.

An expression may have a *language*;
It does not have a *subject*.

A manifestation may have a *typeface*.
It does not have a *subject* or a *language*

An item may have a *condition*
It does not have a *subject*, *language*, or *typeface*.

Foundations

FRBR…

•generalizes current best practice in cataloguing and bibliographic control

•reflects an emerging theoretical consensus

and longstanding bibliographic thinking, e.g.,

Seymour Lubetzky: work vs book

Eva Verona: literary unit vs bibliographic unit

See also Patrick Wilson, Elaine Svenonius, Akos Domanovszky, Martha Yee, et al.

•refines and extends current notions,
proposing new concepts based on new analysis
and anticipating new formats and new technologies

FRBR: Adoption and Influence

Don’t be misled by the example of a traditional paper book, FRBR is not only a major event in cataloguing and technology in libraries

But it is influential in developing ontologies for digital asset management and data curation:

cf. “When a scientist asks ‘what data do you have?’ …”

—Joe Hourcle (NASA/SDAC), “FRBR in a scientific context” 2007

As we will see

The method (simplified), in two steps.

The first step in the entity analysis technique is to:

1) isolate the kinds (types) objects (entities) that are of interest.

   Don't focus on data but on the things the data describe.

   Each entity type becomes a focal point for a cluster of data.

Once this high-level structure for the model has been charted

2) identify the important attributes of each entity.
   - identify relationships between entities
   - identify attributes and values

.

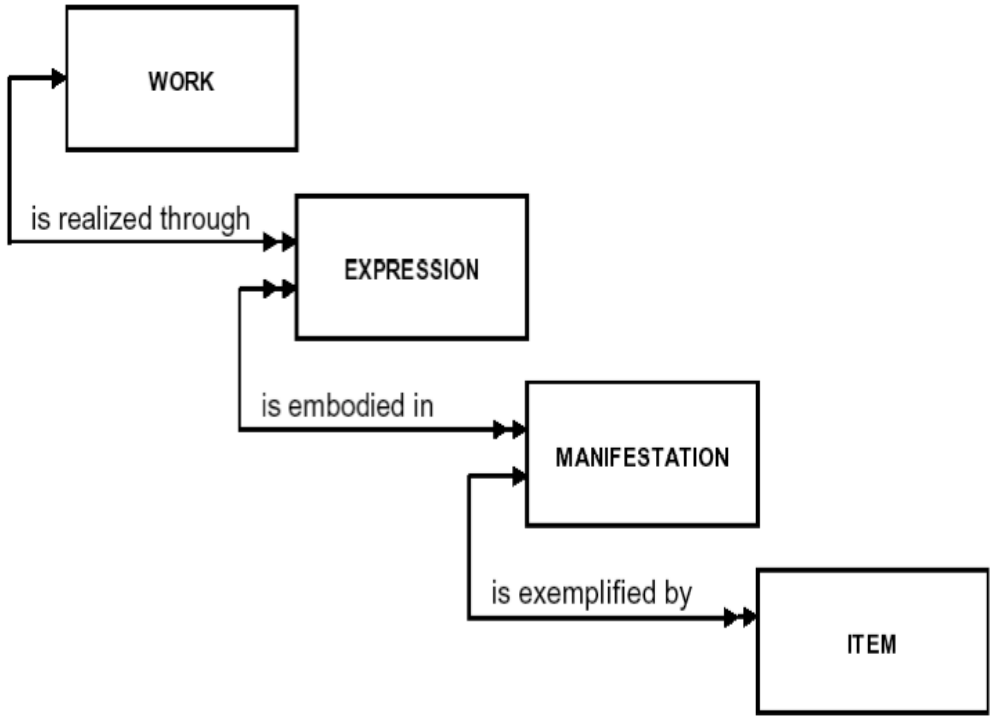
Relationships between the entities types

The novel Moby Dick (a work)

-- is realized through many different expressions (the particular texts of Moby Dick)
(the text of the 1859 edition, the corrected edition, the Murray French translation…)

-- each of which maybe embodied in many various manifestations
(the 1859 edition, a microform of that edition, an ebook with the 1859 text…)

-- each of which may exemplified by many different items
(my copy, the Houghton copy, the Beinecke copy, the Ransom copy)



Overview

Entities

- Group 1: Work, Expression, Manifestation, Item
- Group 2: Person, Family, Corporate Body
- Group 3: Concept, Object, Event, Place

