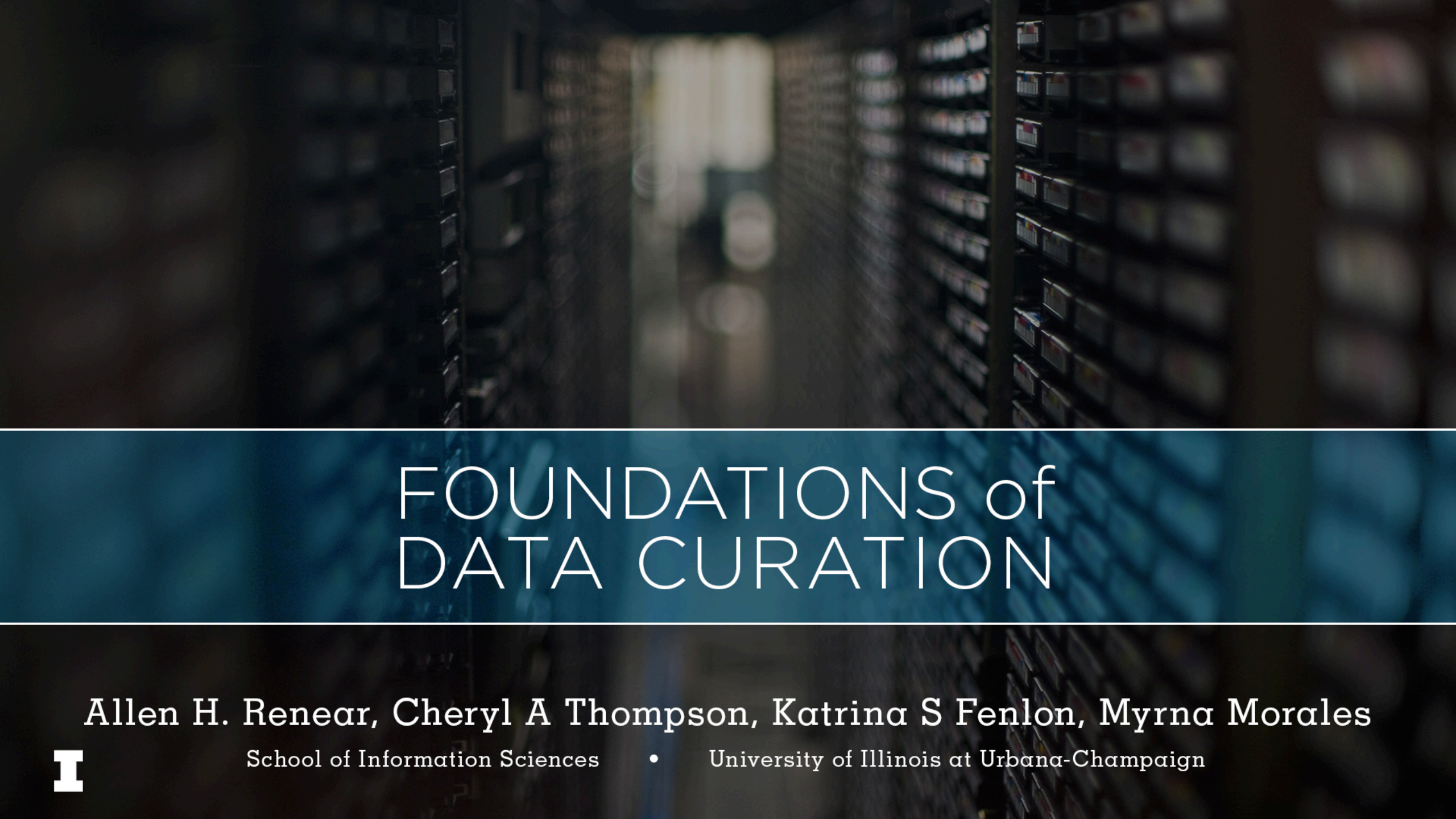


FOUNDATIONS of DATA CURATION

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:

