# Will Formal Preservation Models Require Relative Identity?

## An exploration of data identity statements

**Simone Sacchi, Karen M. Wickett, Allen H. Renear** {sacchi1, wickett2,renear}@illinois.edu

**Center for Informatics Research in Science and Scholarship — Graduate School of Library and Information Science**

**University  of Illinois at Urbana-Champaign**

*In this poster we present the preliminary output of a study on data identity. We analyzed an example of common identity statement and provide plausible interpretation according to established views on identity. The analysis highlights how these interpretations influence our modeling perspective in data curation and digital preservation.*

## The problem

The problem of identifying and re–identifying data put the notion of of **"same data"** at the very heart of preservation, integration and interoperability, and many other fundamental data curation activities.

However, different interpretations of **data identity statements** suggest different formalizations in formal languages and different assessing procedures, influencing how we model digitally–encoded data for curation and preservation.

Here we analyze a fairly common example of identity statement and provide three possible interpretations according to three different views of identity. Each one presents advantages and disadvantages for data modeling.

## An example of problematic identity statement

{ *a and b are the* ***same*** *data*
*but* ***different*** *XML documents* } (**A**)

The general form of such statements is:

*x and y are the same F but different Gs* (**B**)

Statements of this sort relativize identity (sameness) to particular categories, in this case, data or XML Document, and imply that x and y are identical vis–a–vis one category (here, data), but different vis–a–vis another (here, XML Document).

It is easy to see that (B), in our particular case, may be understood as the conjunction of two clauses:

*x is the same data as y* (**C**)

*x is not the same XML Document as y* (**D**)

As an example consider the dataset "Federal Data Center Consolidation Initiative Data Center Closings 2010-2013" available at https://explore.data.gov/d/d5wm-4c37. Anyone can "Download a copy of this dataset in a static format". The available formats include CSV, RDF/XML, RSS, XLS, and XML. Each of this is presumably an encoding of the "same data", while RDF/XML an XML are "different XML Documents".

## 1 – *Classical Identity* interpretation

The classical view asserts the principle known as **Leibniz's Law** (LL):

{ *if x and y are* ***identical****, then every property x has y also has* }

On the classical view this principle is a fundamental feature of our concept of identity and an axiom in most formal logics that include identity. The classical view of identity will formalize (C) snd (D) as follows:

$$\exists(x)\exists(y)[data(x)\ \&\ data(y)\ \&\ x=y] \qquad \textbf{(1a)}$$

$$\exists(x)\exists(y)[(XMLDocument(x)\ \&\ XMLDocument(y)\ \&\ \neg(x=y)] \qquad \textbf{(1b)}$$

### Observations and possible implications

- ✗ It does not seem to respond to the common sense of (A) as it is impossible for one thing to be both data and an XML Document.
- ✓ It complies to the commonsense notion of identity and standard logic.
- ✓ It complies to the standard paradigm of levels of representation, suggesting the need of a FRBR–like set of related abstract entities.
- ✓ If data is the actual target of preservation [3], we need to characterize it in terms that are independent of any specific file format.

## 3 – *Equivalence Class* interpretation

A third view of identity statements attempts to avoid the problems facing any analysis of identity by maintaining that, despite appearances, (A) is not really an identity statement, but rather an **equivalence statement**.

{ *x and y may by* ***different*** *but* ***equivalent*** *with respect to specific equivalence relations.* }

On this account "data" and "XML Document" define *equivalence relations*. This view formalizes the conjunction of (C) and (D) as follows:

$$\exists(x)\exists(y)((x \equiv_{data} y)\ \&\ \neg(x \equiv_{XMLDocument} y)) \qquad \textbf{(3)}$$

### Observations and possible implications

- ✓ It reflects the recently discussed notion of scientific equivalence [4].
- ✓ The connectives `$\equiv_{data}$‘ and `$\equiv_{XMLDocument}$‘ can be better understood as *predicates*, therefore no extension to logic is required.
- ✗ No ontological account of entities for data modeling is provided.

**Acknowledgments**

The research has been conducted at the Center for Informatics Research in Science and Scholarship and funded by the National Science Foundation as part of the Data Conservancy, an NSF funded project (OCI/ITR-DataNet 0830976).

## 2 – *Relative Identity* interpretation

The relative identity view was developed to accommodate the apparent semantics of these commonplace statements. According to this view x and y are *identical* only with respect to a general term (such as *data* or *XML Document*) that provides the **criterion of identity** [1].

{ *x and y can be* ***identical*** *with respect to some general count noun F, but* ***different*** *with respect to some other general count noun G* }

This view formalizes the conjunction of (C) and (D) as follows:

$$\exists(x)\exists(y)((x=_{data} y)\ \&\ \neg(x=_{file} y)) \qquad \textbf{(2)}$$

### Observations and possible implications

- ✓ It accommodates the apparent semantics of commonplace identity statements like (A).
- ✗ It requires a new and very peculiar logical construct.
- ✗ It violates plausible ontological and logical assumptions [2].
- ✗ If we comply to relative identity we have also to abandon established paradigms such that of levels of representation that has proven to be a compelling modeling device to represent "what's really going on" with preservation.

## Conclusion

We have drawn attention to a certain class of very important identity statements commonly made about scientific data in digital form and provide three different possible interpretations of such statements. Each of these interpretations suggest specific implications for data modeling in digital preservation.

Although all are plausible approaches, the classical view of identity seems superior in terms of modeling practices. The application of the classical view suggests the need for a system of distinct FRBR–like entities to correctly represent digitally–encoded data for preservation.

## Reference

[1] P. Geach. Mental acts; their content and their objects. 1957.
[2] J. Perry. The same f. The Philosophical Review, 79(2):181–200, 1970.
[3] S. Sacchi, K. Wickett, A. Renear, and D. Dubin. A framework for applying the concept of significant properties to datasets. Proceedings of the American Society for Information Science and Technology, 48(1):1–10, 2011.
[4] C. Tilmes, Y. Yesha, and M. Halem. Distinguishing provenance equivalence of earth science data. Procedia Computer Science, 4(0):548–557, 2011.