Encoded Descriptions at Face Value

David Dubin
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
ddubin@illinois.edu

ABSTRACT

Information resources seem to be neither fully abstract universals, nor particular concrete arrangements of matter and energy. The subjects of our metadata statements are therefore elusive. This puzzle presents few practical problems for traditional documents, but applications such as scientific data management call for more precise accounts. The problem of relating fully abstract properties to events in time is explored through a comparison of encoding strategies.

KEYWORDS

Data, ontology, properties, encoding.

INTRODUCTION

What kind of thing can have an author? Obviously, a work of authorship, but what kind of thing is that? For a long time there have been two candidates for what, precisely, information resources are: fully abstract universals (e.g., symbol sequences, graphs, trees, relations, automata) or particular concrete arrangements of matter and energy (on paper, magnetic tape, in fiber optic cable, etc). But our usual intuitions about identity, location, and provenance of resources make either account problematic. An information object cannot be identified exclusively with any one of the patterned matter/energy bundles that embodies it. But unlike abstract universals, information resources are anchored to creation and modification events in time.

Indeed, the question of whether information objects are any *type* of thing at all depends on key relationships between abstract universals and contingent facts. Consider, for example, a binary file's encoding of a digital image. Strictly speaking, the property of encoding some particular image is not necessary to that sequence of bits, but contingent on interpretations that guide the execution of computer software. Categories such as "TIFF," "JPEG," or "digital image file," would therefore seem to be roles played by fully abstract sequences, rather than types in their own right (Guarino & Welty, 2000). Similar arguments apply to bibliographic entities such as the expression of a work of authorship or the manifestation of a text (Renear & Dubin, 2007).

Certain physical objects offer a useful analogy: some particular sheet of paper, for example, will have essential properties, such as *being made of paper*, and may instantiate certain non-essential properties, like *being folded into a toy airplane*. If paper airplanes simply *are* sheets of paper folded in particular ways then (following Guarino and Welty) it may seem at first that the property of being a paper airplane is not strictly a type (such as "person") but a role (like "student" or "customer"). But from another point of view one can understand paper airplanes as a category of physical object having both shape and material as essential properties. So the property of *deltry shape*, or *dart shape* can be essential to the toy airplane, even if it's not essential to the sheet of paper.

TEMPORAL, BUT NOT SPATIO-TEMPORAL

There are phenomena which, unlike paper airplanes, have no particular physical locus for their properties, and these fit comfortably in neither the category of concrete particulars nor the category of abstract universals. Documents are one example, along with musical composition (Levinson, 1990), promises (Searle, 1999), and games of chess (Smith, 2008). On the one hand, they don't exist necessarily the way that the number twelve exists in every logically possible world. One can always imagine a possible world in which the musical work was never composed, the promise never pledged, or the game of chess never played. On the other hand, works of authorship, art, etc. are not bounded in space the way that particular physical objects are. Rather, they seem anchored in some crucial way to particular events in time (acts of composition, borrowing, etc.).

Just as documents can be identified with abstract strings and trees, so the other resources can be identified with one or more categories of abstract objects. One can believe, for example, that musical works simply *are* abstract sound patterns or that chess games are nothing more or less than sequences of abstract moves consistent with the rules of the game. On those understandings, works of authorship and art are not truly created by people, but merely discovered, and this is one of several consequence of of the views that some may consider counterintuitive. Abstract objects have no intrinsic non-essential properties So if a textual document is simply an abstract string of characters, it will have the property of a particular length. But it will not, except accidentally, have:

1. the property of being a document

ASIST 2010, October 22-27, 2010, Pittsburgh, PA, USA.