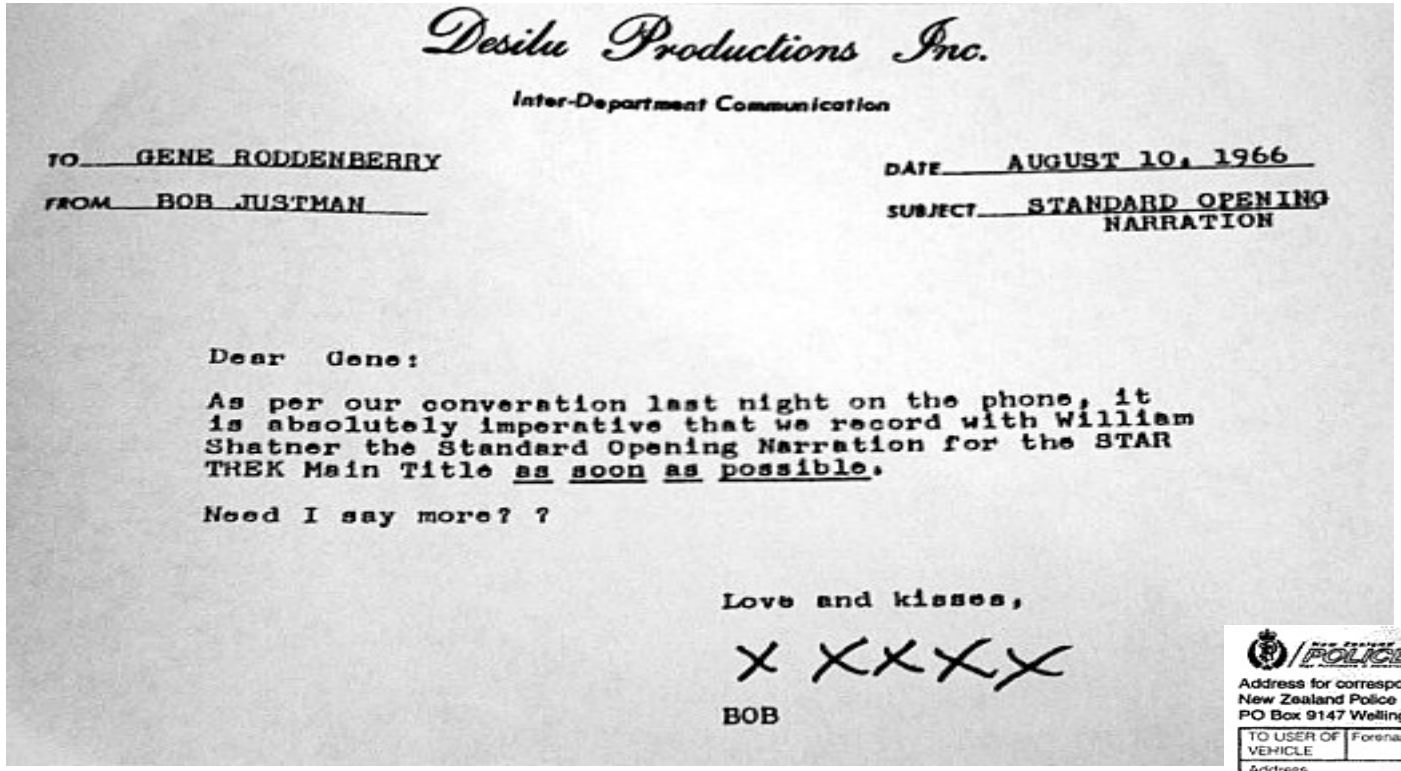
Particularly where the data has this form: *something* has a *value* for an *attribute*

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:

