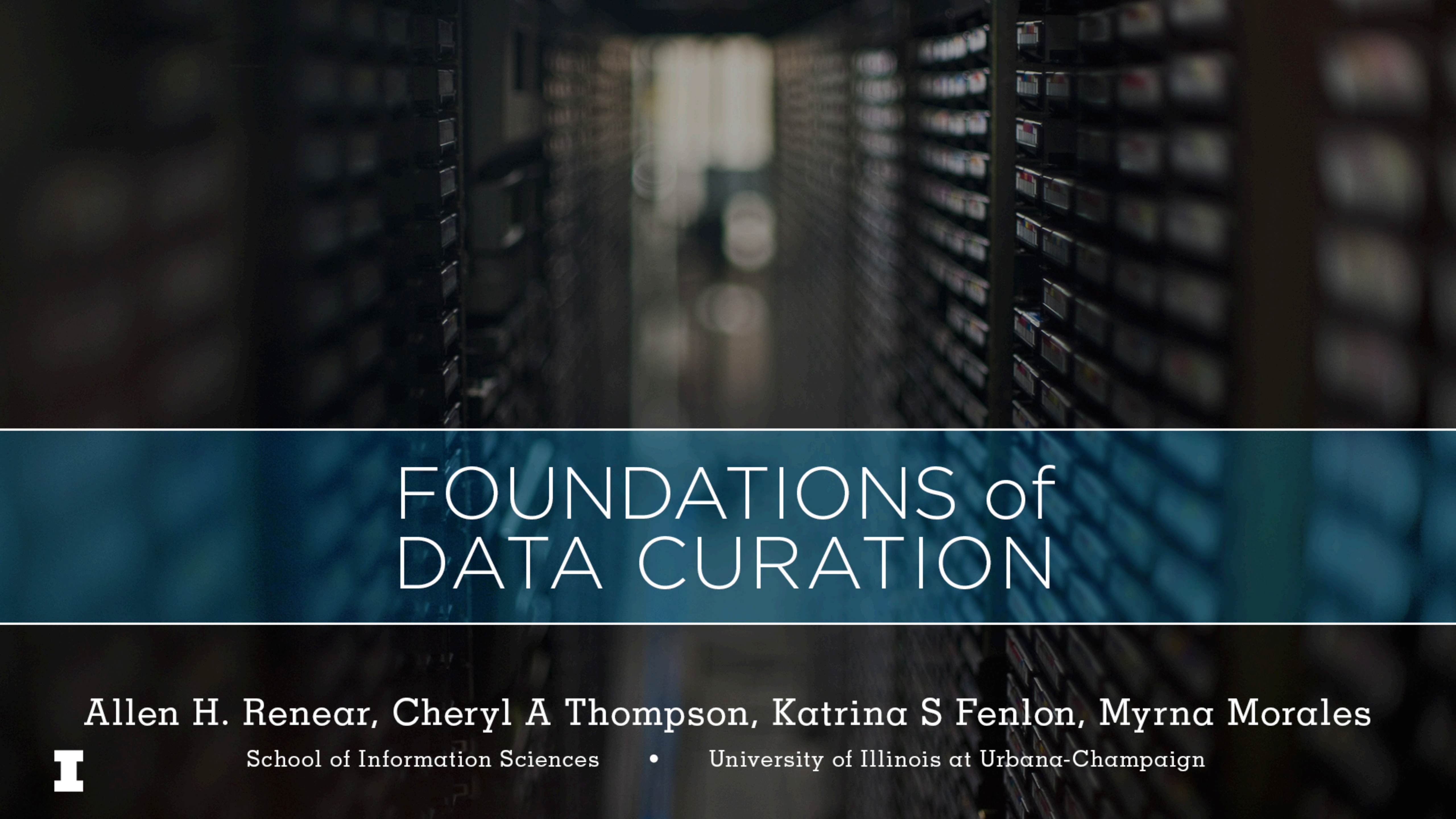


FOUNDATIONS of DATA CURATION

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

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

DATA MODELS: RELATIONS

①

DATA MODELS

Data Models

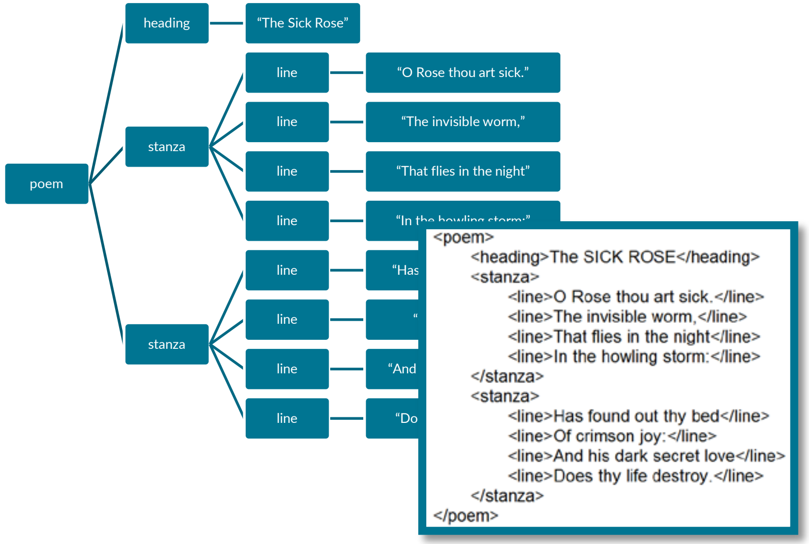
- What is a data model?
- Some examples of data models
- Why data models are important to data curation
- Towards an integrated picture of data model relationships

Some data models you know and love

Relations

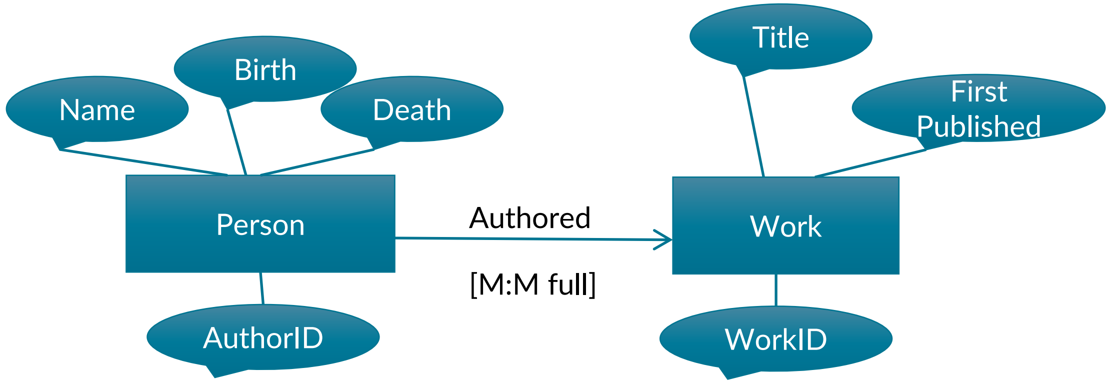
Work	Author	Title	Date
W58425	P42425	Moby Dick	1851
W85246	P24246	The Scarlett Letter	1860
W55427	P24246	Fanshawe	1828

Trees



```
<poem>
        <heading>The SICK ROSE</heading>
        <stanza>
                <line>O Rose thou art sick.</line>
                <line>The invisible worm,</line>
                <line>That flies in the night</line>
                <line>In the howling storm:</line>
        </stanza>
        <stanza>
                <line>Has found out thy bed</line>
                <line>Of crimson joy:</line>
                <line>And his dark secret love</line>
                <line>Does thy life destroy.</line>
        </stanza>
</poem>
```

Entity/Relationship (ontologies)