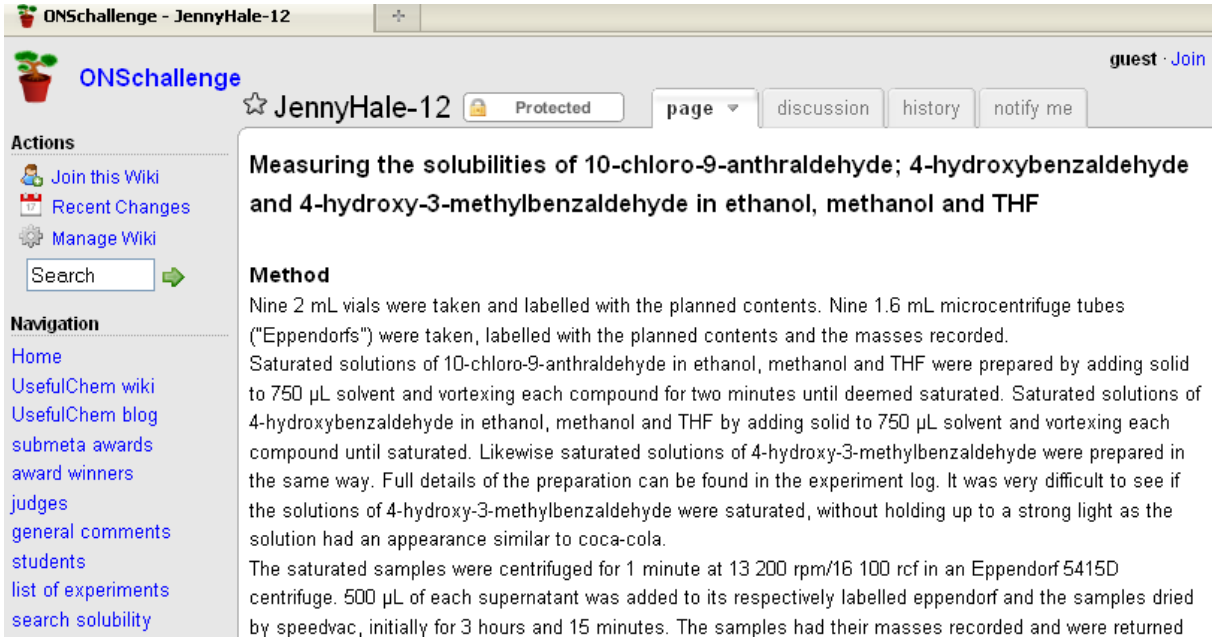


Figure 1. Excerpt from Hale online notebook

2. the property being in a particular language (such as English)
3. a document genre (such as being a business letter or novel)
4. the property of belonging to an era in time, such as the 17th century

And so on. Similarly, if an XML document really is an abstract tree of nodes, then it can have the property of being acyclic intrinsically, but not the property of being valid against a schema or conforming to well-formedness requirements. If a JPEG file is only an abstract stream of bits, then it won't have the property of being a JPEG, encoding a particular image, and so on.

If documents and other information resources are abstract objects like strings, trees, and graphs, then most of the assertions we make about them in encoded descriptions are, strictly speaking, true of nothing. That's not to say that they are false assertions, but that there is no *thing* which has the attributed property essentially. But what, if any, practical problem does that present for information modeling and digital representation? Ascriptions of extrinsic properties such as weight, assigned identification numbers, and home addresses have always been part of information systems. It's not immediately clear how the extrinsic character of these ascriptions is cause for concern.

Indeed, for nearly all the familiar database and document indexing applications, extrinsic property ascription present no serious problem at all. Even if having been authored is not the kind of property that a bit string, character string, or ordered hierarchy could have, acting as if it were the case does nothing to complicate the retrieval of a bibliographic record—quite the opposite! But we've begun to see a need for software support of more demanding forms of information management that call for more precise accounts. We take up one such application in the next section.

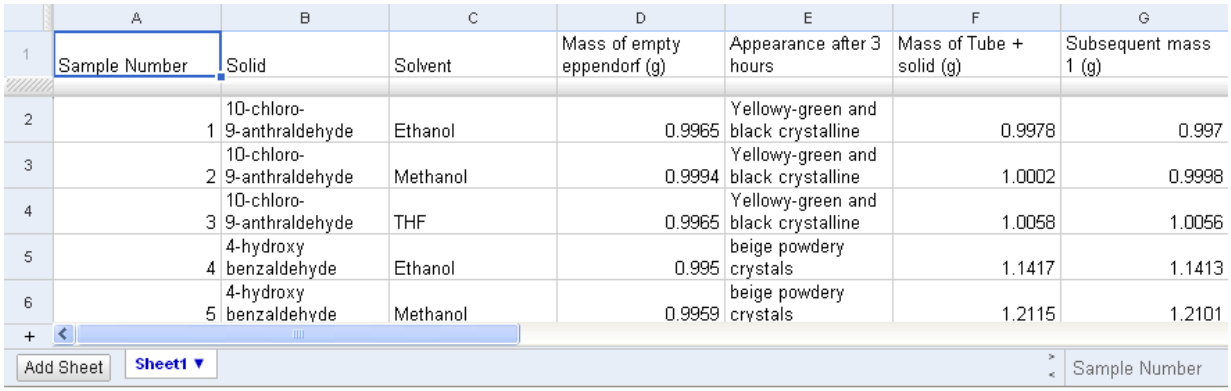
A RESEARCH DATA SET EXAMPLE

In the remainder of this paper we relate the issues examined to one particular encoding problem: the translation of a small scientific data set from a spreadsheet format into RDF. The encoding exercise is part of an ongoing Data Conservancy project funded by the National Science Foundation's Office of Cyberinfrastructure[1]. The project's broader goals are to improve support for collecting and sharing data produced by individual scientists and research groups. This particular formalization exercise is not intended to produce encoding guidelines or tools for information exchange[2], but only to explore basic concepts of what scientific data is.