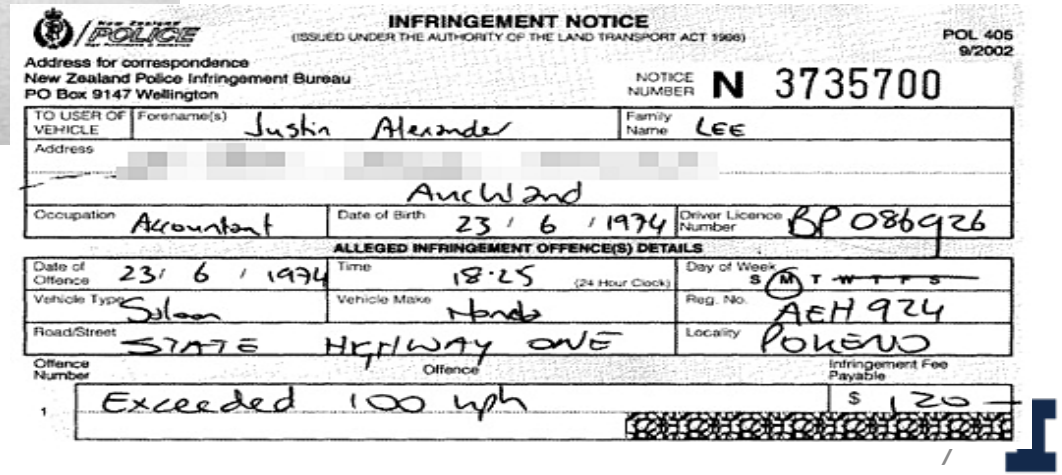
As per our converation last night on the phone, it is absolutely imperative that we record with William Shatner the Standard Opening Narration for the STAR TREK Main Title as soon as possible.

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

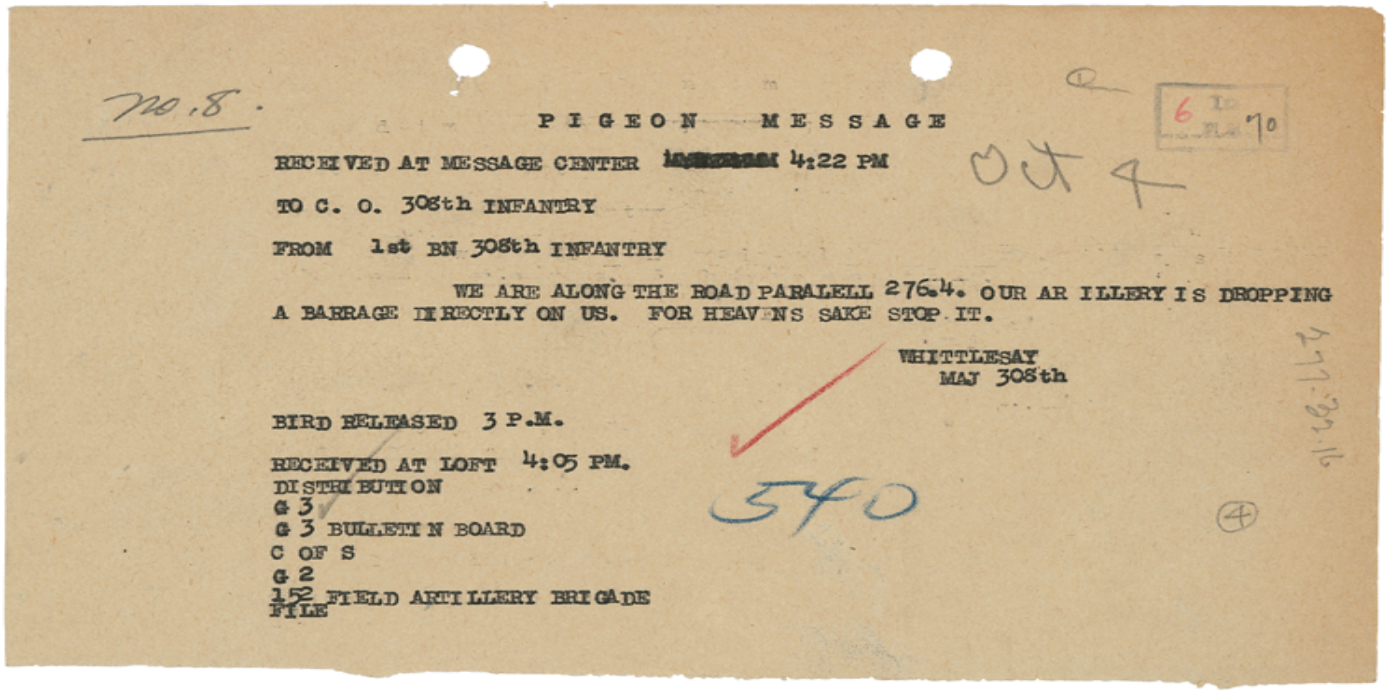
TO USER OF VEHICLE	Forename(s) Justin Alexander	Family Name	LEE
Address	[redacted] Auckland		
Occupation	Accountant	Date of Birth 23 / 6 / 1974	Driver Licence Number BP 086926