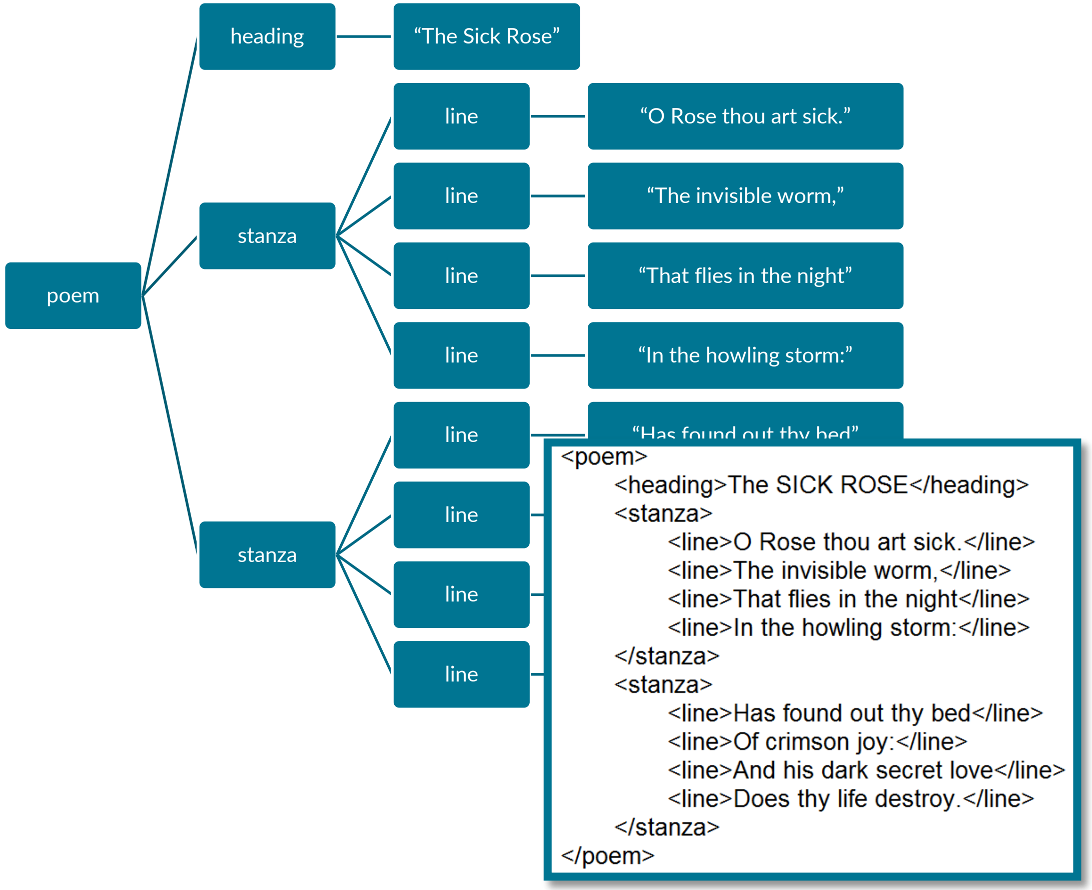
The relational model

Work	Author	Title	Date
W58425	P42425	Moby Dick	1851
W85246	P24246	The Scarlett Letter	1860
W55427	P24246	Fanshawe	1828

Here a **relational** model is being used:

- Relations (tables) are well-suited for data that conceptualized as attribute/value pairs.
- This particular relational model includes the *attributes* **Title** and **Date**.
- It is modeling a state of affairs where a novel, *Moby Dick*, was published in 1851.

The tree model



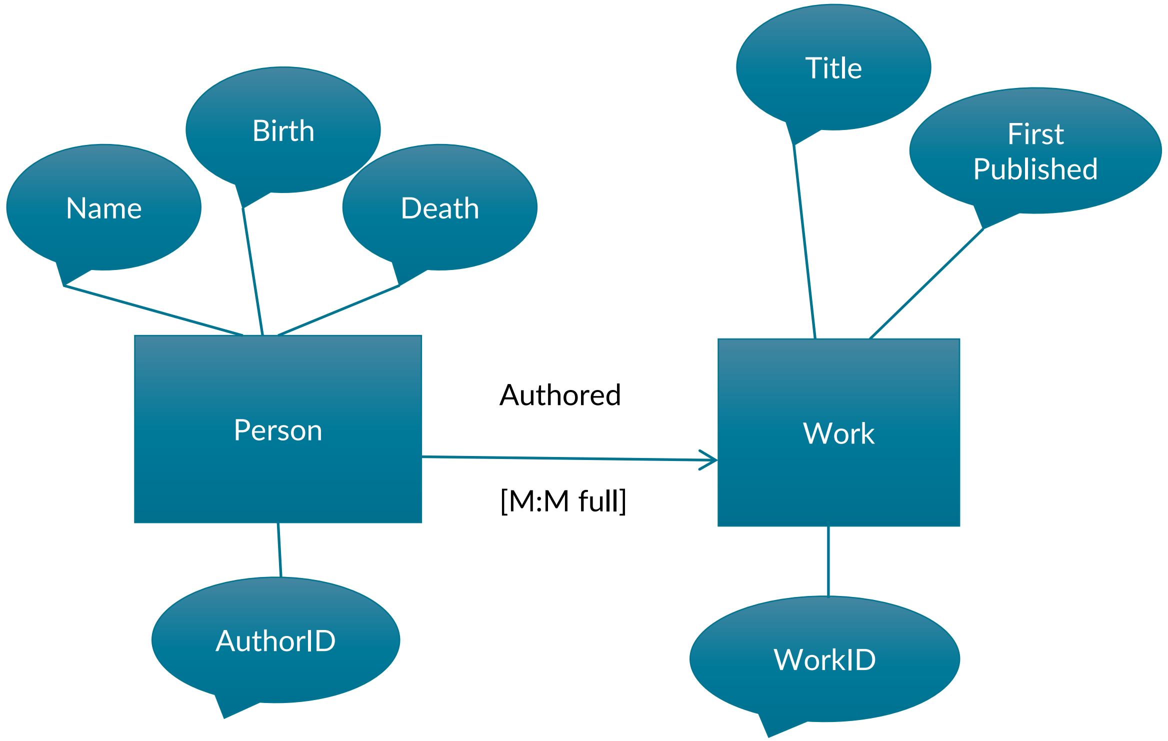
```
<poem>
        <heading>The SICK ROSE</heading>
        <stanza>
                <line>O Rose thou art sick.</line>
                <line>The invisible worm,</line>
                <line>That flies in the night</line>
                <line>In the howling storm:</line>
        </stanza>
        <stanza>
                <line>Has found out thy bed</line>
                <line>Of crimson joy:</line>
                <line>And his dark secret love</line>
                <line>Does thy life destroy.</line>
        </stanza>
</poem>
```

