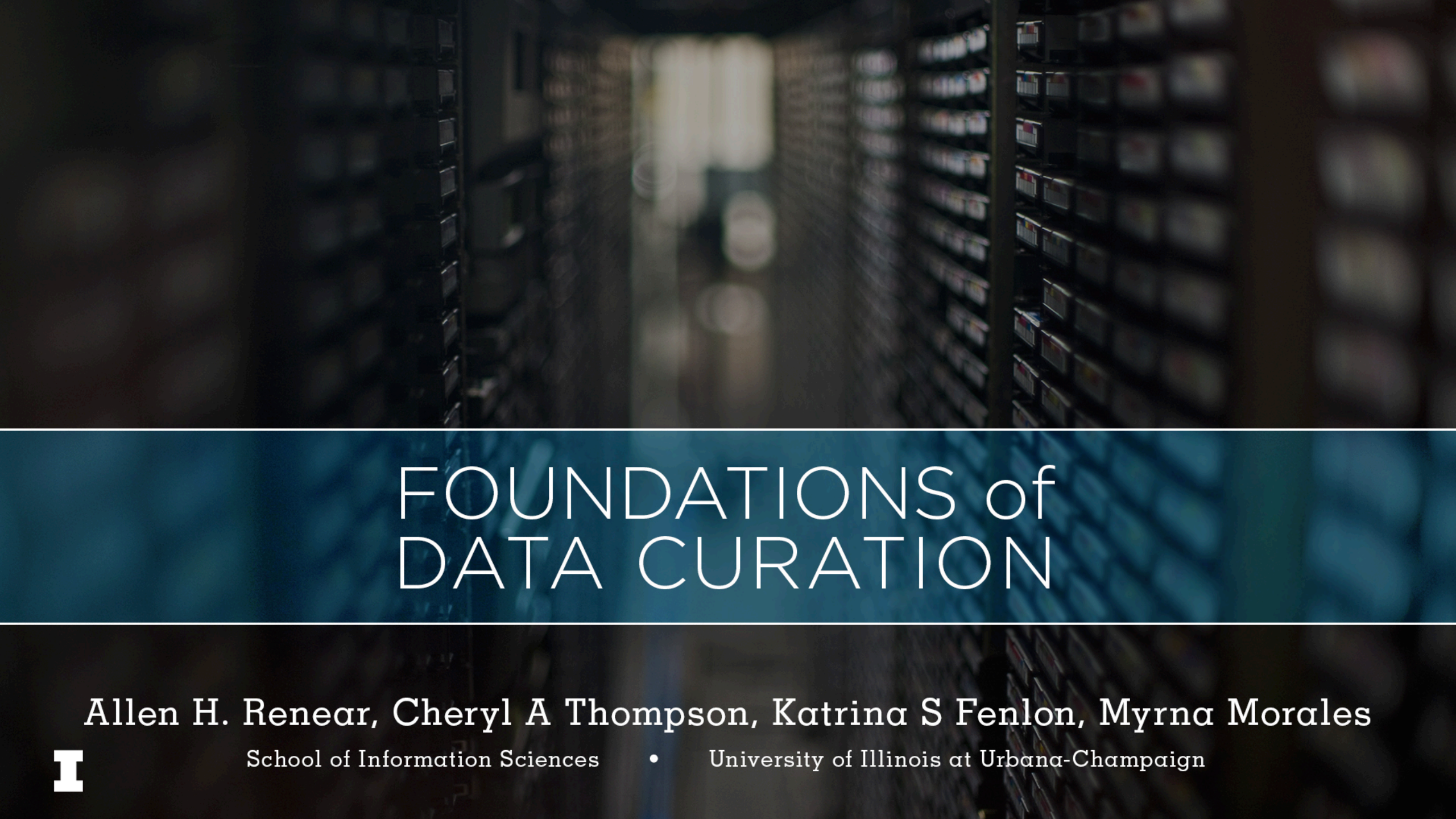

FOUNDATIONS of DATA CURATION

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

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

DATA MODELS: TREES


3
THE SOLUTION:
1. DESCRIPTIVE MARKUP

The Solution: (1) Descriptive Markup (it’s the same solution)

The problem, again

The solution: (1) Descriptive Markup

- How it emerged
- How it works and delivers the goods
- Why it works

Our example:

Recall athe problems we noticed with this approach to organizing text.

Input file contains this data:

```
.pa odd;.font Times;.size 14;
 .it;.ce;.in +5 -5;.sk 2p a;.kp next;.toc include; The Sick Rose[…]
```

After processing the output is rendered like this:

The Sick Rose

[…]

The first improvement
(from the US Government Printing Office in the 1960s)

A macro is defined to abbreviate formatting commands:

```
&format17  =df
   ".pa odd;.font Times;.size 14;.it;.ce;.in +5 -5;.sk 3p;.sk 2p
   a;.kp next;.toc include;"
```

The macro is used in the input:

```
.format17:The Sick Rose […]
```

This helps a bit (why?).

But not as much as it might help (why?).

A Problem

Although format17 helps, it doesn't go far enough.

Since format17 abstracts to a typographic "look" it can be used wherever you want that look: titles, captions, extract labels, . . .