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

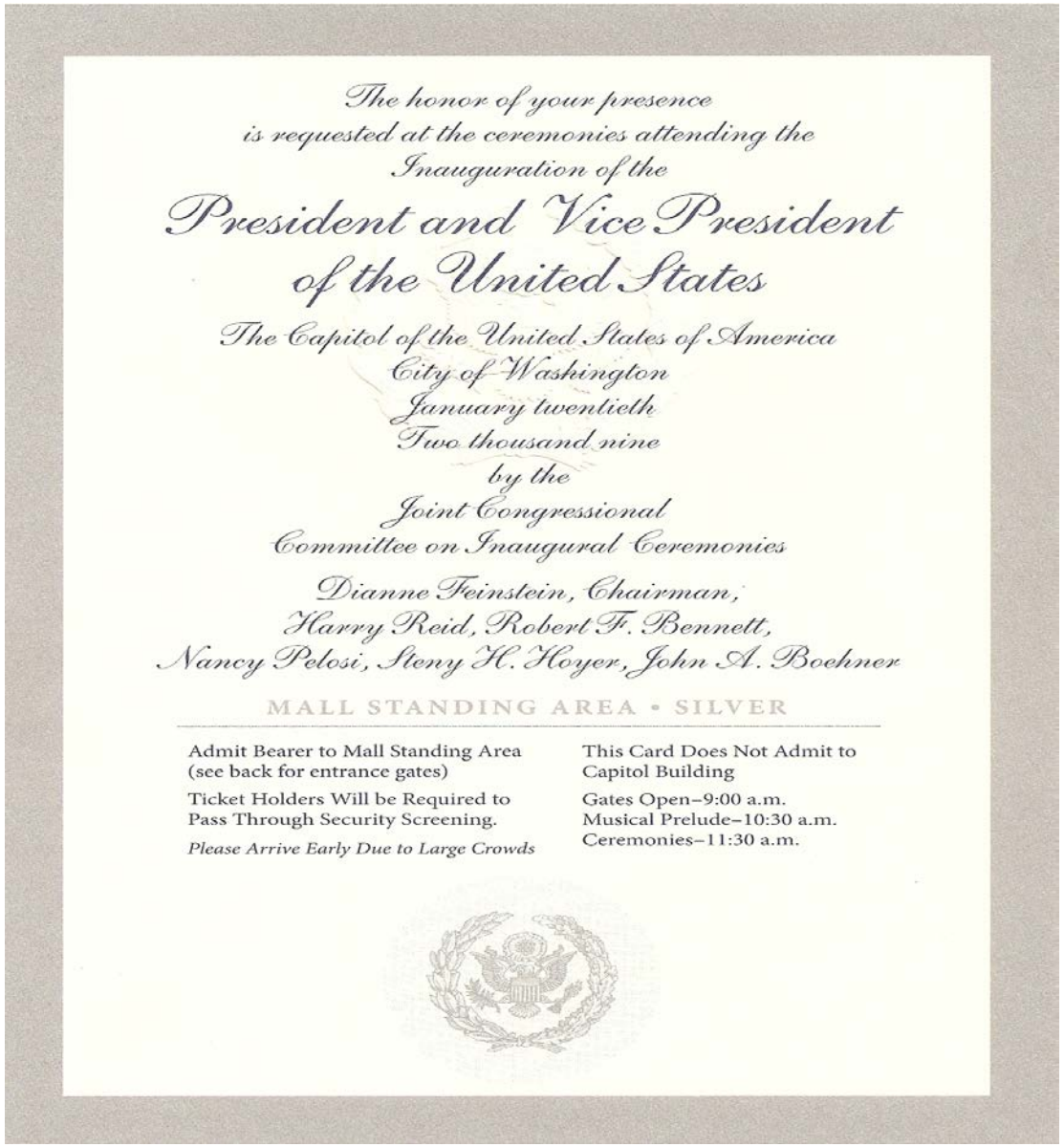
WE ARE ALONG THE ROAD PARALELL 276.4. OUR AR ILLERY IS DROPPING A BARRAGE DIRECTLY ON US. FOR HEAVENS SAKE STOP IT.