Here a **tree** model is being used:

This particular tree is serialized in XML.

- tree models are well-suited for data that has a tree-like or hierarchical structure, such as documents. But they can also be used to serialize relations and other model instances.
- In this tree the *nodes* have labels such as **stanza** and **line**.
- It is modeling a poem that has two stanzas each with four metrical lines.

The entity/relationship model



Here is an ER schema,

- ER models operate at a high of abstraction, representing the things and relationships of a domain.
- In this ER schema there are two *entity* classes: **Person** and **Work; e**ach entity class has several *attributes*, and there is a *relationship* (**Authored**) that obtains between entities in those classes.
- Both relation and tree data models and be used to implement ER models.
  - [For our purposes ER diagrams, UML class diagrams, other conceptual modeling approaches, RDFS and OWL, and other ontology languages are all fundamentally similar and may be all be considered ways of specifying an ontology]

What, exactly, is a data model?

The phrase “data model” has three common senses:

1. A *type* of framework for representing information

2. A *particular* framework for representing information (typically specified by a *schema*)

3. The *application* of a particular framework to represent information

Sense 1: A *type* of framework

1) “The relational model, with attributes, tuples, and values, is a good one for organizing course registration information.”

2) “The tree model, with nodes, labels, and edges, excels at organizing natural language text.”

3) “The entity relationship model, with entities and relationships, identifies the things and relationships in a domain of interest.”

Sense 2: A particular framework (schema)

1) “The registrar’s **relational model** includes these attributes: *course*, *prerequisites*, *credits*, *department* . . . and assigns *credits* the datatype *integer* . . .”

2) “The journal uses an **XML tree model**. It includes the nodes *article*, *title*, *author*, *affiliation* ... It requires that *title* node must (and may only) appear as the first child of an *article* node . . .”

3) “The **ER model** for registration includes the entities *person*, *course*, and *department*, and the relationships *enrolled in*, *sponsored by*, and *teaches*. It allows *person*s to teach multiple *course*s but requires that a *course* be sponsored by just one *department* . . .”

Sense 3: The *application* of a particular framework

1) “The registrar’s [relational] **model** has the value “IS501” for *course* in the only tuple that has “Smith” for *instructor*.

2) “In the [XML tree] **model** for this article the node labelled *author* has the content “Alonzo Church”, and the following sibling node, *affiliation*, has the content “Princeton University”.

3) “The RDF instance of that [ontology] **model** shows that Anton Marty is enrolled in Dr. Brentano’s course”

What is a Data Model? (Elaborated)

Data models typically have three sorts of components:

1. Structure:	sets and tuples, nodes and arcs, …
2. Things:	values, labels, entities, relationships …
3. Constraints:	datatypes, grammars, cardinality …

Often the specification of *operations* is considered essential:

“A data model is a mathematical formalism with two parts:

1. A notation for describing data
2. A set of operations used to manipulate that data” – Ullman, 1988

Why are we talking about data models?

