ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

An Eliminativist Ontology of the Digital World—and What It Means for Data Curation

MITH Digital Dialogues
Maryland Institute for Technology in the Humanities
University of Maryland, October 15nd 2013

Researchers: Dave Dubin, Karen M. Wickett, Simone Sacchi, Allen H. Renear
Mistakes and infelicities: Allen H. Renear

Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Slides: Renear, updated Oct 27 2013

NSF/OCI-ITR DataNet Award #0830976
IMLS/LB Award #RE-05-08-0062-08

INSTITUTE of Museum and Library SERVICES

Papers on modifiability

When Digital Objects Change — Exactly What Changes?
Proceedings of the American Society for Information Science and Technology,
Allen H. Renear, David Dubin, Karen M. Wickett (2008).

Documents Cannot Be Edited.
Proceedings of Balisage: The Markup Conference
Allen H. Renear, David Dubin, Karen M. Wickett (2009).

There are no Documents.
Proceedings of Balisage: The Markup Conference.
Allen H. Renear and Karen M. Wickett (2010).

Definitions of Dataset in the Scientific and Technical Literature.
Proceedings of the American Society for Information Science and Technology
Allen H. Renear, Simone Sacchi, and Karen M Wickett (2010)

Are Collections Sets?
Proceedings of the American Society for Information Science and Technology
Karen M. Wickett, Allen H. Renear, and Jonathan Furner (2011)

4.2.3 Preservation Model 1.0. [Echo DEPository 2008-2010]
Final Report of Project Activities. Sandore, B & Unsworth, JM (eds.)
David Dubin (2010)

Meta-Ontology

Eliminativism

Some of the things we think exist … don't.

Eliminativism in information science

When we try to express our models exactly and literally

we sometimes encounter problems

where elimination seems the best solution

The Modification Puzzle

In particular,

Modeling processes involving *change* and *identity* can lead to eliminativist conclusions

(courage…)

Change in the digital world

The digital world appears to be a place of constant change.

And yet

we are all deluded: *digital objects are immutable*

Our questions

1. When a digital object changes — *exactly what changes?*

2. If digital objects can't change, what is *really* going on in the world when say (speaking loosely) that they change?

3. *What are digital objects anyway?*

The Verona Sentence

Consider the sentence

"I remember Verona."

Let it be the first sentence of the first chapter of a draft of a novel.

Suppose the author edits this sentence to read:

"I remember, but dimly, Verona".

The first sentence of the draft has been modified; it is now longer.

But exactly *what* got longer?

"I remember Verona."? *no…*

"I remember, but dimly, Verona"? *no…*

The paragraph? The chapter? The entire text of the draft? *no*…

In search of *x, the thing that got longer*

What is modification?

A thing, x, loses or gains a property right?

But in the case of the Verona sentence:

there is no plausible candidate for this thing, x.

That is, the following assertion is false:

$(\exists x)$ [hadLength(x,t_1,3) & hadLength(x,t_2,5]

[something was 3 words long at one time, and 5 words long at another]

"there must be a substrate [*ὑποκείμενον*]
underlying all processes of becoming and changing.
What can this be in the present case?"
(Aristotle, Physics Bk II 226a)

What is the x, the subject/substratum, the *ὑποκείμενον,* of change??

The part where we take [some of] it back…(1)

We are not *totally* insane. (No, really we aren't.)

We agree that

"The first sentence was 3 words and is now 5"

can express a true proposition.

But we deny that it expresses the proposition

$(\exists x)$ [hadLength(x,t1,3) & hadLength(x,t2,5]

That is, we deny that

"The first sentence was 3 words and is now 5"

is *literally* true

The part where we take [some of] it back…(2)

"Jane lengthened the first sentence of her novel." is an *idiom*.

Compare: "The average plumber has 3.2 children"

$(\exists x)$ [isaAveragePlumber(x) & NUMCHILDREN[x]=3.2]

Or "There is a scarcity of common sense in this room.