Specifically, we seek to bridge the levels of data content (numbers, strings), data structure (e.g., cases, variables) and research transactions (experimentation, documentation, etc.).

The data selected for this exercise were published on the Worldwide Web as part of the Open Notebook Science Challenge[3]. They record solubility measurements for three compounds in each of three different solvents, and (like other lab notebook entries) include observation event details at a level not available for most of the data sets used in our project. These data are therefore well suited for exploring concepts at a level presumed to be a basis in those data sets that do not record transactions directly. Figure 2 shows part of an

[1] http://cirss.lis.illinois.edu/SciCom/DataConservancy.html
[2] Compare, for example, the OBOE ontology (Madin et al., 2007).
[3] http://onschallenge.wikispaces.com/JennyHale-12



	A	B	C	D	E	F	G
1	Sample Number	Solid	Solvent	Mass of empty eppendorf (g)	Appearance after 3 hours	Mass of Tube + solid (g)	Subsequent mass 1 (g)
2	1	10-chloro-9-anthraldehyde	Ethanol	0.9965	Yellowy-green and black crystalline	0.9978	0.997
3	2	10-chloro-9-anthraldehyde	Methanol	0.9994	Yellowy-green and black crystalline	1.0002	0.9998
4	3	10-chloro-9-anthraldehyde	THF	0.9965	Yellowy-green and black crystalline	1.0058	1.0056
5	4	4-hydroxy benzaldehyde	Ethanol	0.995	beige powdery crystals	1.1417	1.1413
6	5	4-hydroxy benzaldehyde	Methanol	0.9959	beige powdery crystals	1.2115	1.2101

Figure 2. example from Hale 2009 spreadsheet data

online spreadsheet linked from the log, containing the data we propose to encode.

We begin with a naïve, straightforward RDF translation, where each row of the table (i.e., each experimental sample) is deemed an object, and table columns are interpreted as properties. The results of translating one of the rows is shown below.

```
<rdf:Description
rdf:about="ns0:JHale-12#sample3"
ns1:Appearance_after_3_hours="Yellowy-
   green and black crystalline"
ns1:Concentration_(M)="0.075615937512984"
ns1:Mass_of_Tube_+_solid_(g)="1.0058"
ns1:Mass_of_empty_eppendorf_(g)="0.9965"
ns1:Mass_of_solid_in_500_uL_(mg)="9.1"
ns1:RMM_of_solid="240.69"
ns1:Sensible_concentration_(M)="0.076"
ns1:Solid="10-chloro-9-anthraldehyde"
ns1:Solvent="THF"
ns1:Subsequent_mass_1_(g)="1.0056"
ns1:Subsequent_mass_2_(g)="missing_value"
ns1:mMSolidIn500uL="0.037807968756492"/>
```

Although simple and direct, the resulting description is a digital chimera. It suggests an object that has, e.g., a relative molecular mass of 240.69, the property of being the solvent Tetrahydrofuran, and a "Mass of empty eppendorf" property. Clearly different objects and their properties have been confounded in the same description.

Given sufficient time and care, it would be possible to identify each physical object and create a separate description for its properties. The Semantic Interoperability Community of Practice (SICoP) publishes a Common Semantic Model (COSMO) ontology[4] which includes classes and properties that can support a first attempt:

```
<cosmo:Container
rdf:about="ns0:JHale-12#s3Tube"
cosmo:hasMassInGrams="0.9965"/>

<cosmo:ContainerAndContents
rdf:about="ns0:JHale-12#s3Tube+solid"
cosmo:hasMassInGrams="1.0058"/>

<cosmo:AnalyticalSample
rdf:about="ns0:JHale-12#s3solid"
ns1:Appearance_after_3_hours="Yellowy-
     green and black crystalline"
cosmo:hasMassInGrams="0.0091"/>
```

And so on. Some drawbacks to this second approach are more obvious than others. It's very tedious to identify and describe each particular object that can be discriminated by its properties. In the current example, nine solutions are prepared, but each solvent/solid combination can arguably be understood as three distinct objects (each with a different mass), rather than a single object that changes over time as the last of the solvent evaporates. Understanding containers with their contents as objects distinct from either one gives us still more work to do.

Another drawback to the second approach is more subtle: by focusing exclusively on physical objects and their properties we risk throwing away a great deal of revealing structure that is key to the data's interpretation. For example, many of the mass and molarity measurements in the spreadsheet seem redundant: one column's values are a linear function of the values in one or more other columns, suggesting that they contribute nothing new in an information theoretic sense. But these redundancies are essential to document particular calculations the experimenter has made over the data, and these in turn help explain key relationships among the variables in the study (mass, formula weight, and molar concentration, for example).