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

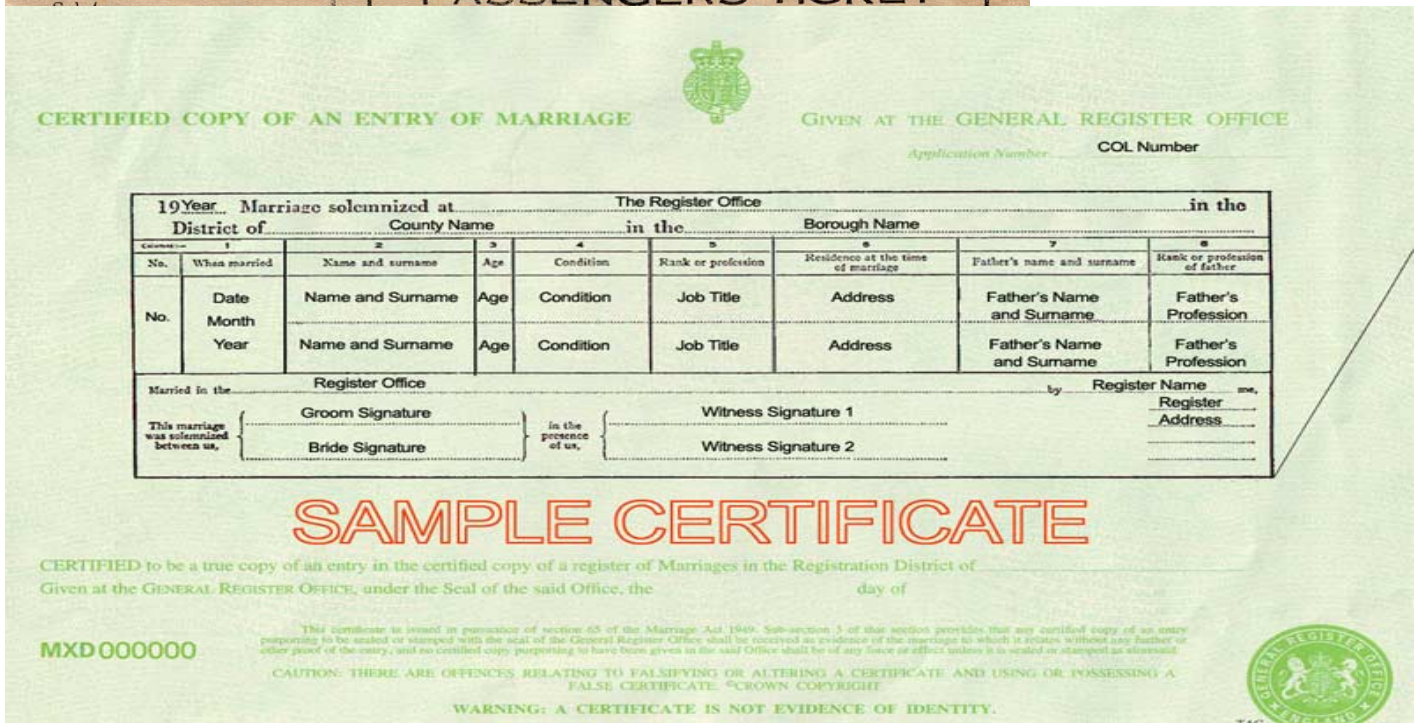
Tele:"DAILYERVICE" Phone: 32.

