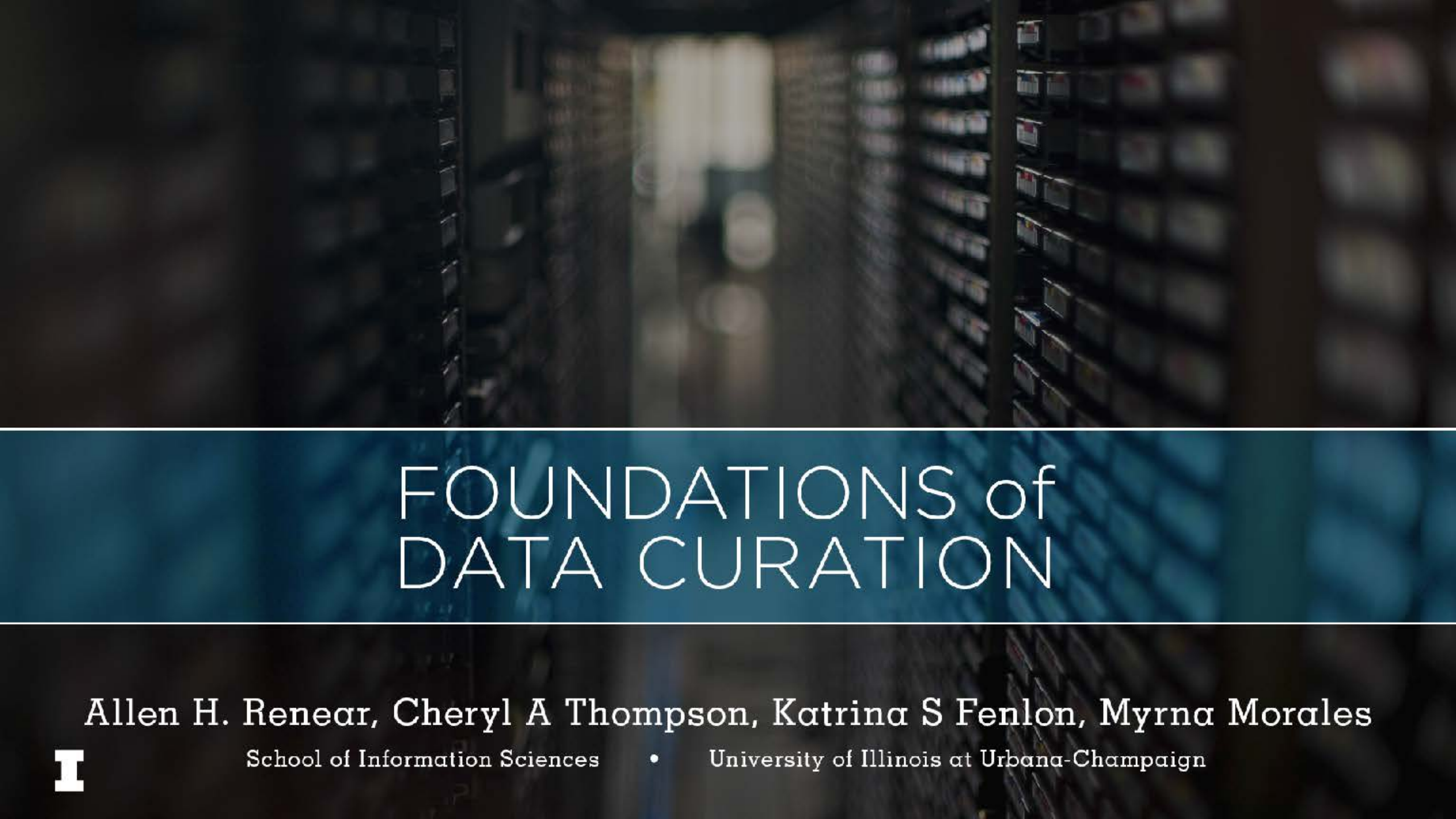


FOUNDATIONS of DATA CURATION

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

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

DATA MODELS: TREES

1

TEXT AND DOCUMENTS

Text and Documents

- Why is text important in data curation?
- Examples of data intensive documents.
- The promised functionality of digital documents . . . but not easily realized

What's so important about documents?

The document is the natural unit of textual information.

Why are documents important? Because. . .

That's where the information is.

Arguably much more information exists in documents and unstructured natural language text then exists in databases.

That's where the action is.

Documents are typically instruments of action; information only has traction on the world when it is communicated in documents

you're hired; you're fired; we agree; you own; you owe;

Even databases typically only have effects when a report (a document) is generated and read by someone (or some processing agent).

That's where we live, work, and play.

We cannot imagine our social lives -- commercial, scientific, cultural everyday -- without the medium of document-based communication.

Documents and data

Relational databases seem a natural fit for certain kinds of data:

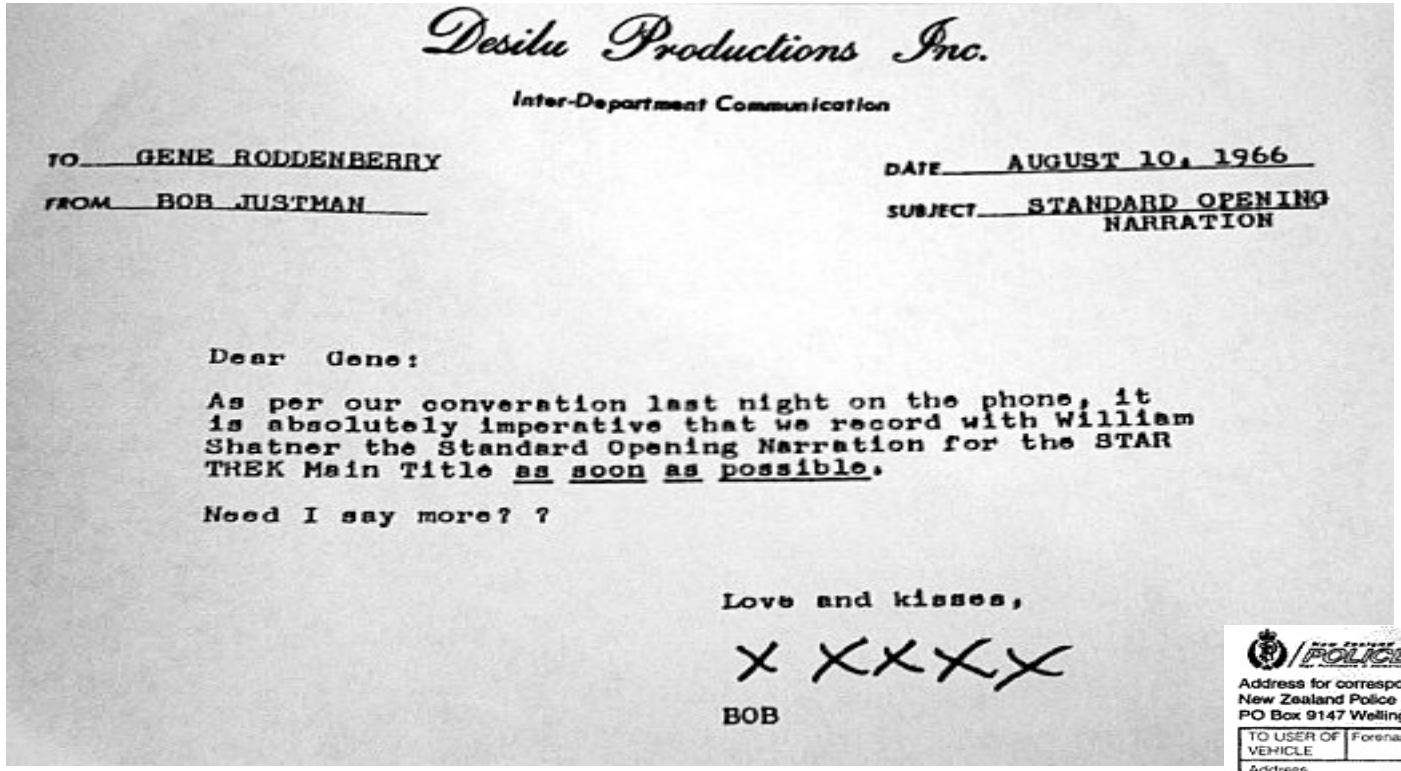
Particularly where the data has this form: *something* has a *value* for an *attribute*

But most the information in the world is contained not in databases, but in the *text* of *documents*.