$(\exists x)$ [isaScarcityofcommonsense(x) & isInthisRoom(x)]

Similarly

"Lumbergh revised the TPS memo." can express a true assertion.

But that assertion is not:

There is something, a TPS memo, that was revised by Lumbergh.

$(\exists x)$ [isaTPSmemo(x) & Revisedby(x,Lumbergh)]

The price of metaphor

The price of metaphor is eternal vigilance.

R. C. Lewontin,
"Models, Mathematics, and Metaphor", *Synthese* 1963.
quoting A. Rosenblueth and N. Wiener

Now let's do this from the top…

What's wrong with this argument:

1) All documents are strings.

2) No string can be modified.

∴ 3) No document can be modified

Classifying our options

An inconsistent triad

1. Documents are strings
2. Strings cannot be modified.
3. Documents can be modified.

Options … and obligations

A) Reject 1):

*Obligatio*n: offer an alternative definition of document,
one that supports modification.

B) Reject 2):

Obligation: reconcile modification with extensionality of strings/sets

C) Reject 3):

Obligation: provide an explanation of apparent modification.

MITH feud (2013): *Audience says* . . . Reject 1? 2? 3? ???

1) All documents are strings

From the XML specification:

> "Definition: A textual object is a well-formed XML document if: Taken as a whole, it matches the production labeled document…"

2.1 Well-Formed XML Documents
Extensible Markup Language (XML) 1.0 (W3C, 2008)

Not buying it?

- Ok, but there is no easy way out…
  - Our argument also works for other mathematical entities (e.g, ….a document is a kind *graph*)
  - And many non-mathematical definitions entail immutability
    - cf FRBR's notion of an *expression*
    - or Tanselle's notion of a *text*

2) Strings cannot modified

Modification requires losing a property (and surviving that loss).

But strings have no properties which it is possible for them to lose!

The string "13571" has properties like:

having a length of five tokens,
having one token type occur twice,
having the substring "35". … etc.

And those are properties that "13571" cannot lose.

This follows both from intuition,

and from the standard mathematical definition of string.

3) Documents can be modified

That documents can be modified is implied by much of what we routinely say and do in text processing and digital publishing.

"The TPS memo has been revised … it used to be three pages"

Some responses

- **The Materialist Strategy**

  Denies all documents are strings

  Argues that documents are material things, which can change

- **The Social Object Strategy**

  Denies all documents are strings

  Argues that documents are social objects

- **The New Document theory**

  Denies documents can be modified

  Argues that modification is a new document)

- ***The String-In-A-Role strategy**

  Denies documents can be modified

  Argues that a document is *a string-in-a-role*

String-in-a-role strategy*

Denies: *Documents can be modified.*

but also finesses the definition of document

A document is a string in a particular communicative role.

being a document is a property that strings have
only in particular *contingent* social situations

On this account **document** is not a *type* of entity (or kind of thing)
but being a document a *role* that some types of entities can have

more specifically *being document* is a role that *strings* have

Cf *person* and *student*,
being a student is a role that *persons* have, in certain circumstances

(Guarino & Welty 2001)

Roles vs types in ontology

Guarino and Welty distinguish *roles* and *types*:

If a property indicates a *type* then it is rigid: it is impossible for anything that has that property to not have that property (and exist).

Compare **person** and **student**

"…the ideal structure of an ontology has *types* in the "backbone" and *roles* hanging off the backbone".

Formally: A property ϕ is *rigid* =df $\Box(\forall x)(\phi x \rightarrow \Box\phi x)$

In possible world model semantics: a property ϕ is rigid =df if ϕ is had by some *x* in some possible world, then it is had by *x* in every possible world in which *x* exists.

Guarino & Welty (2000) A Formal Ontology of Properties.

Summary: A document is a string-in-a-role

First, a modified definition of document:

A document is a string in a communicative role

[draws on J. Searle; N. Guarino & C. Welty]

NB:
this entails that documents are strings (like other definitions)

Next meeting the burden: What is modification on this account?

roughly: a person or persons coming to prefer a different string for a particular communicative role than the string previously preferred.

