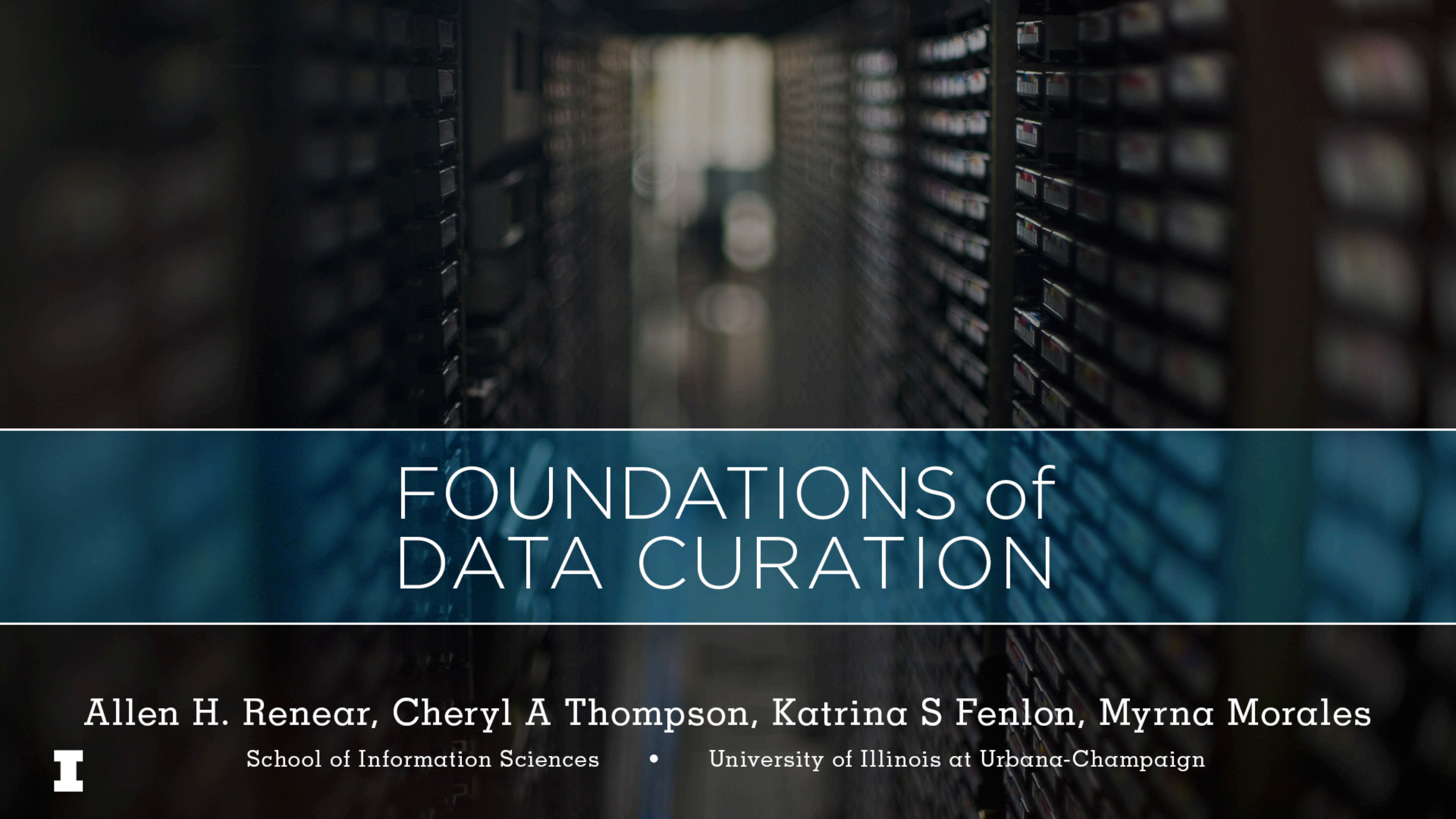


FOUNDATIONS of DATA CURATION

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

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

DATA MODELS: TREES


2
THE PROBLEM

The Problem (it's the same problem)

- The situation (circa 1960)
  - Text is stored and processed in radically different ways
  - Interaction with text is immediately and directly via storage and processing methods
  - Explicit and formal conceptualization of text components as such is rare (and typically only in human memory)
- Why is this a problem?
  - Huge operational inefficiencies
  - Lack of functionality
  - Lack of data independence

The promise of digital documents – unfulfilled

In the last video we noted the importance of documents and the long-standing promise of digital documents to provide exciting new functionality.