This poses challenges for the relational model:

1. In the text of a document is not obvious, and may require considerable human analysis, to see *what* is being said about *what*.
2. The document's text *itself* often needs to be organized and managed (as in publishing applications), rather than the data being asserted by the text; most documents do not appear to be tabular in nature.

Examples



Desilu Productions Inc.

Inter-Department Communication

TO GENE RODDENBERRY

DATE AUGUST 10, 1966

FROM BOB JUSTMAN

SUBJECT STANDARD OPENING NARRATION

Dear Gene:

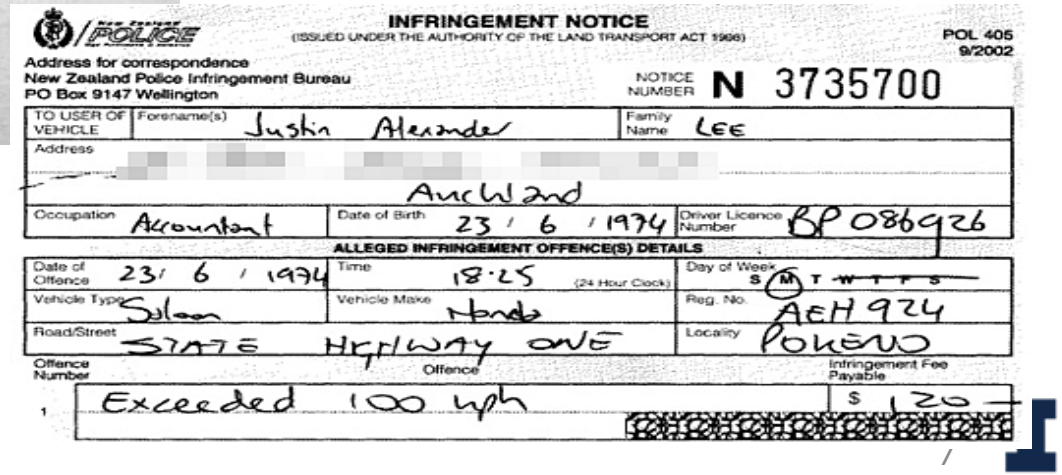
As per our converation last night on the phone, it is absolutely imperative that we record with William Shatner the Standard Opening Narration for the STAR TREK Main Title as soon as possible.

Need I say more? ?

Love and kisses,

X XXXX

BOB



INFRINGEMENT NOTICE

(ISSUED UNDER THE AUTHORITY OF THE LAND TRANSPORT ACT 1998)

POL 405
9/2002

Address for correspondence
New Zealand Police Infringement Bureau
PO Box 9147 Wellington

NOTICE NUMBER N 3735700

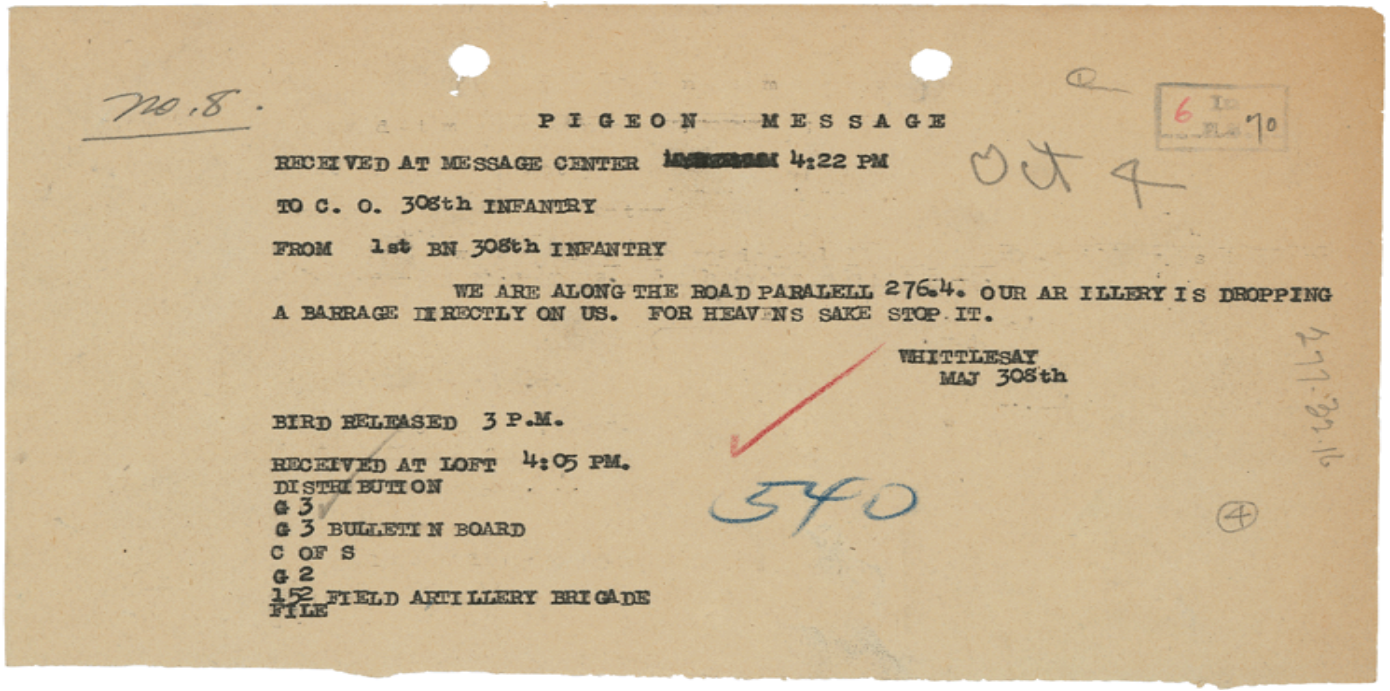
TO USER OF VEHICLE	Forename(s) Justin Alexander	Family Name	LEE
Address	[redacted] Auckland		
Occupation	Accountant	Date of Birth 23 / 6 / 1974	Driver Licence Number BP 086926

ALLEGED INFRINGEMENT OFFENCE(S) DETAILS

Date of Offence	23 / 6 / 1974	Time 18·25 (24 Hour Clock)	Day of Week S M T W T F S
Vehicle Type	Saloon	Vehicle Make	Reg. No. AEH 924
Road/Street	STATE HIGHWAY ONE		Locality POKENO

Offence Number	Offence	Infringement Fee Payable
1	Exceeded 100 kph [redacted]	$ 120

Examples



No. 8.

