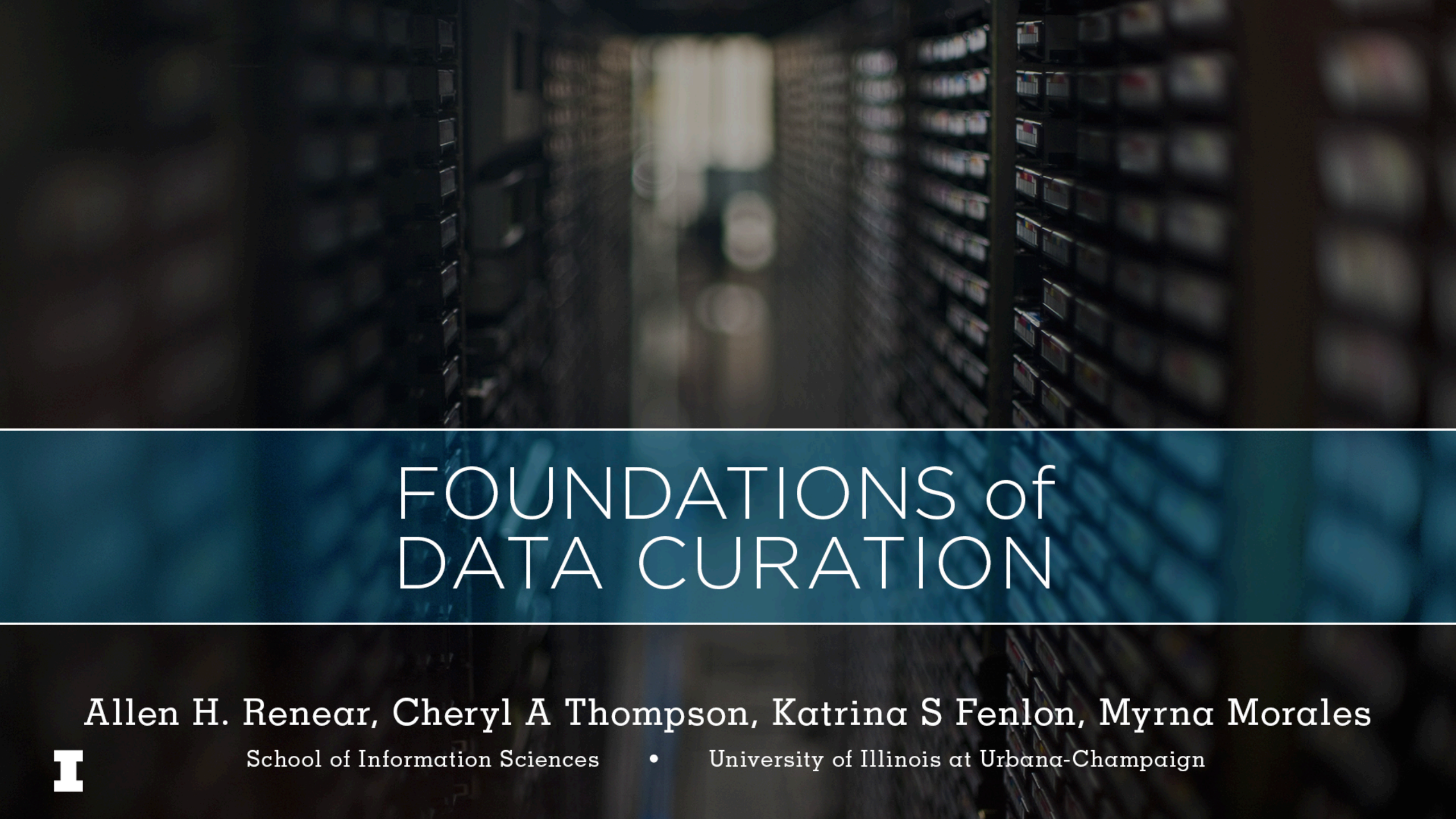


FOUNDATIONS of DATA CURATION

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

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

DATA MODELS: TREES


4
THE SOLUTION:
2. TREES

The Solution: (2) Trees

- The OHCO model of text
- The data structure here is a tree, a kind of graph
- Trees can be serialized in formal languages defined by context free grammars

