fhiso

CFPS 94
(Call for Papers Submission number 94)

Recording provenance using statements about statements

Submitted by: Smith, Richard

Created: 2013-11-14

URL: Most recent version: http://fhiso.org/files/cfp/cfps94.pdf
This version: http://fhiso.org/files/cfp/cfps94_v1-0.pdf

Description: This paper extends CFPS 77's statement mechanism with a means of grouping statements into bundles for the purpose of description which it proposes to use to record provenance infomation and other metadata.

Keywords: data model, statement, provenance, sources, metadata, inference, hypotheses, RDF

Abstract

This paper extends the statement formalism of CFPS 77 by proposing a way of grouping statements into *bundles* which can be made the subject of further statements. This provides a general and extensible means of storing metadata about genealogical data, which can be used to record the provenance of genealogical data. Specific mechanisms are proposed for recording which statements are extracted directly from a source, which are inferred from other statement, and which are hypotheses advanced by a particular researcher. The proposal is influenced by and is compatible with the W3C's PROV specifications for discussing provenance, although this proposal stops considerably short of requiring full support for PROV.

1 Introduction

The formalism introduced in the paper *A unified formalism for genealogical statements* (CFPS 4) provides a way of defining a genealogical data model in terms of thing nodes, their connections, properties and so on [1]. In CFPS 77, it was shown that these concepts could largely be unified into a single notion of a *statement* [2]. Statements were defined as a tuple containing a *subject*, a *predicate*, and an *object*. Individually they are rather low-level, conveying claims such as "a person existed (or exists)" or "something (not necessarily someone) was called 'John Smith'". But in combination they allow meaningful facts to be represented, such as in the example in table 1 taken verbatim from CFPS 77.

Subject	*Predicate*	*Object*
`#I1`	*type*	*person*
`#I1`	*name*	John Smith
`#I1`	*participated in*	`#E1`
`#I2`	*type*	*person*
`#I2`	*name*	Mary Brown
`#I2`	*participated in*	`#E1`
`#E1`	*type*	*marriage event*
`#E1`	*happened at*	`#P1`
`#E1`	*date*	`1850-01-01`
`#P1`	*type*	*place*
`#P1`	*name*	Dunny-on-the-Wold

Table 1: "John Smith married Mary Brown on 1 Jan 1850 at Dunny-on-the-Wold"

There are relatively few material differences between the statement formalism of CFPS 77 and the node formalism of CFPS 4. Representing each of the types of node in CFPS 4 as statement does not imply that an application must treat them iden-

tically. Indeed, an application written in an object oriented fashion might well choose to represent statements as types of node, exactly as per CFPS 4. Similarly, an application backed by a relational database might choose to have separate tables depending on whether the object of a statement is a class, literal or reference (which are differentiated in table 1 by italic, roman and monospaced fonts); they correspond almost precisely to CFPS 4's notion of thing nodes, properties and connections.

Statements about statements can also be used for expressing belief or disbelief in a statement, though discussion of this is left to future paper. This deferral is partly because CFPS 79 introduces the useful principle of *sensible disbelief* [3], and in the process raises some objections to the formalism of CFPS 77. To discuss those issues properly would require the introduction of many of the concepts in this paper, and it seems logical to separate the general mechanism of statements about statements and its simpler uses into a separate paper from the more complex and perhaps controversial matters associated with of truth, belief and justification.

The handling of source citations was one important aspect of CFPS 4's formalism that CFPS 77 acknowledged was not included in its statement formalism. It was left to a future paper, and is taken up here. This paper proposes that source citations are handled as part of a more general mechanism for making statements about statements. That mechanism can also record details such as who extracted the data from the source, and when. Such metadata is not catered for in CFPS 4.

2 Statements about statements

One simple strategy for making statements about statements would be to attach a label to the statement being discussed, and then use that label as the subject of another statement. (Such a mechanism exists in the RDF/XML serialisation format where an `rdf:ID` attribute can be placed on the XML element representing the predicate, and that ID made the subject of another statement [4].) Unfortunately this strategy makes many of the most common applications unduly verbose. This is demonstrated by trying to specify the source from which the statements in table 1 were extracted, the researcher who extracted them, and the date on which this was done. Because there are 11 statements in the table 1, each would need a separate label, and then 33 further statements would be required to state the source, creator and creation date of the 11 original statements. This seems unsatisfactory.

This proliferation of sources is present to a degree in the CFPS 4 data model. In CFPS 4, the 11 statements in table 1 would be represented by 11 nodes (four thing nodes, four property nodes, and three connection nodes). And in CFPS 4, each of these nodes separately references the source. Whilst it is important that each and

every statement can be linked back to the originating source, this paper takes the view that sourcing each statement individually results in too fine-grained sourcing.