How does it fare in the competition?

The string-in-a-role theory
a) introduces no new objects or relationships into our ontology,
b) has clear identity conditions (well, the string does)
c) provides a plausible alternative account of apparent document modification

Our short answer to the modification puzzle:

Apparent changes "*in digital objects*" are actually changes *in us*,

in the person or persons interacting with those objects,

and not changes in the objects themselves.

What changes when a digital object changes? *You do.*

(Ok, maybe it takes a village.)

The Eliminativist Response

Having trouble accepting that documents cannot change?

There's another way out.

Eliminativist responses to conceptual puzzles

The problem of "material constitution".

Consider a statue of a horse (S) and the bronze that composes it (B)

Does S = B? It would seem so.

But S has properties that B does not (e.g., being created this morning)

And suppose the statue is melted and reformed to a statue of a cat (S*)

Same bronze, different statue.

But if B=S, and B=S*, then S=S*; which we agree is false.

Eliminativist solution: *There are no statues*

And here's an equivalent statue-free paraphrase for "statue creation"

some bronze had its parts arranged horse-statuewise
and then later . . . cat-statuewise

Eliminativism as a solution to the puzzle

If there are no documents our puzzle is solved.

This is itself perhaps a reason to believe that there are no documents.

But we also have an alternative solution, the string-in-a-role theory.

So be preferred over the string-in-a-role solution the *no documents* solution needs additional support.

Let's see what can be provided along those lines.

Is there a positive argument for *no documents*?

A positive argument for *no documents…*
[suggested by Dan Korman, adapting Trenton Merricks]

It is commonly believed that documents can be
revised, edited, shortened, lengthened, and modified in various ways.
This belief is widespread, and deeply rooted.

Perhaps so deeply that it is integral to our concept of a document.

If so then we can express this relationship, in neutral terms, this way:

"if there are documents, then there are modifiable documents"

But we've shown that there are no modifiable documents.

So therefore there are no documents

The Argument from extensionality of sets

Digital objects are defined as kinds of strings, tuples, relations, graphs, etc.

These in turn are defined as kinds of *sets*

Sets are *extensional*: they cannot lose or gain members

This is a formal consequence of all standard set theories

.

$$(\forall S)\,(\forall T)\,(\forall x)\,[\,(S{=}T) \equiv (\forall x)(x \varepsilon T \equiv x \varepsilon S)\,]$$

sets S and T are the same if and only if they have the same members

Everyone believes that this follows immediately from the ZFC axiom of extensionality.
It doesn't, but it does eventually follow given a few other plausible assumptions

— James van Cleve "Why do sets have their members essentially?" (1985)

Collections are sets, therefore…

A collection is a set of objects

"Definition 17. A collection C {d1, d2 … dn} is a set of digital objects."

"Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries".
ACM Transactions on Information Systems,
Goncalves et al. (2004)

So a collection is a set

And therefore collections cannot lose or gain items

When "a" is added to {b,z} to get {a,b,z},
exactly *what* is the thing that once was {b,z} and now is {a,b,z}??

See "Are Collections Sets?" Karen Wickett, Allen H. Renear, Jonathan Furner.
ASIS&T Proceedings 2011.

Database tables are sets, therefore…

A database table (instance) is defined as a mathematical relation

r(R) $\subseteq$ (dom(A1) x dom(A2) x …. x dom(An))”

— El Masri & Navathe, *Fundamentals of Database Systems* (2006)

So a database table is a *relation* … a *set*; its records are *tuples*

And therefore tables therefore cannot lose or gain records

Adding a record to a table (or modifying a record)
is really just mapping from one table to another,
not modifying a persistent underlying entity

Cries from the heart

" the terms '**Data Product**', '**Data Set**,' and '**Version**' are overlaid with multiple meanings between communities." (Barkstrom, 2009)

"There is ambiguity in **what type of object a dataset is**; with different groups of users applying different connotations

There needs to be an explicit statement of what the intended preservation of a dataset will imply." (Pepler, 2008)

[is there any more unanimity for *text* or *document*?]