PIGEON MESSAGE

RECEIVED AT MESSAGE CENTER [redacted] 4:22 PM

TO C. O. 308th INFANTRY

FROM 1st BN 308th INFANTRY

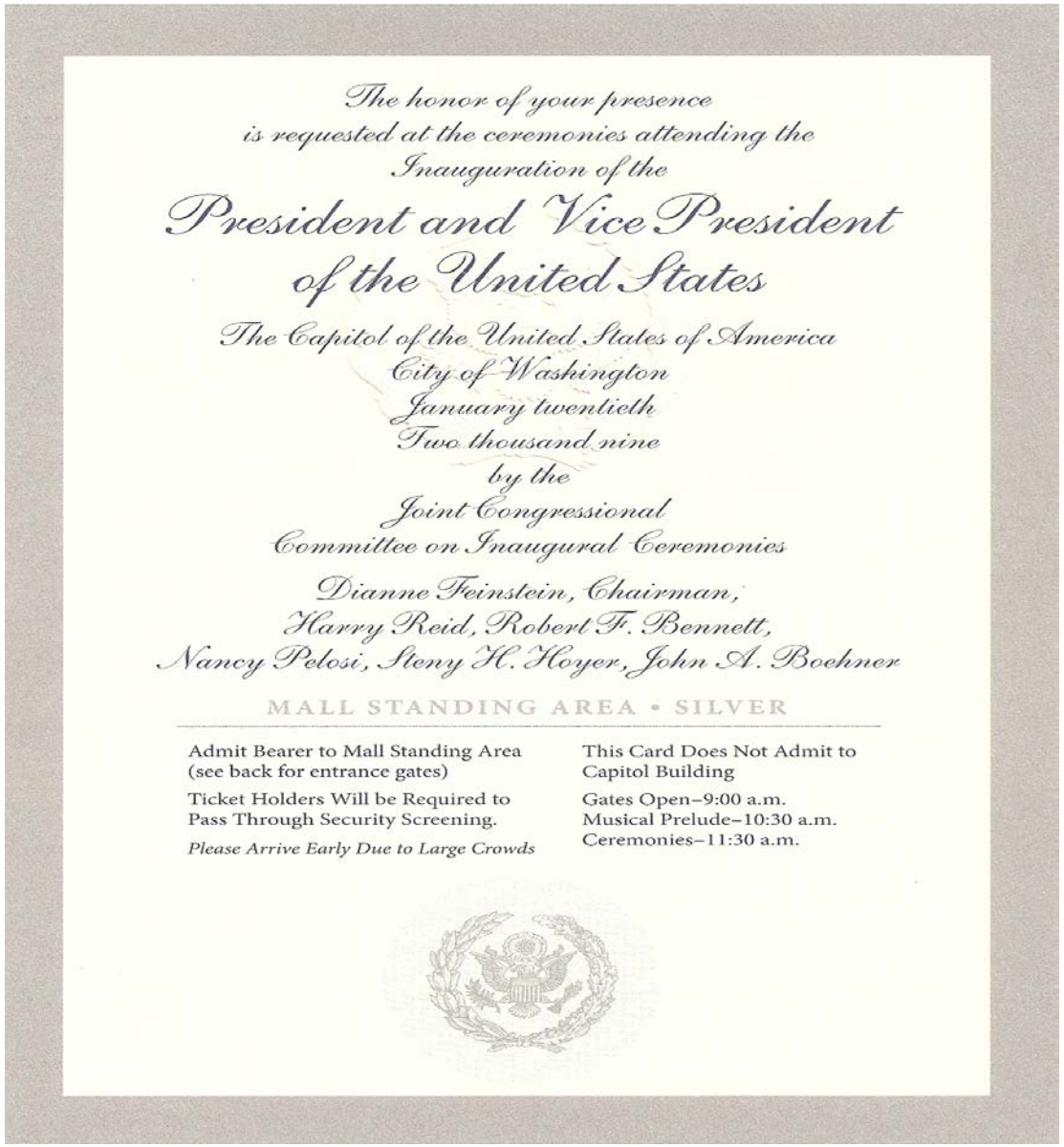
WE ARE ALONG THE ROAD PARALELL 276.4. OUR AR ILLERY IS DROPPING A BARRAGE DIRECTLY ON US. FOR HEAVENS SAKE STOP IT.

WHITTLESAY
MAJ 308th

BIRD RELEASED 3 P.M.

RECEIVED AT LOFT 4:05 PM.

DISTRIBUTION
G 3
G 3 BULLETIN BOARD
C OF S
G 2
152 FIELD ARTILLERY BRIGADE
FILE



The honor of your presence
is requested at the ceremonies attending the
Inauguration of the

President and Vice President
of the United States

The Capitol of the United States of America
City of Washington
January twentieth
Two thousand nine
by the
Joint Congressional
Committee on Inaugural Ceremonies

Dianne Feinstein, Chairman,
Harry Reid, Robert F. Bennett,
Nancy Pelosi, Steny H. Hoyer, John A. Boehner

MALL STANDING AREA • SILVER

Admit Bearer to Mall Standing Area
(see back for entrance gates)

Ticket Holders Will be Required to
Pass Through Security Screening.

Please Arrive Early Due to Large Crowds

This Card Does Not Admit to
Capitol Building

Gates Open–9:00 a.m.
Musical Prelude–10:30 a.m.
Ceremonies–11:30 a.m.

Examples

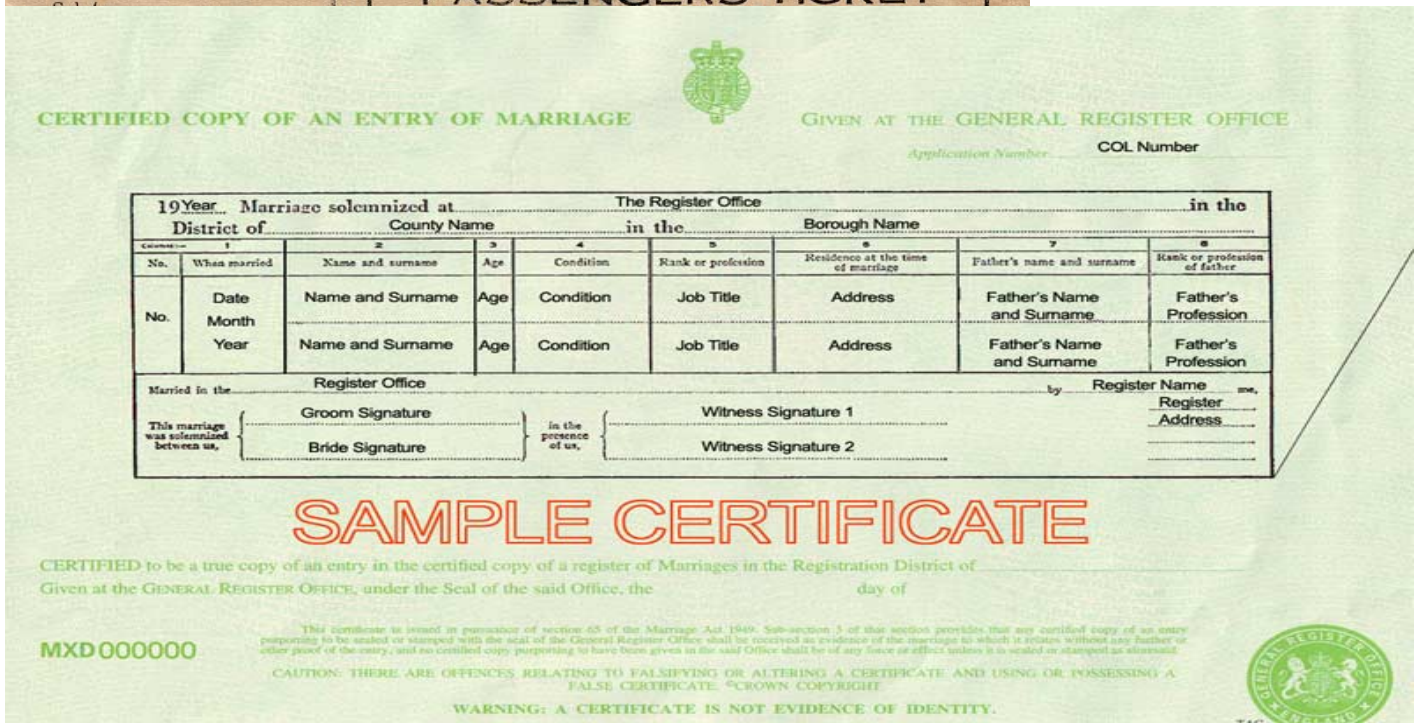
Tele:"DAILYERVICE" Phone: 32.