`.format17:The Sick Rose`	`.format17:Fig. 1. A tea rose`
The Sick Rose *[…]*	 *Fig. 1. A tea rose*

So what might be even better?

Don't identify the *look*, identify the component itself

A much better improvement

A macro is defined to identify the logical component of the text itself not the the intended processing, or the appearance of the that component

```
&title. =df
                “.pa odd;.font Times;.size 14;.it;.ce;.in +5 -5;.sk
                3p;.sk 2p a;.kp next;.toc include;”
```

The macro is used in the input file like this:

```
.title:The Sick Rose […]
```

Abstraction from storage

Consider this text:

An example of the Tea rose is: http://www.example.org/rose/hybrid/tea/pink/42

It's a mouthful. And what if the image moves?

A entity name is defined to abbreviate a location:

```
&rose42 =df http://www.example.org/rose/hybrid/tea/pink/42
```

The entity name is used in the input:

```
An example of the Tea rose is: &rose42;
```

After processing the output is rendered like this:

An example of the Tea rose is:

Again: Abstraction and Indirection

Again our solution is abstraction

In explicitly identifying recurring logical objects (like titles) we abstract away from the varied and varying details of processing and storage

We then exploit this abstraction by using indirection:

- Mapping object instances to storage locations
- Mapping types of objects to processing rules

Achieving efficiencies and new functionality

Examples of text components

- Title
- Author
- Date
- Abstract
- Section, subsection, subsubsection
- Section title, subsection title … etc
- Paragraph
- Extract (long quotation)
- Equation
- Diagram
- Footnote

Genre-specific text components

Scientific article:

- Title, author, affiliation, address, date submitted, date revised, keywords, abstract, introduction, methodology, results, discussion, conclusion, diagram, equation, plate, graph, chart, bibliography, bibliography item, date

Playscripts:

- Act, scene, stage direction, line, character, cast list

Poetry:

- Title, author, verse, stanza, couplet, line, half-line

Also:

- Legal and financial documents such as contracts, deeds, licenses, writs, tickets, receipts
- Office documents such as project proposals, monthly reports, position descriptions, performance evaluations, other forms
- Etc...

An XML example

```
<anthology>
  <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>
  <!-- more poems go here -->
</anthology>
```

**Example from the Text Encoding Initiative (P5)*

Other examples

NeXML for phylogenetic information

```
<characters otus="tax1" id="m1" xsi:type="nex:DnaSeqs">
   <!-- ... -->
   <matrix aligned="1">
      <row id="r1" otu="t1"><seq>AACATATCTC</seq> </row>
      <row id="r2" otu="t2"><seq>ATACCAGCAT</seq> </row>
      <row id="r3" otu="t3"><seq>GAGGGTATGG</seq> </row>
      <row id="r4" otu="t4"><seq>GGTCTTAGAG</seq> </row>
      <row id="r5" otu="t5"><seq>CGTCACAGTG</seq> </row>
   </matrix>
</characters>
```

Music Markup Language

```
<section>
  <measure n="0" xml:id="m0" type="upbeat">
    <staff n="1">
      <layer n="1">
        <rest dur="4"/>
      </layer>
    </staff>
    <staff n="2">
      <layer n="1">
        <beam>
          <note xml:id="m0_s2_e1" pname="e" oct="5" dur="8" stem.dir="down"/>
          <note xml:id="m0_s2_e2" pname="f" oct="5" dur="8" stem.dir="down"/>
        </beam>
      </layer>
    </staff>
    ...
```

Medical Markup Language

```
...
<clinical_document_header>
<document_type_cd DN="MML Document" S="1.2.392.114319.1.1" V="0300" />
   <provider>...</provider>
   <patient>...</patient>
<!-- ...-->
</clinical_document_header>
<body>
   <!-- ...-->
   <mml:docInfo contentModuleType="report"
        moduleVersion="http://www.medxml.net/MML/report/1.0">
      <mml:securityLevel>...</mml:securityLevel>
      <mml:title generationPurpose="reportRadiology">CT scan Report</mml:title>
      <mml:docID><mml:uid>JPN432101234567RR20020823_CT_20020851501</mml:uid></mml:docID>
   </mml:docInfo>
</body>
...
```

Descriptive Markup

Descriptive markup describes the *logical components* of documents.

It does not specifiy *processing*.

Advantages of descriptive markup

Authoring, Editing, Transcribing:

- Composition is simplified
- Writing tools are supported
- Alternative views and links facilitated

Publishing:

- Formatting generically specified and modified
- Apparatus automated
- Output device support enhanced
- Portability maximized

Retrieval and Analysis:

- Information retrieval supported
- Analytical procedures supported

How does this help with data curation

Descriptive markup makes digital documents...

- easier to create
- easier to maintain
- easier to convert (new formats, new delivery software)
- better integrated with workflow in organization
- better integrated with other applications and tools (databases, word processing templates, indexes,)etc.
- more accessible to varied audiences
- easier to accommodate different technological circumstances (varying hardware, operating systems, browser software (brands and versions both), connectivity (bandwidth), etc.
- Easier to accommodate different perceptual abilities (blindness, other sight disabilities, dyslexia, 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.