Documents are sets, therefore . . .

For instance, digital documents have been defined as strings:

From the XML specification:

"Definition: A textual object is a well-formed XML document if: Taken as a whole, it matches the production labeled document…"

2.1 Well-Formed XML Documents
Extensible Markup Language (XML) 1.0 (W3C, 2008)

So a document is a string

A string is a function f:N→A
from natural numbers into some codomain of elements.

So a string is subset of NxA, i.e. a string is a set.

And therefore strings cannot lose or gain elements

"Editing" strings is mapping from string to string,
not modifying a persistent underlying entity

Why bother?

Automated inferencing over formal ontologies is increasingly important

> "Strategic Reading, Ontologies and the Future of Scientific Publishing"
> Allen H. Renear, Carole L. Palmer,
> *Science*, **325**, 828 (2009); Aug. 16, 2009.

Such inferencing requires assertions that allow only literal interpretation, with compositional semantics and existential instantiation.

Humans communicate with natural language sentences such as

> "The sun rose in the east",
> "An fog of anxiety descended upon the congregation",
> "The average plumber has 3.2 children", or
> "The TPS memo was revised"

Naive formalization of our familiar discourse about documents fails this requirement.

Identity Problems

Two historians, Jill and John,
read *the same text*.

What does that mean?

And how can well tell?

Identity Problems

Two linguists, Jill and John,
used the same TEI document.

What does that mean?

And how can well tell?

Identity Problems

Two scientists, Jill and John,
used *the same data*.

What does that mean?

And how can well tell?

Identity Problems

Compare:

Two scientists, Jill and John,
used *the same statistician*.

Identity Problems

Compare:

Two scientists, Jill and John,
used the same *centrifuge*.

Identity and Representation Levels

Consider two files with the

… same data,

but relational tables in one case

and RDF triples in another

… same data and the same RDF triples,

but an XML serialization in one case,

an N3 serialization in another

… data, the same RDF triples, the same N3 serialization,

but UTF-8 character encoding in one case

and UTF-16 encoding in another

How many of levels do we need?

How do we define and manage them?