Date .....
Bus No. .....
From AML.—KTM.
Class .....
Fare Rs .....
Extra Lug. Frt. .....
S. No.

Nepal Transport Service

BUS NO.	Class	DATE
FROM AMLEKHGUNJ	TO KATHMANDU	
S.	Rs.	Sd.

PASSENGERS TICKET



CERTIFIED COPY OF AN ENTRY OF MARRIAGE

GIVEN AT THE GENERAL REGISTER OFFICE

Application Number COL Number

19 Year Marriage solemnized at The Register Office in the District of County Name in the Borough Name

No.	When married	Name and surname	Age	Condition	Rank or profession	Residence at the time of marriage	Father's name and surname	Rank or profession of father
No.	Date Month Year	Name and Surname	Age	Condition	Job Title	Address	Father's Name and Surname	Father's Profession
		Name and Surname	Age	Condition	Job Title	Address	Father's Name and Surname	Father's Profession

Married in the Register Office by Register Name me, Register Address

This marriage was solemnized between us, Groom Signature / Bride Signature — in the presence of us, Witness Signature 1 / Witness Signature 2

SAMPLE CERTIFICATE

CERTIFIED to be a true copy of an entry in the certified copy of a register of Marriages in the Registration District of

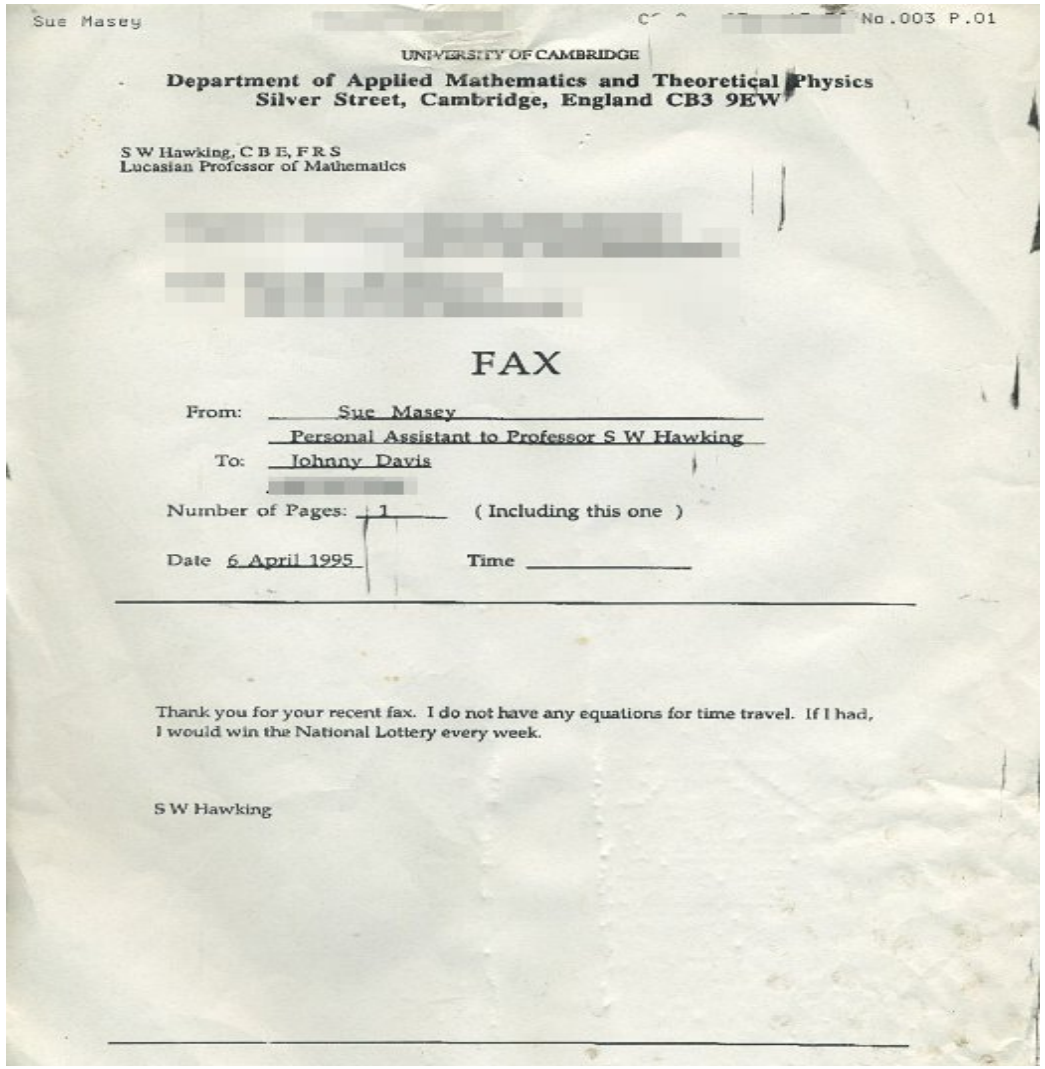
Given at the GENERAL REGISTER OFFICE, under the Seal of the said Office, the day of

MXD 000000

This certificate is issued in pursuance of section 65 of the Marriage Act 1949. Sub-section 3 of that section provides that any certified copy of an entry purporting to be sealed or stamped with the seal of the General Register Office shall be received as evidence of the marriage to which it relates without any further or other proof of the entry, and no certified copy purporting to have been given in the said Office shall be of any force or effect unless it is sealed or stamped as aforesaid.

CAUTION: THERE ARE OFFENCES RELATING TO FALSIFYING OR ALTERING A CERTIFICATE AND USING OR POSSESSING A FALSE CERTIFICATE. ©CROWN COPYRIGHT

WARNING: A CERTIFICATE IS NOT EVIDENCE OF IDENTITY.



Sue Masey [redacted] No.003 P.01

UNIVERSITY OF CAMBRIDGE
Department of Applied Mathematics and Theoretical Physics
Silver Street, Cambridge, England CB3 9EW

S W Hawking, C B E, F R S
Lucasian Professor of Mathematics

[redacted]

FAX

From: Sue Masey
Personal Assistant to Professor S W Hawking
To: Johnny Davis
[redacted]
Number of Pages: 1 (Including this one)
Date 6 April 1995 Time

Thank you for your recent fax. I do not have any equations for time travel. If I had, I would win the National Lottery every week.

S W Hawking

is done during the 'Reynolds averaging' process, corresponds to covariance terms such as $\langle C_i' v_j' \rangle$ for a reaction involving two gas phase components i and j. The second-order equation describing the flux term takes the form

$$\frac{\partial \langle w'C_i' \rangle}{\partial t} + \ldots = \ldots \quad (2)$$

(1) (2)

$$\ldots = \ldots \quad (3) \quad (4) \quad (5)$$

Here term (1) is the differential transport term; (2) is the vertical gradient of the turbulent flux; (3) is the change due to buoyancy (Gi); (4) is known as the pressure scrambling term and represents correlations of pressure fluctuations with concentration (P); and the (5) term represents the effect of change in flux due to chemical reactions.

The dynamic evolution of the 1-D PBL is described in the current model by using an algebraic stress model (ASM) assumption to close the second-order transport terms. This approach yields equations that would conform to a level-2.5 closure of the Reynolds averaged equation. We adopted a similar approach (ASM-like) and assumptions to close the second-order chemical equation. This yields the following equation for the flux term:

$$\frac{\partial \langle w'C_i' \rangle}{\partial t} = \ldots \quad (6)$$

$$\ldots \quad (7)$$

where the respective correlation terms are derived as in (3). As is the usual practice in ASM-based modeling, the flux terms can now be represented in a form similar to the eddy mixing coefficient formulation, for example as follows:

$$\langle w'C_i' \rangle = \ldots \quad (8)$$

Here A is a constant derived from rearranging (2). This constant is dependent on the flow conditions and correlation terms discussed above.