We'd like to encode our description at a level that connects fully abstract properties like masses and concentrations to the transaction events at which times they were observed or computed. Our descriptions should include *measurements* and *data values* as entities, but what exactly are those? Are they numbers? Classes of event? Statements in a language of some kind? Can we express relationships between proper-

[4] http://semanticommunity.wik.is/

ties, measurement events, measurement units, numbers, and numeral strings without creating chimeras of the kind we saw in the first encoding attempt?

Ultimately we need to accommodate scenarios where detailed transaction logs aren't available, whether due to issues of scale or to generations of post-processing and derivations between the data set and the original recorded observations. So rather than moving directly to another RDF description of the molarity data, we begin with a set of terminological axioms. The aim is to propose a high-level account of data and data content that could govern knowledge structures expressed in RDFS or OWL. The aim is not only to clarify relationships between non-repeating events and fully abstract objects, but for those expressions to support limited inferences about missing information, such as transaction events for which no asserted record appears in the knowledge base.

$$Proposition \sqsubseteq AbstractThing$$
$$SymbolStructure \sqsubseteq AbstractThing$$
$$Obs \sqsubseteq Event$$
$$Comp \sqsubseteq Event$$
$$Assertion \sqsubseteq Event$$
$$Claim \equiv Proposition \sqcap \exists substanceOf.Assertion$$
$$DataContent \equiv Claim \sqcap \exists supportedBy.(Obs \sqcup Comp)$$
$$Datum \equiv SymbolStructure \sqcap \exists expresses.DataContent$$

By this account, a datum is an abstract symbol structure that expresses data content. Data content is a claim for which someone has cited one or more observation or computation events as support. And a claim is a proposition that is the substance of some agent's assertion event. An ontology based on axioms like these might govern an encoding like the one below, where propositional content is explicitly linked to provenance events.

```
<ns0:proposition
rdf:about="ns1:propJH999">
<ns0:subject
resource="ns1:JHale-12#s3solid"/>
<ns0:predicate
resource="cosmo:hasMassInGrams"/>
<ns0:object rdf:datatype=
"xsd:float">0.0091</ns0:object>
<ns0:substanceOf
resource="ns1:JHAssert19"/>
<ns0:supportedBy
resource="ns1:JHComput441"/>
</ns0:proposition>
```

Of course, the list of axioms is not complete: the figure omits event agency properties, for example, inverse properties (e.g., *has substance*) and details of domain and range constraints. But the three definitions of claim, data content, and datum are the core of the proposal.

CONCLUSION

The definitions explored in the last example do not resolve the the metaphysical puzzles with which we began the paper. On this account, scientific data simply are a kind of abstract symbol structure, and their content are a subset of fully abstract propositions. The categories are extrinsic properties based on contingent facts. So to say that a particular symbol string "is" a datum is to ascribe a property that, strictly speaking, abstract symbol structures don't have. But we need not read the biconditional operator as expressing strict identity: a more appropriate gloss might be something like "a string of symbols has *data status*, just in case…" Such a reading would be similar to Searle's theory of social objects or social facts (Smith & Searle, 2003).

By working toward a satisfying general account of scientific data, we hope to inform the development of guidelines for recording, description, and archival practice.

ACKNOWLEDGEMENTS

This research was supported by NSF Grant OCI-0830976. Special thanks to Carole Palmer, Allen Renear, Dan Korman, Karen Wickett, Simone Sacchi, Trevor Munoz, Richard Urban, Thomas Dousa, Alejandro Gutierrez, Aaron Fleisher, and the anonymous reviewers of earlier drafts for their assistance and feedback.

References

Guarino, N., & Welty, C. A. (2000). A formal ontology of properties. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management* (pp. 97–112). London, UK: Springer-Verlag.

Levinson, J. (1990). Music, art, and metaphysics: Essays in philosophical aesthetics. In (p. 63-88). Ithaca, NY: Cornell University Press.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, *2*(3), 279 - 296.

Renear, A. H., & Dubin, D. (2007). Three of the four FRBR Group 1 entity types are roles, not types. In A. Grove (Ed.), *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology*. Medford, NJ: Information Today, Inc.

Searle, J. R. (1999). *Mind, language, and society: Philosophy in the real world*. New York: Basic Books.

Smith, B. (2008). Searle and de soto: The new ontology of the social world. In B. Smith, D. M. Mark, & I. Ehrlich (Eds.), *The mystery of capital and the construction of social reality* (pp. 35–51). Chicago/La Salle IL: Open Court.

Smith, B., & Searle, J. (2003). The construction of social reality: An exchange. *American Journal of Economics and Sociology*, *62*(1), 285–309.