But we also noted that we are only part way there

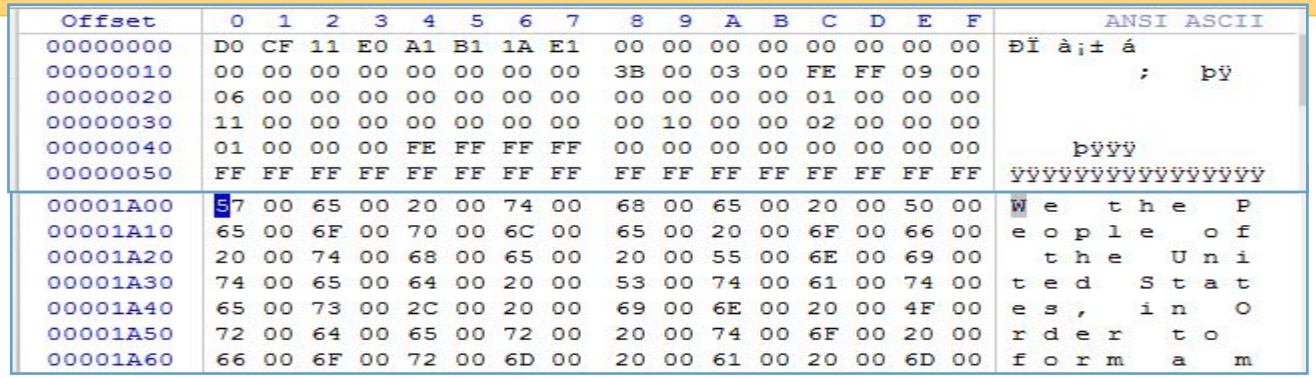
Creating complex, high-performance documents remains arduous and the results are hardly the sophisticated high-performance information environments we were promised.

The Problem *(it's the same problem!!)*

There are many many ways to represent text in documents.

```
.ll 3i
.mk a
.ce
Preamble
.sp
We, the people of the United States, in order to form a more perfect Union
```

```
{\rtf1\ansi{\fonttbl\f0\fswiss Helvetica;}\f0\pard
This is some {\b bold} text.\par }
```



Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	ANSI ASCII
00000000	D0	CF	11	E0	A1	B1	1A	E1	00	00	00	00	00	00	00	00	ÐÏ à¡± á
00000010	00	00	00	00	00	00	00	00	3B	00	03	00	FE	FF	09	00	; þÿ
00000020	06	00	00	00	00	00	00	00	00	00	00	00	01	00	00	00	
00000030	11	00	00	00	00	00	00	00	00	10	00	00	02	00	00	00	
00000040	01	00	00	00	FE	FF	FF	FF	00	00	00	00	00	00	00	00	þÿÿÿ
00000050	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00001A00	57	00	65	00	20	00	74	00	68	00	65	00	20	00	50	00	W e t h e P
00001A10	65	00	6F	00	70	00	6C	00	65	00	20	00	6F	00	66	00	e o p l e o f
00001A20	20	00	74	00	68	00	65	00	20	00	55	00	6E	00	69	00	t h e U n i
00001A30	74	00	65	00	64	00	20	00	53	00	74	00	61	00	74	00	t e d S t a t
00001A40	65	00	73	00	2C	00	20	00	69	00	6E	00	20	00	4F	00	e s , i n O
00001A50	72	00	64	00	65	00	72	00	20	00	74	00	6F	00	20	00	r d e r t o
00001A60	66	00	6F	00	72	00	6D	00	20	00	61	00	20	00	6D	00	f o r m a m

```
<w:body>
    <w:p w:rsidR="00146B24" w:rsidRDefault="00146B24">
      <w:bookmarkStart w:id="0" w:name="_GoBack"/>
      <w:bookmarkEnd w:id="0"/>
    </w:p>
    <w:p w:rsidR="00146B24" w:rsidRDefault="00146B24“  w:rsidP="00146B24">
      <w:pPr>
        <w:pStyle w:val="Heading1"/>
      </w:pPr>
      <w:r>
        <w:t>Preamble</w:t></w:t>
      </w:r>
    </w:p>
    <w:p w:rsidR="001C180C" w:rsidRDefault="00146B24">
      <w:r>
        <w:t xml:space="preserve">We the People of the United States, </w:t></w:t>
      </w:r>
    </w:p>
      …
```

Interaction is typically directly with these storage structures

The fundamental principles of abstraction and indirection are not implemented

The problems that result

- Training does not transfer
- Tools are not interoperable
- Data from multiple sources cannot be integrated
- Applications development is arduous
- Documentation is difficult
- Validation and assurance is difficult
- Specialized applications (searching, analysis, etc.) are not supported
- Schemas are typically non-existent, or unhelpful

And so on. . .

Sound familiar?

A closer look at one example: electronic publishing in the 1960s

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

[…]

What are some problems with this approach to organizing text?

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.