The transport terms in the chemical tendency equations are solved by using a fully implicit finite difference scheme. The vertical grid employed has 43 levels (and put on a log-normal axis to give highest resolution in the lowest 50 m of the model. The model is solved with a time step of 5 s for the chemistry and 5 min for the dynamic equations. The preliminary focus of these calculations is on evaluating the effect of chemistry on the calculated flux of NO from soil.

3. PRELIMINARY RESULTS

In this initial set of calculations, the mixing coefficient derived as described by (8) were employed in solving the transport equation for NO, NO_2, and O_3, keeping the eddy mixing coefficients for the rest of the 72 trace gases same as that calculated for the temperature in the model. This set of calculations is referred to as 'with reaction'. In a second set of calculations, the eddy mixing coefficients for all of the trace gases in the model are set to those derived for the temperature and are referred to as "no reaction" in the following discussion. In both of these calculations, the

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-HEP-12-046		

4. Title and Subtitle	5. Report Date
Asset Sustainability Index: A Proposed Measure for Long-Term Performance	July 2012
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Gordon D. Proctor, Sincera Varma, and Varnedoe	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Gordon Proctor & Associates, Inc 1825 Wilkshire Drive Dublin, Ohio 43016 Starlias Corp. 3737 Woodstone Drive Lewis Center, Ohio 43035 National Center for Pavement Preservation 3847 Jolly Road Okemos, MI 48864	11. Contract or Grant No. DTFH61-10-C-00036

12. Sponsoring Agency Name and Address	13. Type of report and period covered
FHWA Surface Transportation Environment Planning Cooperative Research Program, and the Office of Asset Management, Pavements and Construction	2011
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report examines the concept of asset sustainability metrics. Such metrics address the long-term performance of highway assets based upon expected expenditure levels. It examines how such metrics are used in Australia, Britain and the private sector. It also reviews asset management data from selected states to illustrate that long-term sustainability metrics could be produced using available US asset management data.

17. Key Words	18. Distribution Statement
Asset Sustainability, Asset Management, Long-term Performance, Sustainable Infrastructure Performance Management	No restrictions. This document is available to the public from the: FHWA Surface Transportation Environment and Planning Cooperative Research Program and the Office of Asset Management, Pavements and Construction www.fhwa.dot.gov/hep/step www.fhwa.dot.gov/infrastructure/asstmgmt

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	116	Free

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PRAHA NEUTAJOVANÉ M9-0144G-50420-00

Slaňovací rám SR-1

Provozní kontroly

Strana

tabulek Strana

obrázků Strana

duly/Dokumentace

Požadavk

any stav

sk na pracovní síl

Skupina
Mechanik

ní opatření

Před začát

ie

M9-0144G-50420-00 NEUTAJOVANÉ LOM PRAHA

Postup

1 Provozní kontrola funkčnosti slaňovacího rámu SR-1

1.1 Ujistěte se, že SR (Obr. 1 [1]) je v transportní poloze.

1.2 Vytáhněte držadlo (Obr. 1 [2]) a rukou umístěte SR (Obr. 1 [1]) do sklopné polohy.

1.3 Uvolněte držadlo (Obr. 1 [2]).
- Držadlo (Obr. 1 [2]) se musí automaticky zasunout ve sklopné poloze SR.

1.4 Ujistěte se, že SR (Obr. 1 [1]) je zajištěn ve sklopné poloze.

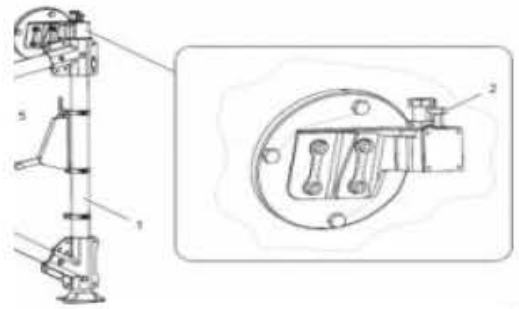
1.5 Zvednětě konzolu závěsu (Obr. 1 [3]) do horní polohy a zajistěte čepem kluzáku (Obr. 1 [4]).
- SR (Obr. 1 [1]) je ve sklopné poloze.

1.6 Podržte konzolu závěsu (Obr. 1 [3]), demontujte čep kluzáku (Obr. 1 [4]) a umístěte konzolu závěsu (Obr. 1 [3]) do spodní polohy a čep kluzáku (Obr. 1 [4]) do držáku (Obr. 1 [5]).
- SR (Obr. 1 [1]) je ve sklopné poloze.

1.7 Vytáhněte držadlo (Obr. 1 [2]) a rukou umístěte SR (Obr. 1 [1]) do transportní polohy.

1.8 Uvolněte držadlo (Obr. 1 [2]).
- Držadlo (Obr. 1 [2]) se musí automaticky zasunout v transportní poloze SR.

1.9 Ujistěte se, že SR (Obr. 1 [1]) je zajištěn v transportní poloze.



Obr. 1 R

ií rám

A-A

Odkazy

4 Direct very axiomatized theories

3. Given a derivation of the *sentence* φ from the axioms of the theory T using the background logical proof system, we will say that φ is a *theorem* of T. Using the standard abbreviatory symbol, we write: $T \vdash \varphi$.

2. A theory T is *sound* iff every theorem of T is true (i.e. true on the interpretation built into T's language). Soundness is, of course, normally a matter of having true axioms and a truth-preserving proof system.

3. A theory T is *effectively decidable* iff the property of being a theorem of T is an effectively decidable property – i.e. iff there is an algorithmic procedure for determining, for any given sentence φ of T's language, whether or not $T \vdash \varphi$.

4. Assume now that T has a standard negation connective '$\neg$'. A theory T *decides* the sentence φ iff either $T \vdash \varphi$ or $T \vdash \neg\varphi$. A theory T *correctly decides* φ just when: if φ is true (on the interpretation built into T's language), $T \vdash \varphi$, and if φ is false, $T \vdash \neg\varphi$.

5. A theory T is *negation-complete* iff T decides every sentence φ of its language (i.e. for every sentence φ, either $T \vdash \varphi$ or $T \vdash \neg\varphi$).

6. T is *inconsistent* iff for some sentence φ, we have both $T \vdash \varphi$ and $T \vdash \neg\varphi$.

Note our decision to restrict the theorems, properly so called, to the derivable *sentences*: no wffs with free variables derived as we go along though a proof don't count. This decision is for convenience as much as anything, and nothing deep hangs on it.

Here's a very elementary toy example to illustrate some of these definitions. Consider a trivial pair of theories, T_1 and T_2, whose shared language consists of the (interpreted) propositional atoms 'p', 'q', 'r' together with all the wffs that can be constructed from them using the familiar propositional connectives, whose shared underlying logic is a standard natural deduction system for propositional logic, and whose axioms are respectively

T_1 : $\neg p$,

T_2 : $\neg p$, q, $\neg r$.

T_1 and T_2 are then both axiomatized formal theories. For it is effectively decidable what is a wff of the theory, and whether a purported proof is a proof from the given axioms. Both theories are consistent. Moreover, both are decidable theories; just use the truth table test to determine whether a candidate theorem really follows from the axioms.

However, note that although T_1 is a *decidable theory* that doesn't mean T_1 *decides every wff*: it doesn't decide e.g. the wff '$(q \wedge r)$', since T_1's sole axiom doesn't entail either '$(q \wedge r)$' or '$\neg(q \wedge r)$'. To stress the point: it is one thing to have an algorithm for deciding what is a theorem; it is another thing for a

(NTI), things identical in one possible world cannot be distinct in another. Sharvy offers a proof that NTI follows from LL (1968, p. 311).

Armed with these principles, Sharvy argues as follows. C-SC and C-ESC had the same members in 1968; therefore, by PE, they were identical in that year. So we have:

[1] In 1968, C-SC = C-ESC.