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

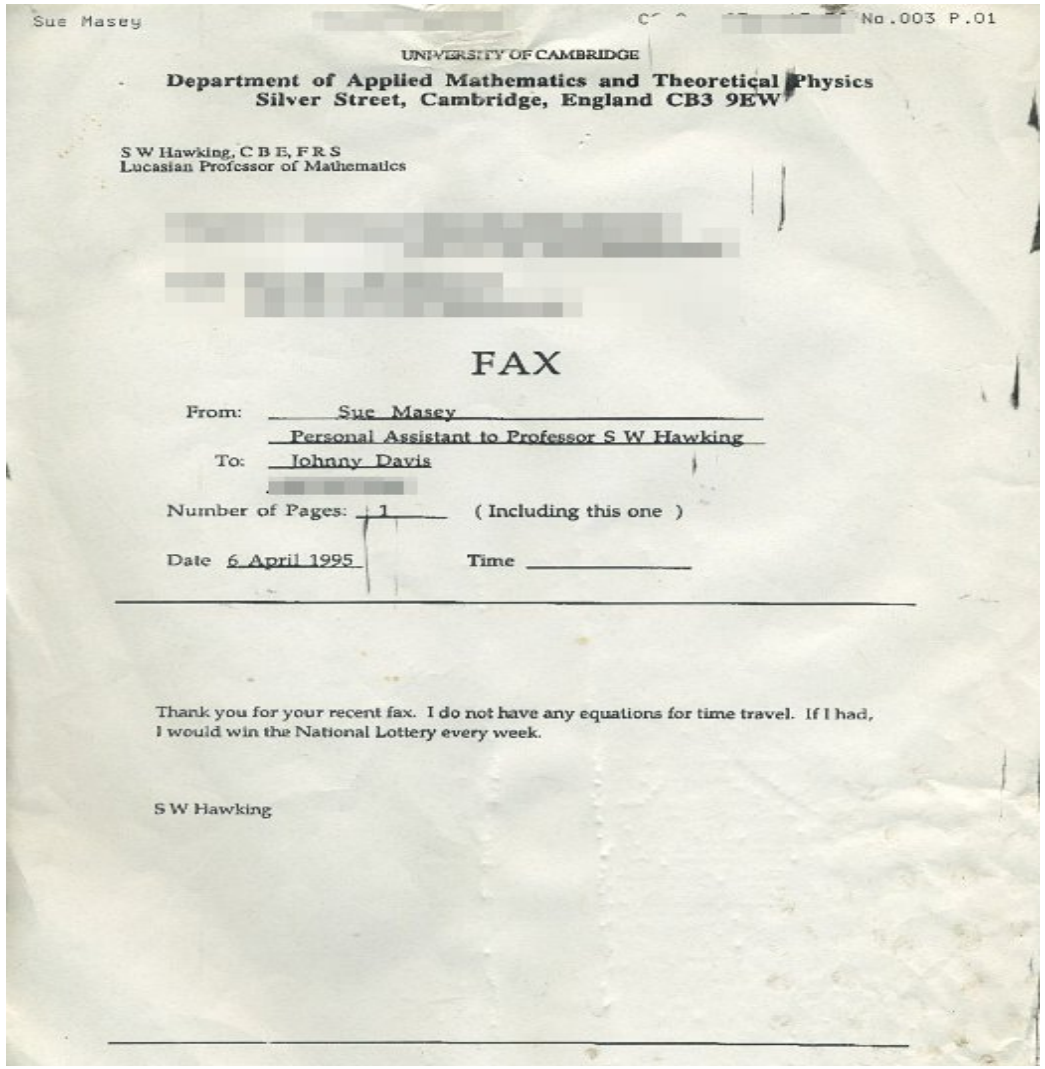
Given at the GENERAL REGISTER OFFICE, under the Seal of the said Office, the day of

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 $\langle C_i' c_j' \rangle$ for a reaction involving two gas phase components i and j. The second-order equation describing the flux term takes the form

$$\frac{\partial \langle w'C_i' \rangle}{\partial t} + \langle w \rangle \frac{\partial \langle w'C_i' \rangle}{\partial z} - \frac{\partial \langle w'w'C_i' \rangle}{\partial z} - \langle C_i' \frac{\partial P'}{\partial z} \rangle + \langle w'R(c_i) \rangle \quad (2)$$

(1) (2) (3) (4) (5)

Here term (1) is the diffusional transport term; (2) is the vertical gradient of the turbulent flux; (3) is the change due to buoyancy (G); (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:

$$\frac{\partial \langle w'C_i' \rangle}{\partial t} = \frac{g}{\theta} \frac{\partial \langle w'w'C_i' \rangle}{\partial z} + \langle w'R(c_i) \rangle - \frac{\partial \langle w'C_i' \rangle}{\partial z} \frac{2\langle w'w' \rangle_{(1)} + 2\langle A' \rangle_{(2)}}{\frac{P'+C'}{\tau}} \quad (3)$$

$$\langle w'R(c_i) \rangle = K_{11}\langle wNO \rangle \langle wO_3 \rangle + K_{12}\langle O_3 \rangle \langle wNO \rangle \quad (7)$$