Because critical activities in data curation include

Select data model types

Select data model schemas

Develop data model schemas

Revise data model schemas

Document data model schemas

Validate dataset instances with schemas

Transform data in one model (type) to another data model (type)

Transform data in one model (schema) to another data model (schema)

Transform data from one representation (e.g. serialization) to another (with same schema)

Integrating data from two different data models (schema or type)

and more

Looking ahead: data model relationships

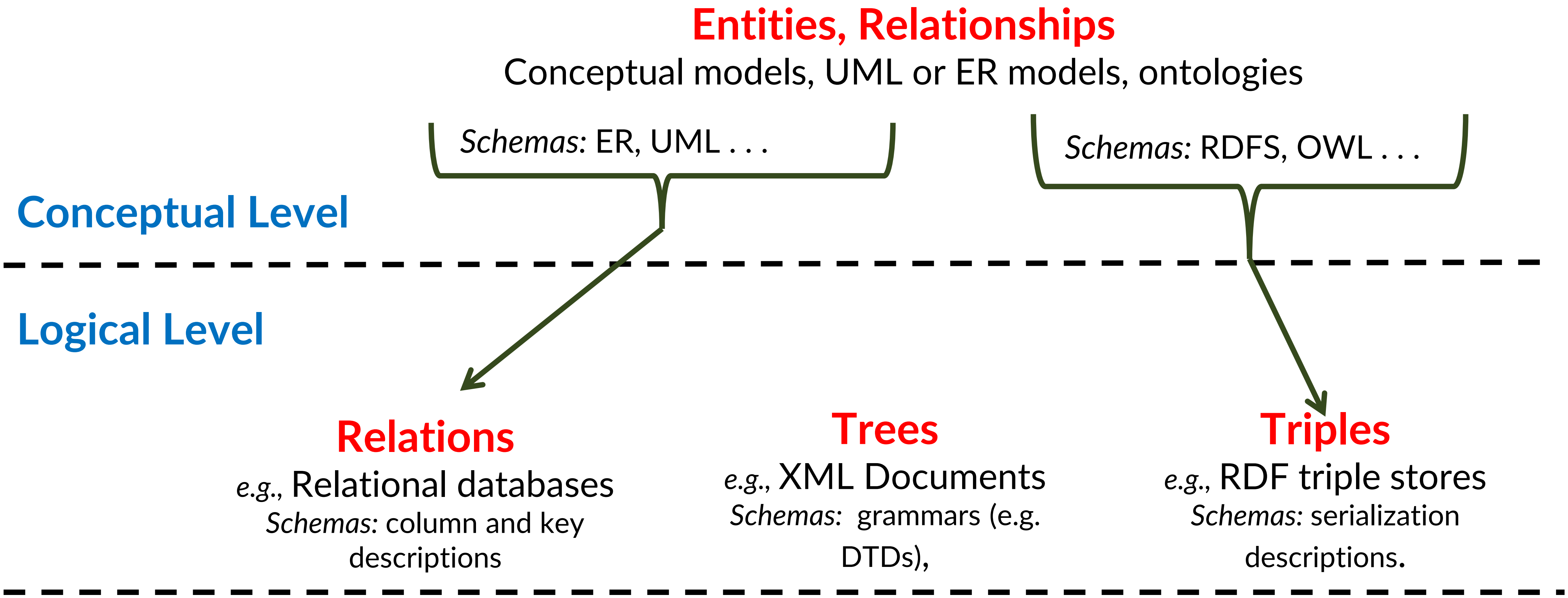
A critical issue in data curation has to do with:

How different types of data models, and different data models of the same type, are related to one another

This is a question we will take up in detail later in the course

But here’s a partial diagram of the territory ahead . . .

Data model relationships



Entities, Relationships

Conceptual models, UML or ER models, ontologies

Schemas: ER, UML . . .

Schemas: RDFS, OWL . . .

Conceptual Level

Logical Level

Relations

e.g., Relational databases
Schemas: column and key descriptions

Trees

e.g., XML Documents
Schemas: grammars (e.g. DTDs),

Triples

e.g., RDF triple stores
Schemas: serialization descriptions.

Physical Level [or: Storage]

[files, records, delimiters, data structures, indexes, etc.]

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.