From (2) and NTI we next deduce that C-SC and C-ESC are identical *simpliciter*:

[2] C-SC = C-ESC.

Now the assumption about C-ESC was

[3] C-ESC never changes in membership,

and this in combination with [2] yields

[4] C-SC never changes in membership.

That is, the supposedly variable class cannot vary after all. Q. E. D.

This is fine as far as it goes, but we are left to wonder why premise [2] is true. No doubt it is true, but why? Isn't that just the question we want answered? Sharvy may have chosen to rely on classes here to be invariant in some art, but why must classes such as C-SC, be invariant in the first place?

A similar problem would arise if we tried to adapt Sharvy's strategy to answer my question. The adaptation would presumably use in place of NTI the principle of no *universe* transfer (NCI): things identical in one possible world cannot be distinct in another. And in place of (2) it would use the assumption that there are certain sets that have the same membership in all possible worlds. But what is the ground for this assumption? That is the very thing we want explained.

Whatever the answer, it still have to involve principles in addition to PE and NCI. To see this, consider another PIu, the *Identity of Indiscernibles*:

(ID) Necessarily (if F_x has F $\leftrightarrow$ y has F) $\rightarrow$ $x = y$.

This is analogous to PE, so if PE and NCI implied that a set has all its members essentially, ID and NCI ought likewise to imply that a thing has all its properties essentially. But of course there is no such implication.

The promise of digital documents

This is the grand old dream of radical new functionality.
(cf Paul Otlet, Vannevar Bush, Douglas Engelbart, and Ted Nelson)

- computationally available data items accessible with discipline-specific tools (chemical formulae, proteins, equations, etc.)
- advanced navigation and viewing optimized for domain-specific browsing and analysis,
- typed hypertext linking with links as first class objects,
- data-driven interactive diagrams and graphics
- computable equations,
- supportive ontological inferencing
- thoroughgoing interoperability with other tools

… and so on, and on, and on.

Are we achieving the promise of digital documents?

We are not.

What the problems are,

and why we have these problems

is the topic of the next video.

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.

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

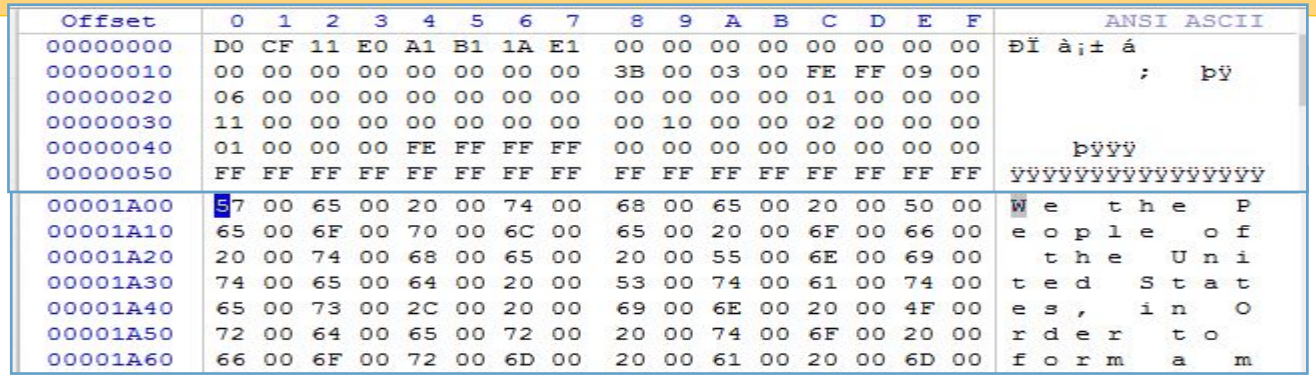
Creating complex, high-performance documents remains arduous and the results are hardly the sophisticated high-performance information environments we were promised.

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.


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...

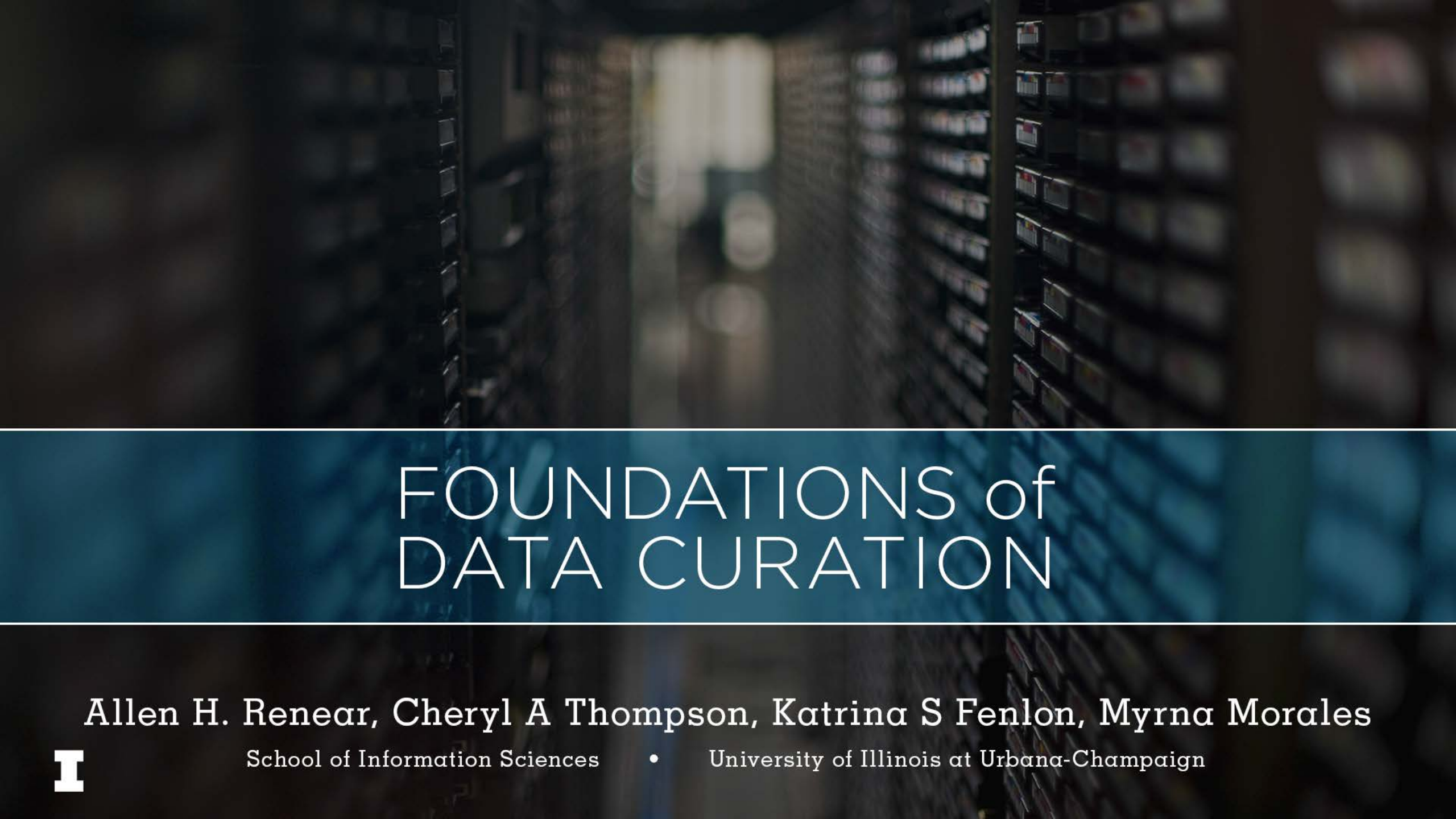
- easier to create
- easier to maintain
- easier to convert (new formats, new delivery software)
- better integrated with workflow in organization
- better integrated with other applications and tools (databases, word processing templates, indexes,)etc.
- more accessible to varied audiences
- easier 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.