The OHCO model of text emerges

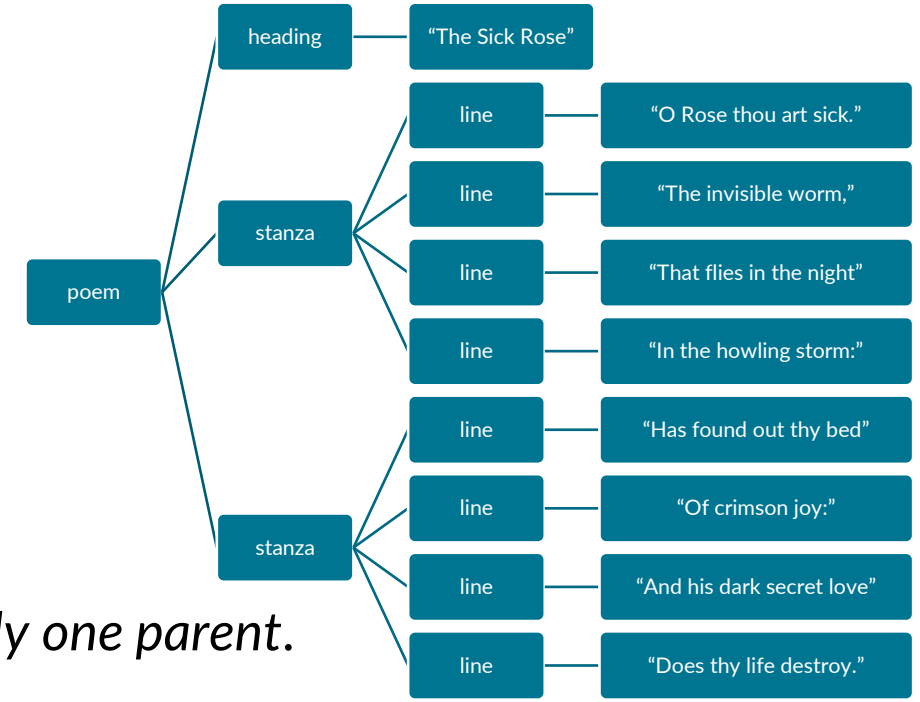
Text is an *O*rdered *H*ierarchy of *C*ontent *O*bjects

- *content* objects = things such as chapters, paragraphs, sentences, stanzas, lines, speeches, equations, titles, headings, abstracts

- *hierarchy* = sentences inside paragraphs, paragraphs inside sections, sections inside chapters, etc., nesting with no overlaps

- *ordered* = objects proceed or follow one another

Trees



Most documents can be modeled as *trees*.

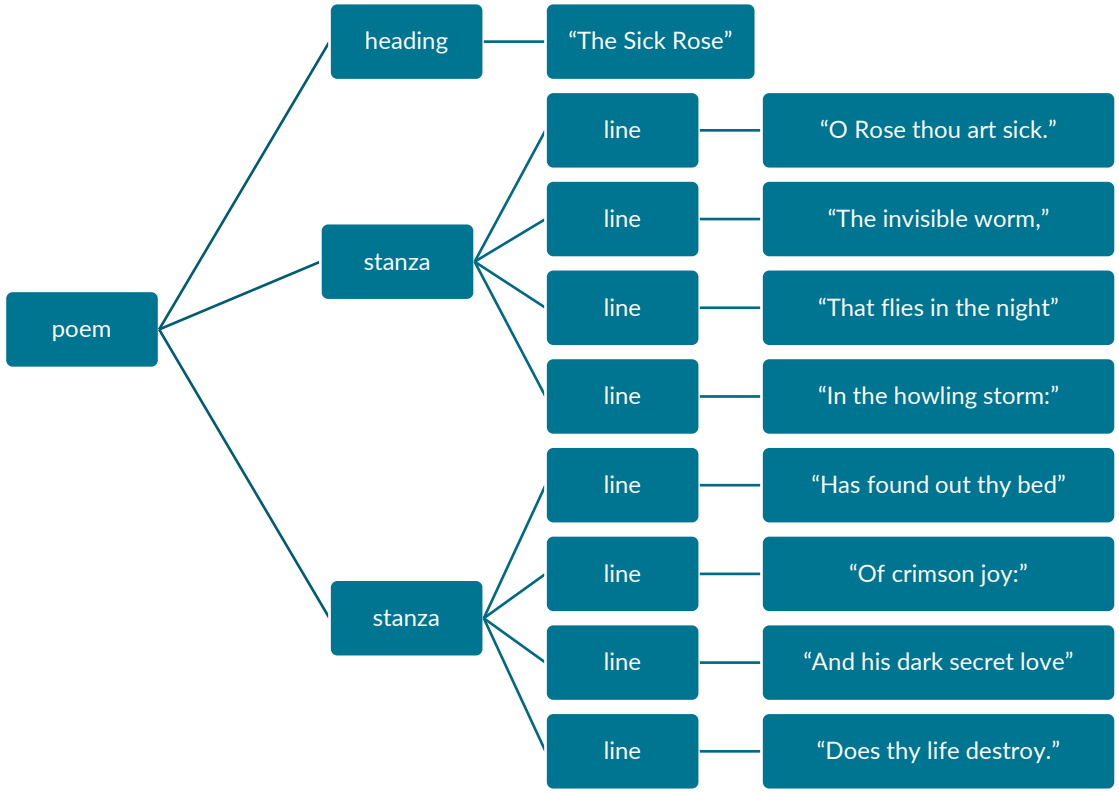
A tree, in our sense*, is

a *directed acyclic graph* with *ordered branches*
and *all nodes except one having exactly one parent*.

The nodes are labelled (e.g. *title*)
and, also, typically annotated with additional information (such as lang=English).

* A specialization of the usual definition.

Using XML to serialize a tree




```
<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>
```

A tree can be serialized with a formal language defined by a context free grammar, such as an XML language.

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.