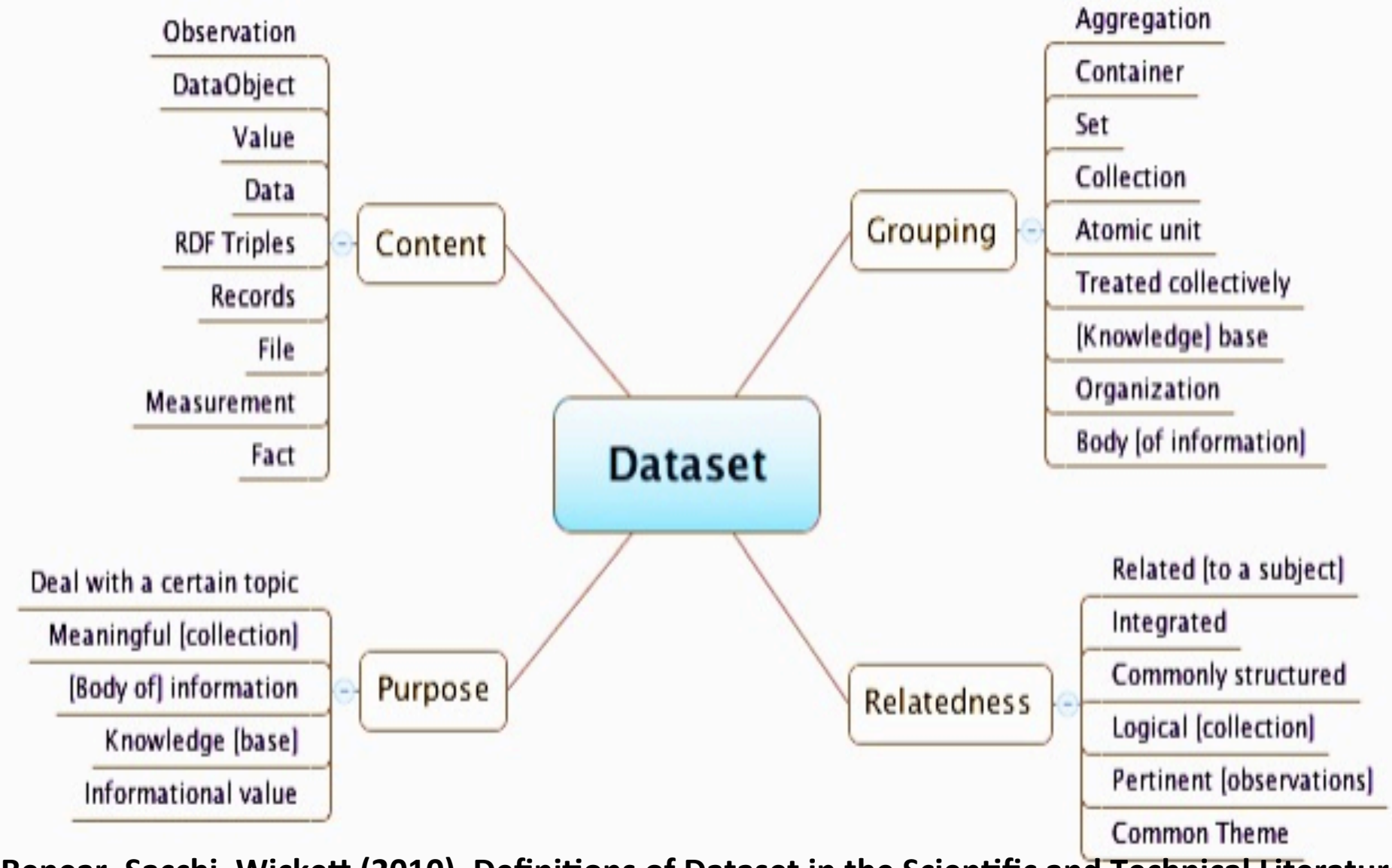
How can they be identified and re-identified?

What *is* a dataset anyway?

Maybe we should ask a scientist

They'll have an answer, right?

There are almost as many answers as scientists



Renear, Sacchi, Wickett (2010), Definitions of Dataset in the Scientific and Technical Literature

Concluding remarks

Ordinary discourse is full of idiom and metaphor
(and a good thing too)

But sometime literal precision is needed.

Ontological foundations can provide that.

But will replace our familiar world with a stranger one.

Papers on modifiability

When Digital Objects Change — Exactly What Changes?
Proceedings of the American Society for Information Science and Technology,
Allen H. Renear, David Dubin, Karen M. Wickett (2008).

Documents Cannot Be Edited.
Proceedings of Balisage: The Markup Conference
Allen H. Renear, David Dubin, Karen M. Wickett (2009).

There are no Documents.
Proceedings of Balisage: The Markup Conference.
Allen H. Renear and Karen M. Wickett (2010).

Definitions of Dataset in the Scientific and Technical Literature
Proceedings of the American Society for Information Science and Technology
Allen H. Renear, Simone Sacchi, and Karen M Wickett (2010)

Are Collections Sets?
Proceedings of the American Society for Information Science and Technology
Karen M. Wickett, Allen H. Renear, and Jonathan Furner (2011)

4.2.3 Preservation Model 1.0. [Echo DEPository 2008-2010]
Final Report of Project Activities. Sandore, B & Unsworth, JM (eds.)
David Dubin (2010)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Questions?

Research in these areas is being carried out by the *Data Concepts Group* and the *Conceptual Foundations Group* at the *Center for Research in Informatics and Scholarship* (CIRSS), Graduate School of Library and Information Science, at the University of Illinois at Urbana-Champaign,

Principal contributors include

David Dubin, Karen M. Wickett, Simone Sacchi,, Allen H Renear

NSF/OCI-ITR DataNet Award #0830976
IMLS/LB Award #RE-05-08-0062-08

INSTITUTE of Museum and Library SERVICES