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

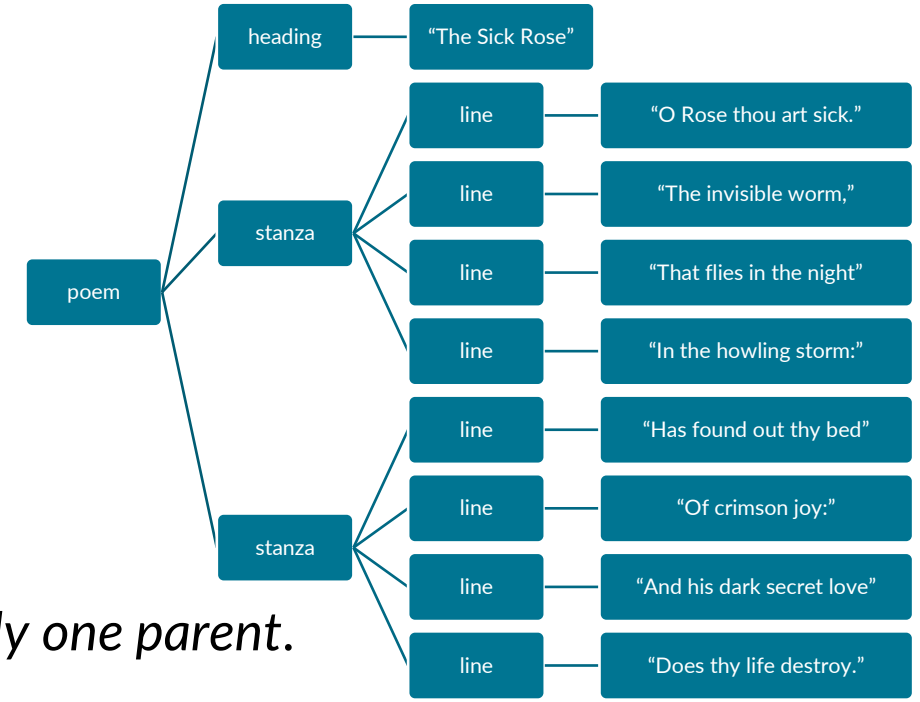
Text is an *O*rdered *H*ierarchy of *C*ontent *O*bjects

- *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*.

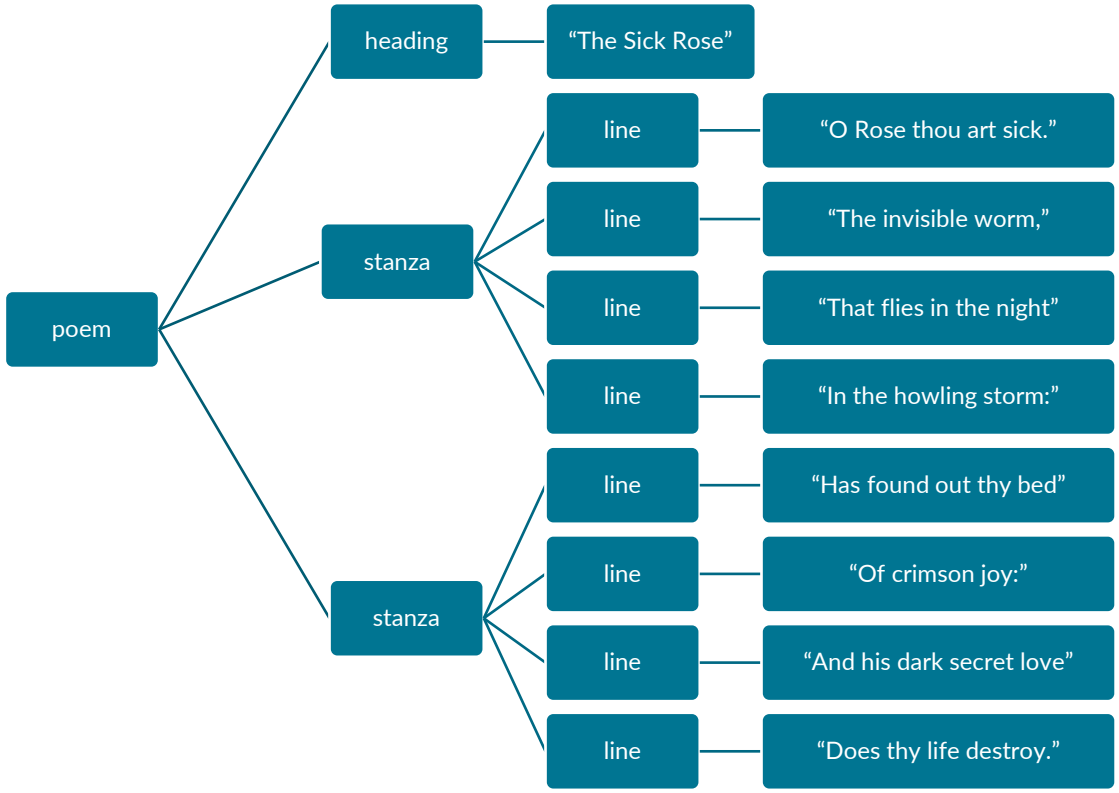
A tree, in our sense*, is

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.


5
WHY THE SOLUTION WORKS

Clarification: these models have two parts

The heart of the relational model can be understood as the combination of the relational data structure with attributes for data values.

Similarly the tree model, as it is typically implemented, combines the tree data structure with descriptive markup node labels (or in XML terminology “generic identifiers”) such as “`stanza`”.

We will now refer to the tree model as the ”tree/DM” model.
The phrase “relational model” does not need the clarification as the role of attributes is commonly understood.

Drum roll

Can we say in general terms why the tree/DM model works so well?

Why it succeeds so well in meeting data management challenges?

We can.

It works the same way the relational model works: *Abstraction* and *Indirection*

Abstraction

Both models focus on the data *itself*, separate from storage and processing.

This explicit identification of *data attributes* in one case, and *logical text objects* in the other, brings enormous new functionality and efficiency

Once again: *Indirection*

Both models support an indirect relationship to storage and processing,
but in practice the emphasis is often different:

For the relational model abstracting away from storage is dominant
For the tree/DM model abstracting away from processing is dominant

In both case the separation is mediated by a mapping:
logical schema to physical schema in the case of the relational model
text component (type) to processing instructions in the case of trees/DM

Formal vs colloquial understanding of these data models

Although in the relational model we commonly think of attributes as representing dyadic properties or relationships in the world, technically they are names for domains of values,

.

Similarly in the tree/DM model we think of these node labels as indicating the kind of enclosed text object (stanza, formula, etc), but in the model they are really simply names.

It is our colloquial understanding of these models that enables us to actually use them to secure useful abstraction and data management..

This may seem a small point now, but it motivates a further advance in abstraction, as we well see when we discuss ontologies.

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.

6

IMPLEMENTING THE SOLUTION: XML

Implementing the solution: XML

- XML: a schema language (or: a meta-grammar)
    - XML schemas
    - XML documents
- Example of an XML schema (DTD)
- Example of XML processing
- XML tools
- XML languages

XML

An XML document uses a defined set of delimiters with arbitrary element names and attribute value pairs to nest spans of text.

A *well-formed* XML document fits a formal grammar (along with other constraints) that ensures the document can be parsed as a tree by an XML parser.

> NB: a well-formed XML document need not have a schema that defines the element vocabulary and grammar;
>
> > it may use arbitrary element, attribute, and value names and arrange text objects in any way that does not violate the tree data structure.

The two main things in the XML world

Schemas [such as Document Type Definitions (DTDs)]

- One for each document type (class, category, genre)
- Defines a markup language for document structures by specifying its vocabulary and syntax (grammar)
- What elements can occur in documents of a particular type, what patterns these elements may form, what other information can be included about these elements?

