Identifiers and Change

Acknowledgements: This unit draws on work carried out collaboratively at the School of Information Sciences, University of Illinois at Urbana-Champaign. The principal researchers (Allen Renear and Karen M Wickett) are grateful for conversations with many people, but are particularly indebted to Dan Korman and Dave Dubin. Thanks also to Simone Sacchi and Richard Urban and the other members of the iSchool seminar: *Identity and Change in Information Systems* and the research group *Conceptual Foundations of Information Science.*

The image part with relationship ID [illegible] was not found in the file.

The image part with relationship ID [illegible] was not found in the file.

Change

Managing data means managing change.

… transcoding, format conversion, reduction, extraction, correction, addition, generalization, specialization, and so on and on and on and on . . .

We manage change with fine-grained machine readable workflow documentation.

And identifiers are at the heart of this.

But which of the things in the workflow *actually change?*

Or, more generally, what do we mean by change?

Change in the digital world

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

digital documents are modified as sentences are revised

datasets undergo corrections

database tables are updated are records are inserted, and deleted

files get larger as more data is included

and so on.

And yet all this is an illusion

Most things in the digital world are absolutely immutable

Understanding immutability

The demonstration is simple and not restricted to digital objects:

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*.
(It was three words long and it is now five words long).

But exactly *what* got longer? What is it that was *three* words and is *five* words?

“I remember Verona.”? no, it was three words, but *it still is three words*

“I remember, but dimly, Verona”? no, it *is* five words, but *has always been*

The paragraph? chapter? the entire draft? no, those are just longer character strings

In search of ***x***

Before going further: what is *change*?

Losing or gaining a property right?
So, something (*x*) is modified (changed) iff
there is some time when it has a property F
and some later time when it does not have that property F.

And we don’t count merely relational changes. You don’t change (really, intrinsically), when someone in Iceland stops talking about you, even though you had the property of being talked about by someone in Iceland, and then you lost that property when they stopped talking about you.

But in the case of the Verona sentence:
there is no plausible candidate entity for the thing (x) that is modified

That is, nothing seems to fit ***x*** in the logical expression:

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

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

From Aristotle: “there must be a substrate [*ὑποκείμενον*] underlying all processes of becoming and changing *What can this be in the present case?*” (Physics Bk II 226a). But what is the x, the subject/substratum, the *ὑποκείμενον,* of change in the case of the Verona sentence.

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

We are not *totally* insane. (really we aren't.)

We don't deny that

“The first sentence was 3 words and is now 5”

can express a true proposition.

What we deny is that it expresses the proposition it appears to express:

(∃*x*) [hadLength(x,t1,3) & hadLength(x,t2,5]

We are denying that

“The first sentence was 3 words and is now 5 words”

is *literally* true

Idiom, metaphor, and other logical disasters

Compare: “The average plumber has 3.2 children”

$(\exists x)\ [\ \text{isaAveragePlumber}(x)\ \&\ \text{NUMCHILDREN}[x]=3.2\ ]$??

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

$(\exists x)\ [\ \text{isaScarcityofcommonsense}(x)\ \&\ \text{isInthisRoom}(\text{x})\ ]$??

Similarly in our case:

“Jane lengthened the first sentence.”

Does not mean: $(\exists x)\ (\exists y)\ (\exists z)\ [\text{hadLength}(\text{x,t1,y})\ \&\ \text{hadLength}(\text{x,t2,z}\ \&\ \text{z}>\text{y}\ ]$

Or, more generally

“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.

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

Is this just a trick?

Are we exploiting a “scope ambiguity” in “the first sentence in the novel”?

For comparison:

The *first person in the coffee queue* has changed:

A. At 1:00pm *the first person in the queue* was an old man
B. At 1:01pm *the first person in the queue* was a young woman.

But no one thinks that this means,

there exists an x such that x was an old man at one time and then at a slightly later time x was a young woman

And that is exactly the point:

Digital object modification is just like coffee queue processing.

The queue changes, the people don’t, despite sentences A and B above.

Note that you can’t replace the phrases *“the first person in the queue” with a uniquely referring (coreferential) proper name and get the same result.*

Why worry? *The price of metaphor*

The price of metaphor is eternal vigilance.

R. C. Lewontin,
“Models, Mathematics, and Metaphor”, *Synthese* 1963.
attributed to A. Rosenblueth and N. Wiener

The unforgiving nature of logic-based languages

Inferencing over formal ontologies is increasingly important

But it is based assertions in a logic-based data representation language
and such assertions allow only *literal* interpretation.

Humans on the other hand communicate with natural language sentences the deploy idiom, metaphor, and other rhetorical features than conceal logical form. For instance:

"The sun rose in the east",
"A fog of anxiety descended upon the congregation",
"The average plumber has 3.2 children",
"I edited the draft"
"The TPS memo was revised".

But computational inferencing requires literal interpretation*, including:

compositional semantics,
existential instantiation,
valid deductive inference.

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

*The underlying ontology need not reflect the latest theories of modern physics, but it should nevertheless at least be consistent.

Stitching this together

Now let's connect this with earlier discussions of identity and change

Usually when we are documenting change with respect to a digital object

The digital object itself is not changing (even though we say it is)

However its *relationships* are changing

And new things are appearing and playing new roles

For instance

We identify a dataset as temperature observations at a place and time

We convert that dataset from a JSON representation to a XML representation

Now what, precisely and literally, has changed (intrinsically)?

The observations haven't, they are the same.

And the representations haven't changed either,

yes, we have a new and different representations

but each is exactly what it is, has been, will be.

and no thing was once the old representation but is now the new.

Some practical conclusions

In a slogan: ➔ **Use identifiers to identify, not to describe** 🡐

Use non-descriptive ("opaque") identifiers to identify digital objects

e.g., "DS0021" for a dataset

Do not use descriptive expressions as identifiers

e.g., `temperatureArea32notValidated.csv`

such a descriptive identifier may

be factually incorrect
or need to be revised as things change
[and it makes lousy data storage system in any case!]

Recognize that digital objects do not undergo intrinsic (nonrelational) change*

But that they do undergo relational change,

e.g., DS0021 can change from being unreviewed to being reviewed.

(But such changes do not affect original object identity)

*obviously there are some challenges to this assertion that are best handled by compromise: sometimes we need a system model than allows *files* (for instance), to undergo byte-based changes. But when we do this we court trouble if we also, in the same model, define a file as a sequence of bytes, as we now have an inconsistent model.

If you are interested in more of this sort of thing. . .

"When Digital Objects Change — Exactly What Changes?"
In *Proceedings 71st Annual Meeting of the American Society for Information Science and Technology*.
Allen H. Renear, David Dubin, and Karen M. Wickett.. (2008).

“Documents Cannot Be Edited.”
In *Proceedings of Balisage: The Markup Conference*.
Allen H. Renear, and Karen M. Wickett. (2009).

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

The Argument: From extensionality of sets

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

All of these are in turn are defined mathematically as kinds of *sets*

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

This is a formal consequence of all standard set theories*

Two sets are the same if and only if they have the same members.

Many digital objects, like documents, can be defined as a string

A string is a function f:$\mathcal{N}$→A from the natural numbers into some codomain of elements.

So a string is subset of $\mathcal{N}$xA, i.e. a string is a set.

And therefore strings cannot lose or gain elements

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

*Although not immediately or without additional axioms. It is widely believed that that claim that sets have their members essentially 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).