Will Formal Preservation Models Require Relative Identity?

An exploration of data identity statements

Simone Sacchi, Karen M. Wickett, Allen H. Renear
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana–Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211 USA
{sacchi1,wickett2,renear}@illinois.edu

Keywords
Data, Identity, Scientific Equivalence, Data Curation, Digital preservation,

1. INTRODUCTION

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, it is also a profoundly challenging notion because the concept of *data* itself clearly lacks a precise and univocal definition. When science is conducted in small communicating groups, with homogeneous data these ambiguities seldom create problems and solutions can be negotiated in casual real-time conversations. However when the data is heterogeneous in encoding, content and management practices, these problems can produce costly inefficiencies and lost opportunities. We consider here the *relative identity* view which apparently provides the most natural interpretation of common identity statements about digitally–encoded data. We show how this view conflicts with the curatorial and management practice of "data" objects, in terms of their modeling, and common knowledge representation strategies.

In what follows we focus on a single class of identity statements about digitally–encoded data: "same data but in a different format". As a representative example of the use of this kind of statements consider the dataset "Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013"[1], available at Data.gov. Anyone can "Download a copy of this dataset in a static format". The available formats include CSV, RDF, RSS, XLS, and XML. Each of this is presumably an encoding of the "same data". We explore three approaches to formalization into first order logic and for each we identify distinctive tradeoffs for preservation models. Our analysis further motivates the development of a system that will provide a comprehensive treatment of data concepts. [3].

[1] https://explore.data.gov/d/d5wm-4c37

2. PROBLEMATIC IDENTITY STATEMENTS

An example of the sort of statement we are considering is

$$\begin{array}{c} a \text{ and } b \text{ are are the same data} \\ \text{but different XML documents} \end{array} \quad \text{(A)}$$

Where "a" and "b" are identifiers or names of some sort and the object(s) they refer to are described as being different XML Documents but the same data, as would be for the RDF and XML files. The general form of such statements is:

$$x \text{ and } y \text{ are the same F but different Gs} \quad \text{(B)}$$

Statements of this sort *relativize* identity (sameness) to particular categories such as, 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 the (B) may be understood as the conjunction of two clauses.

$$x \text{ is the same data as } y \quad \text{(C)}$$

$$x \text{ is not the same XML Document as } y \quad \text{(D)}$$

We now present three different approaches to understand these familiar sentence patterns.

2.1 The Classical View

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 one that lies behind much ordinary reasoning; it is in fact an axiom in most formal logics that include identity. The classical view of identity will formalize (C) as follows:

$$\exists(x)\exists(y)(data(x) \ \& \ data(y) \ \& \ x = y) \quad \text{(1a)}$$

This reads: "There exists an x and a y such that x is data and y is data and x is identical to y". On the Classical view x and y are the same "absolutely": if they are the same "data", they are the same (are identical) and so the

same with respect to any other possible characteristics. The classical view of identity will instead formalize (D) as follows:

$$\exists(x)\exists(y)(XMLDocument(x)\ \&\ XMLDocument(y)\ \&\ \neg(x=y)) \quad (1b)$$

This reads: "There exists an x and a y such that x is an XML Document and y is an XML Document and x is NOT identical to y. The function of the term "data" and "XML Document" is only to qualify the referents of x and y, not to describe the *kind* of identity asserted. Both (1a) and (1b) are ordinary expression in standard first order logic. On to this account, it follows from (1a) and (1b) that if x is data and y is an XML Document x is not the same thing as y. Yet there is "something" that *is* data and "something" that *is* an XML Document.

The classical view seems to imply that the natural analysis of our problematic identity sentences will result in a FRBR-like conceptual model with some number of closely related abstract entities — one of which is data, and another an XML Document — but no object that has all the properties that we seem to be implied in our ordinary colloquial sentences. This is the significance of our observing, above, that it is impossible for one thing to be both data and an XML Document, the conjunction of (C) and (D) is false for all values of x and y. Among the implications for data preservation is that if *data* is the actual target of preservation [3], we need to characterize it in terms that are independent, for example, of any specific file format. All approaches that rely on file–level definitions of data are fundamentally incomplete — if not flawed — and do not entirely support a correct representation of essential data transformations, like, for example, format migration.

2.2 Relative Identity View

Clearly the classical view does not respond to the sense of (A). The *relative identity* view was developed to accommodate the apparent semantics of these commonplace statements. According to the relative identity 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]. Therefore a statement like "x is identical with y" is an incomplete expression, for which it "makes no sense to judge identity" unless we provide a criterion under which we can judge identity [1]. A consequence of this approach in that 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. The relative identity view formalizes the conjunction of (C) and (D) like this:

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

Although at first glance this view seems to match the grammar of how we often talk about digital objects, relative identity requires a new and very peculiar logical construct (an identity relationship that has three argument places: the terms identity is being applied to, and the sortal criterion). However, in a famous paper John Perry constructs a argument showing that relative identity is inconsistent with a number of very plausible assumptions[2], both at ontological and the logical levels [2]. From a modeling perspective, 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 [3].

[2]See: http://plato.stanford.edu/entries/identity-relative/

2.3 Equivalence Class View

A third view of identity statements such as (A) attempts to avoid the problems facing any analysis of identity by maintaining that, despite appearances, (A) is not really an identity statement at all, but rather an equivalence statement. According to the Equivalence Class View x and y may by different but *equivalent* with respect to specific equivalence relations. In our examples "data" and "XML Document" will both define equivalence relations: *data–equivalent* and *XMLDocument–equivalent* respectively. This view formalizes the conjunction of (C) and (D) like this:

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

We note that although (3) appears to use distinctive connectives it is plausible that they are best understood as *predicates*, therefore requiring no extensions to standard first order logic. The recently discussed notion of *scientific equivalence* [4] seems to reflect this approach. However, it leaves open the issue of a precise ontological representation of the entities involved in modeling digital objects for preservation.

3. CONCLUSION

We have drawn attention to a certain class of very important statements commonly made about scientific data in digital form. Although there are three plausible approaches to making logical sense out of these statements, the classical view of identity is decidedly superior to the others. The application of the classical view suggests the need for a system of distinct entities to correctly represent digitally–encoded data for preservation.

4. ACKNOWLEDGMENTS

The research is funded by the National Science Foundation as part of the Data Conservancy, a multi–institutional NSF funded project (OCI/ITR-DataNet 0830976).

5. REFERENCES

[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.