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:

$$\langle w'C_i' \rangle = \frac{2K}{\alpha l} \langle w'^2 \rangle \frac{\partial \langle C_i \rangle}{\partial z} \quad (8)$$

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 45 levels laid out on a log-normal axis to give highest resolution in the lowest 50 m of the model. The model is solved with 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 Inex: A Proposed Measure for Long-Term Performance	July 2012
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Gordon D. Proctor, Shobna Varma, Steve Varnedoe	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Gordon Proctor & Associates, Inc. 7825 Wiltshire Drive Dublin, Ohio 43016 StarIsis Corp. 3737 Woodstone Drive Lewis Center, Ohio 43035 National Center for Pavement Preservation 2857 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

NEUTAJOVANÉ M9-0144G-50420-00

Slaňovací rám SR-1

Provozní kontroly

Strana

tabulek Strana

obrázků Strana

Odkazy

M9-0144G-50420-00 NEUTAJOVANÉ

Postup

1 Provozní kontrola funkčnosti slaňovacího rámu SR-1

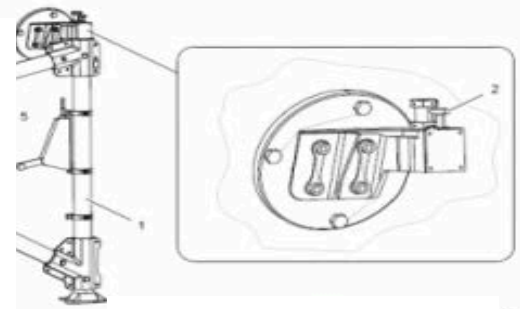
1.1 Ujistěte se, že SR (Obr. 1 [1]) je v transportní poloze.
1.2 Vytáhněte držadlo (Obr. 1 [2]) a rukou umístěte SR (Obr. 1 [1]) do sklopné polohy.
1.3 Uvolněte držadlo (Obr. 1 [2]):
- Držadlo (Obr. 1 [2]) se musí automaticky zasunout ve sklopné poloze SR.

1.4 Ujistěte se, že SR (Obr. 1 [1]) je zajištěn ve sklopné poloze.
1.5 Zvedněte konzolu závěsu (Obr. 1 [3]) do horní polohy a zajistěte čepem kluzáku (Obr. 1 [4]):
- SR (Obr. 1 [1]) je v pracovní 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

4 Effectively axiomatized theories

1. 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*: so wffs with free variables derived as we go along through 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

$T_1 : \neg\mathsf{p},$

$T_2 : \neg\mathsf{p}, \mathsf{q}, \neg\mathsf{r}.$

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 '$(\mathsf{q} \wedge \mathsf{r})$', since T_1's sole axiom doesn't entail either '$(\mathsf{q} \wedge \mathsf{r})$' or '$\neg(\mathsf{q} \wedge \mathsf{r})$'. To stress the point: it is one thing to have an algorithm for deciding what is a theorem; it is another thing for a

588 NOÛS

(NTI): things identical at one time cannot be distinct at some other time. Sharvy offers a proof that NTI follows from LL (1968, p. 511).

Armed with these principles, Sharvy argues as follows. C-SC and C-69SC had the same members in 1969; therefore, by PE, they were identical in that year. So we have:

(1) In 1969, C-SC = C-69SC.

From (1) and NTI we next deduce that C-SC and C-69SC are identical *simpliciter*:

(2) C-SC = C-69SC.

Now the assumption about C-69SC was

(3) C-69SC 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 (3) is true. No doubt it *is* true, but *why*? Isn't that just the question we want answered? Sharvy may have shown us why all classes have to be invariant if some are, but why must classes such as C-69SC 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 *contingent* identities (NCI): things identical in one possible world cannot be distinct in another. And in place of (3) 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 will have to involve principles in addition to PE and NCI. To see this, consider another PIs, the *Identity of Indiscernibles*:

(ID) $\mathrm{Nec}(\forall x)(\forall y)\,[(\forall F)(x \text{ has } F \iff y \text{ has } F) \Longrightarrow x = y]$.

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.