V2: What is data preservation?

The common definitions are misleading

The problem: ***What exactly is preserved?***

Our answer: ***Nothing*** **is preserved**

Our definition: **[you'll see]**

Why the definition is important

Some common definitions of data preservation

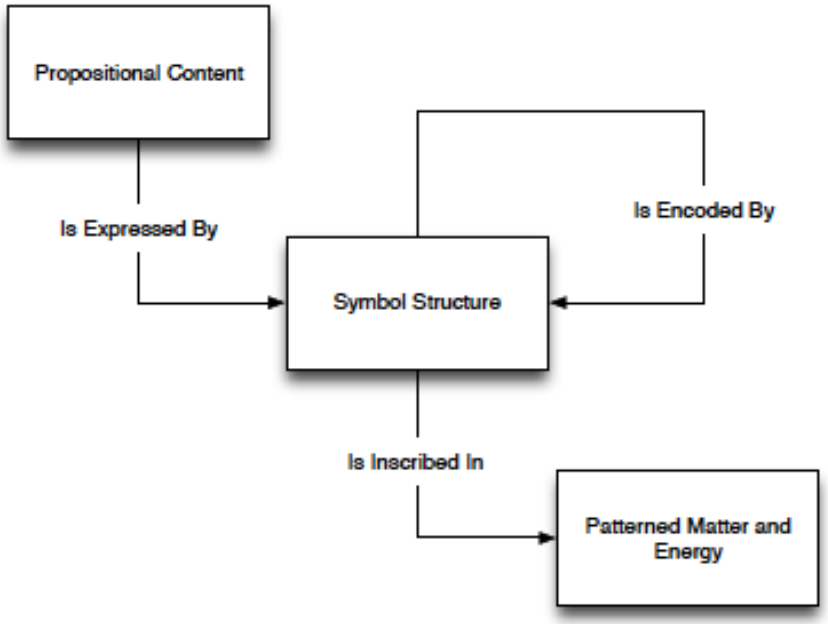
Preservation:

> ". . . The professional discipline of protecting materials by minimizing chemical and physical deterioration and damage to minimize the loss of information and to extend the life of cultural property . . . "
> (Society of American Archivists)

Digital preservation:

> "… the active management of digital content over time to ensure ongoing access." (US Library of Congress)

Recall our data ontology



Now ask yourself:

In a successful preservation scenario, *what exactly is preserved?*

What . . . exactly . . . is . . . preserved?

No, really, ***what?***

The physical thing?

Maybe things in the patterned matter & energy entity type?

That would be actual disks, thumbdrives, hard drives, raid arrays, etc.

Sure, we often do need to take care of these, absolutely

But over time preservation of data (or information) may be successful, even as the individual physical media disappear or become inoperable.

So the continuing existence and integrity of some physical object is not necessary for preservation.

Propositional Content?

Maybe things belonging Proposition Content entity type?

That would be that data or information itself.

e.g., observations, claims, assertions, “facts” etc.

But do those things really need preservation?

Consider this assertion:

100,000 people live in zip code area 61820

Why does its “preservation” require our intervention?*?*

(it can’t crumble, oxidize, dissolve, mold, etc.)

Such things cannot decay, and so cannot be the object of preservation

(at least not in the original and narrow sense of *preservation*).

Symbol Structures?

Maybe things of the Symbol Structure entity type?

But symbol structures also don't need preservation

consider: `up:Z61820 cb:population “100000”`

this is a repeatable sequence of tokens (or characters)
it also cannot oxidize fall apart, dissolve, get moldy, etc

And so it also does not need our help surviving the vicissitudes of nature.

[And in any case:
we often *support preservation* of data by deliberately *changing* symbol structures;
e.g., replacing an obsolete language or format with a different or newer one.]

But if those things then . . . ?

So what is preserved?

We just observed that it can't be any of the things in our data ontology!

The answer is:

Nothing is preserved.

Or perhaps less paradoxically:

No *thing* is preserved.

.

Data preservation is not about preserving data

Preserving physical objects (in the traditional sense of preservation) is part of data preservation, but only part.

> Preservation of physical media must occur for some intervals of time, but data preservation can also take place even when particular media decay — and, indeed, often we deliberately discard media as part of a data preservation strategy.

Preserving propositions (assertions etc) and symbol structures (encodings) is not part of data preservation all: because they cannot decay.

> And, moreover, their continued existence does not ensure preservation.

Preservation is not about preserving any *thing*

Data preservation is not about preserving the existence of objects

It is about *communication with the future*. [1]

The best simple definition is

Ensuring reliable communication with the future [2]

[1] Reagan Moore, Towards a theory of digital preservation, *The International Journal of Digital Curation*, 2008

[2] Simone Sacchi, *What do we mean by preserving digital information'? Towards sound conceptual foundations for digital stewardship* (Doctoral dissertation, University of Illinois at Urbana Champaign), 2015.

The definition expanded. . .

Preservation is not about preservation.

Preservation is *ensuring reliable communication with the future*

More exactly: preservation actions are intended to ensure that future researchers (or other users*)

1) will come into possession of physical media and encodings

2) from which they will correctly recognize the originally intended propositional content

3) and from which they will be justified in believing that this propositional content is in fact the intended propositional content

* this can be adapted to more explicitly accommodate software agents and automatic processing. The key thing is that the process is reliable: all interpretations are (1) correct and (2) justified.

Why is this important?

Because when you emphasize that data preservation is:

Ensuring reliable communication with the future

You see that it is primarily about *understanding* and *credibility*,
not the physical persistence of objects

And that focuses our attention on preservation actions that ensure understanding and credibility, *actions that are too often neglected*.