Document Instances

- Particular documents, marked up with a markup language that meets well-formedness constraints, and, perhaps, also meets the constraints of a relevant schema.

A well-formed XML document

```
<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)

Schemas for Trees

This is an XML Document Type Definition (DTD), defining an XML document type

```
<!ELEMENT anthology (poem+)>
<!ELEMENT poem (title?, stanza+)>
<!ELEMENT title (#PCDATA)
<!ELEMENT stanza (line+) >
<!ELEMENT line (#PCDATA) >
```

This schema specifies element vocabulary and grammar

The DTD schema language is based on (extended) Backus Naur Form (BNF) grammars

Some XML schema languages provide additional validation and constraints on content, including data typing.

Another DTD

```
<!ELEMENT poem    (title, author? verse) >
<!ATTLIST poem
              editor    CDATA             #REQUIRED>
<!ELEMENT verse   (stanza+)>
<!ELEMENT stanza  (line+)>
<!ELEMENT title   (#PCDATA | italic | persname)*>
<!ELEMENT author  (#PCDATA)
<!ATTLIST author
              sex       (male | female)   #IMPLIED
              dates     CDATA             #IMPLIED
              bio       IDREF             #IMPLIED>
<!ELEMENT line    (#PCDATA | italic | persname)
<!ATTLIST line
              lang      CDATA             "ENGLISH">
<!ELEMENT italic  (#PCDATA)>
<!ELEMENT persname(#PCDATA)>
```

Another XML document, with attribute/value pairs

```
<!DOCTYPE text SYSTEM ”poem.dtd">

<poem editor=“Sara Porter”>


<title>Terence</title>
<author person=“N320”>A. E. Houseman</author>

<verse>
<stanza>
<line>Terence this is stupid stuff  </line>
<line>you eat your victuals fast enough</line>
<line>there can’t be much amiss ‘tis clear</line>

</stanza>
<stanza>
[…]
<line  lang=“latin”>The old lie: </line>
<line>  in vino veritas  </line>
</stanza>
</verse>

</poem>
```

Valid XML Documents

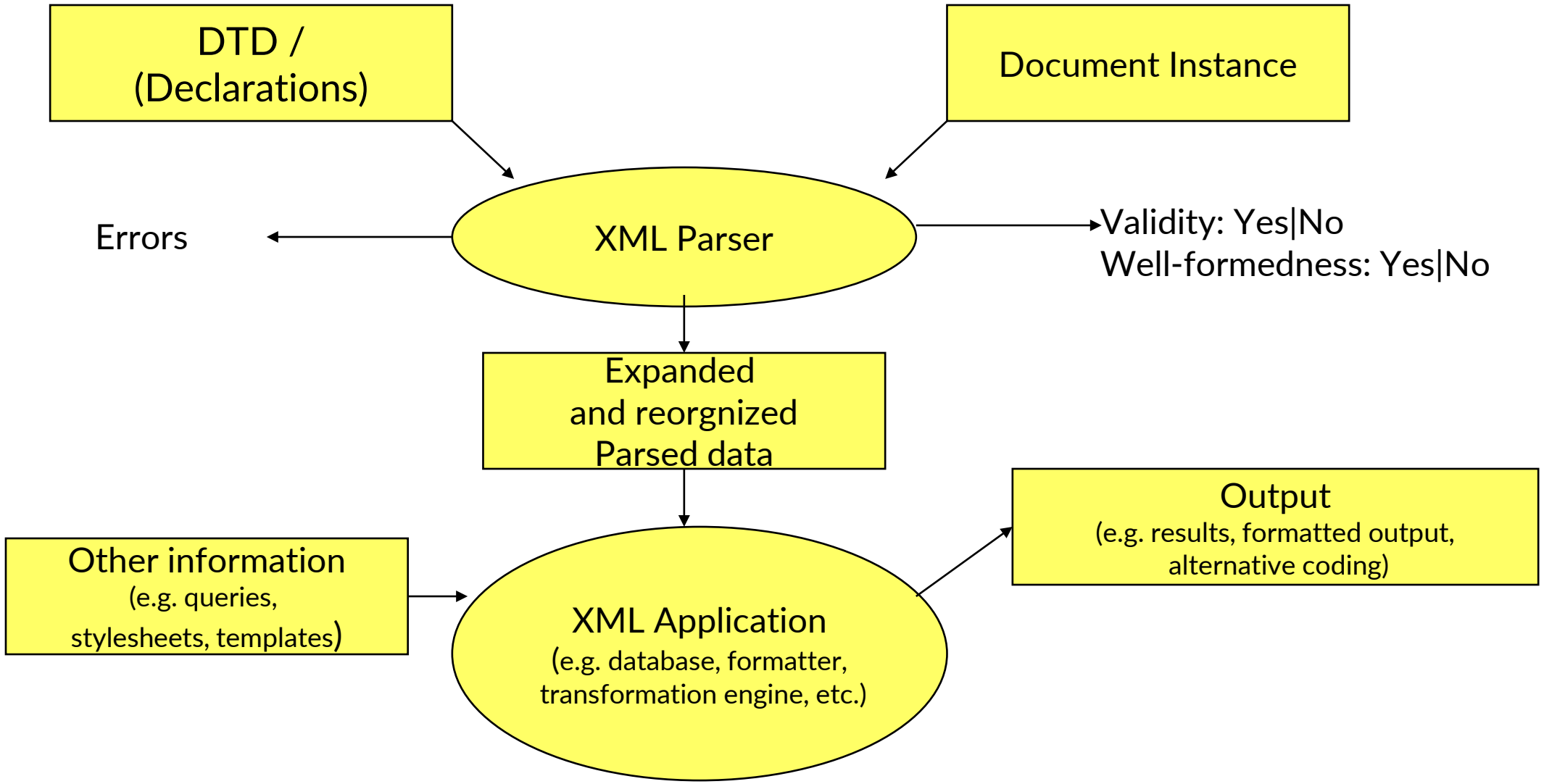
A document instance is *valid* with respect to some schema if it conforms to the declarations in that schema, which is to say, matches the grammar and other constraints.

> . . . nothing out of place, nothing missing, no attributes with values the wrong type, no references that fail, and so on.

A *validating parser* applies an XML schema to an XML document and determines whether or not the document conforms to the constraints specified in the schema.

All valid XML documents are well-formed, but not vice versa.

XML Processing



Some XML tools and schema languages

Two important XML transformation tools

XSLT: “a language for transforming XML documents into other XML documents” https://www.w3.org/TR/xslt

Xquery: “a standardized language for combining documents, databases, Web pages, and almost anything else” https://www.w3.org/XML/Query/

Two other XML schema languages

XML Schema (XSD): “the XML Schema Definition Language…offers facilities for describing the structure and constraining the contents of XML documents”

http://www.w3.org/XML/Schema

- A more complex schema language than DTDs, but does more than validate
- Written in XML
- Common for business applications

RelaxNG: “a schema language for XML” http://relaxng.org/

- Similar expressiveness to XSD, with simpler syntax
- Less commercial application support

Some important XML languages for documents

You should be familiar with these. Please explore the websites.

XHTML: “a family of current and future document types and modules that reproduce, subset, and extend HTML”

https://www.w3.org/TR/xhtml1/

TEI: Text Encoding Initiative – “a standard for the representation of texts in digital form”

http://www.tei-c.org/index.xml

JATS: Journal Article Tag Suite “defines a set of XML elements and attributes for tagging journal articles”

https://jats.nlm.nih.gov/

XML Languages and Interchange

There are an enormous number of XML markup languages.
See https://en.wikipedia.org/wiki/List_of_XML_markup_languages

Most of these languages are not for text, but are interchange and preservation formats for structured data.

XML is an important preservation format. It use of simple ASCII text with inline tags, can be parsed without a schema, and if a schema is available can be validated ensuring a correct grammar (nothing missing, nothing out of place) and data typing.

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.