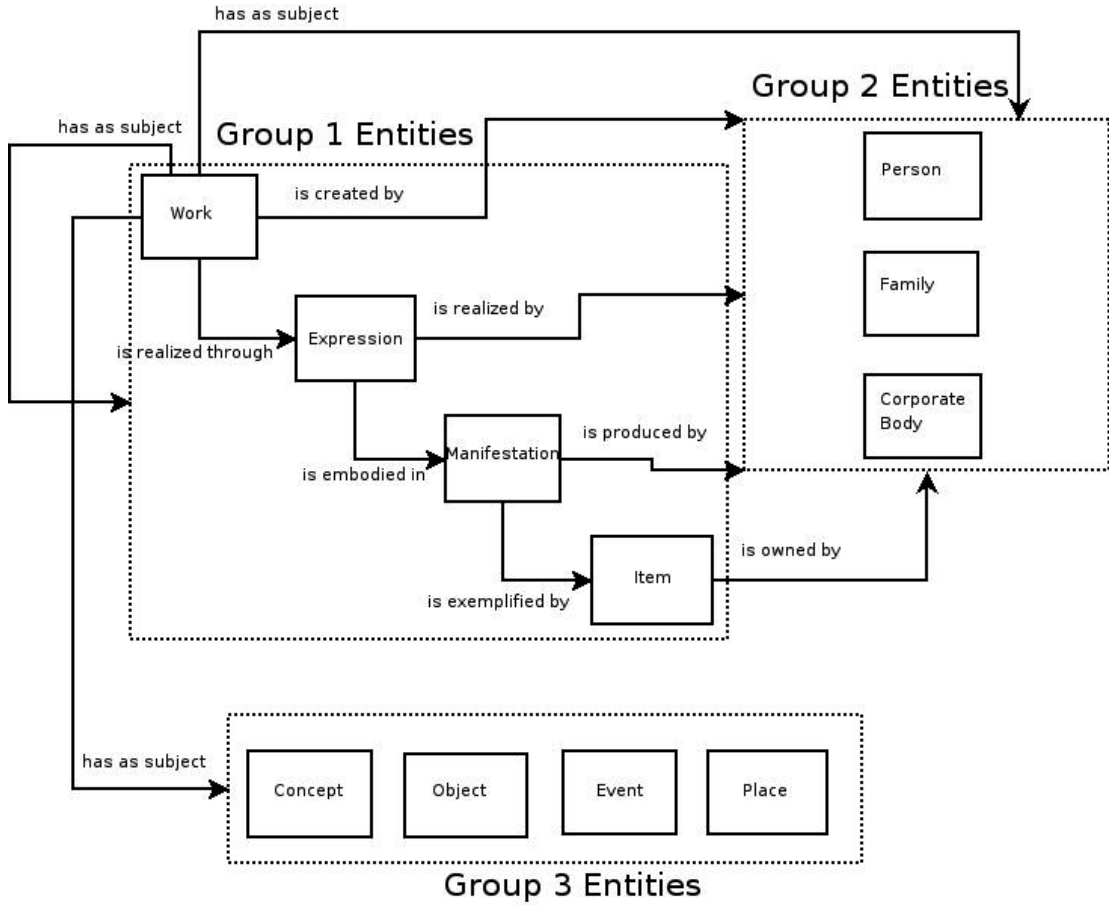
Attributes

- Clusters of appropriate attributes for each entity.
- Some “inherent”, other “externally imputed”

Relationships

- Some “integral to the definition of the entities”: e.g. exemplifies, embodies, realizes. Others are not: adaption, translation, abridgement, part of.
- Some exists between entities of the same type, others between entities of different types.

The FRBR Entity Groups



Not shown:
-- attributes
-- lower-level relationships

At http://www.frbr.org/files/entity-relationships.png; unattributed, pls send email to renear@uiuc.edu if you know the attribution.

FRBR as an ontology (using FOL) (work with Karen Wickett)

Definitions

work(x)		… x is an artistic or intellectual creation
expression(x)	=df	(∃y)[realizes(x,y)]
manifestation(x)	=df	(∃y)[embodies(x,y)]
item(x)	=df	(∃y)[exemplifies(x,y)]

Primitives:

work(x), realizes(x,y), embodies(x,y), exemplifies(x,y)

Cardinality Axioms

(realizes(y,x))	→	[(∀z)(realizes(y,z) → (z=x)]
(exemplifies(y,x))	→	[(∀z)(exemplifies(y,z) → (z=x)]

Domain/Range Axioms

realized(x,y)	→	work(y)
embodies(x,y)	→	expression(y)
exemplifies(x,y)	→	manifestation(y)

Disjointness Axioms

expression(x)	→	~ [work(x) v manifestation(x) v item(x)]
manifestation(x)	→	~ [work(x) v expression(x) v item(x)]
item(x)	→	~ [work(x) v expression(x) v manifestation(x)]

Theorems

realizes(x,y)	→	~(∃z)[embodies(x,z) v exemplifies(x,z)]
embodies(x,y)	→	~(∃z)[realizes(x,z) v exemplifies(x,z)]
exemplifies(x,y)	→	~(∃z)[realizes(x,z) v embodies(x,z)]

all relationships are now irreflexive and asymmetric; and trivially, transitive.

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.