This paper instead proposes that related statements can be grouped into a *bundle*, which can be made the subject of further statements. A label may be useful for referencing a bundle, much as it may be useful to use labels in the reference of individuals or events. The details of this are largely a detail of the serialisation, and this paper does not specify what form the label should have: it might be a simple identifier, a URI or a UUID; or in an application, it might be implemented as a pointer. For the purpose of the examples in this paper the label will be written as an ID prefixed by a '`#`', just for the labels of individuals and events.

A statement may not be part of more than one bundle, although identical statements may appear in different bundles. For simplicity's sake, this paper assumes that all statements are part of a bundle. (An earlier draft of this paper allowed for statements that were not part of a bundle; however such statements can be thought of as part of some default bundle.) One possible implementation is to store the bundle label as a fourth member of the subject–predicate–object tuple.

The idea of grouping statements into a bundle and making those bundles subject to additional statements is one used in several recent Semantic Web standards. In particular, the World Wide Web Consortium's PROV family of specifications does this [5]; these define a data model and serialisation formats for exchanging information about the provenance of resources on the Internet, and as such is of particular relevance to genealogy. This paper does not propose that the FHISO should adopt the PROV data model: such a proposal would be premature until the requirements for genealogical provenance are established. However this paper does propose sensible alignment between the PROV and a future FHISO specification. Even if the FHISO chooses not to use PROV, it would be unfortunate if unnecessary incompatibilies between the data models made it difficult for statements about provenance in genealogy to be re-expressed in PROV, and vice versa.

3 An example

Supposing the statements in table 1 are part of a bundle, `#B1`, an application wishing to make statements about the creator, creation date and source of those statements can do it as shown in table 2. In this example, `#R1` represents the researcher. This paper proposes a class of object (a thing node in the terminology of CFPS 4) provisionally called *agent* to represent researchers. The term 'agent' is chosen rather than, say, 'researcher' because it is general enough to refer to a company or organisation, rather than a single researcher. An agent serves the role of the submitter record and `SUBM` tag in GEDCOM [6].

Subject	*Predicate*	*Object*
`#B1`	*extracted from*	`#S1`
`#S1`	*type*	*source*
`#B1`	*creator*	`#R1`
`#R1`	*type*	*agent*
`#R1`	*name*	Joe Genealogist
`#B1`	*created*	`2013-11-10`

Table 2: Describing the provenance of the statements in table 1

This paper proposes a *creator* predicate (or connection node per CFPS 4) to link a bundle to the researcher who created it, and a *created* predicate (or property node) to note when the bundle was created. The names of these predicates has been chosen so that it is aligned with the Dublin Core metadata specification [7], which has become a *de facto* standard for recording authorship. The current GEDCOM X draft has also chosen to align these metadata terms with those of Dublin Core [8].

The object model proposed in CFPS 4 make no mention of whether its property and connection nodes should be used to represent metadata about sources as well as to represent data derived from sources. This paper explicitly proposes that the same data model be used for both. Thus an agent (researcher) is represented by a thing node or a type statement, just as an individual mentioned in a source is.

The final remaining new predicate used in the example is the *extracted from* predicate. This is essentially a source citation, but the choice of predicate name is intended to make its precise semantics clear. If bundle `#B1` is extracted from source `#S1`, then this implies that every statement in `#B1` is explicitly stated in `#S1`. (A minor relaxation is made to this definition in §4.) No interpretation is involved beyond the conversion the natural language used in the source to the machine-readable statement used in the bundle. Inferences that were not explicitly stated in the source do not belong in the bundle of extracted information. They are a slightly later stage of the reasoning process, and belong in a separate bundle.

4 Bundle granularity

This paper suggests a general guideline on how statements are grouped into bundles: a bundle should correspond to what a genealogist might regard as a single piece of evidence, together with its metadata. When combining data from multiple sources, applications should be required to include either the whole of a bundle or none of it. (Additional requirements on copying them may also be required.) Thus "John Smith married Mary Brown on 1 Jan 1850 at Dunny-on-the-Wold" is a single piece of evidence, and belongs in one bundle. It may be that a partic-

ular researcher thinks some part of it (perhaps Mary's surname, for example) is untrue, but in order to provide proper context to the evidence, the whole bundle should still be included. Similarly, the metadata provides important context that allows a future researcher to assess its likely accuracy; it should be kept with the extracted data, and therefore logically belongs in the same bundle.

In the example the statements in table 1 were part of bundle `#B1`. For the reasons just given, the associated metadata belongs in the same bundle, which suggests the statements in table 2 belong in `#B1` too. However, the 'extracted from' predicate says that each statement in `#B1` comes from the given source, and table 2 includes statements about a modern-day genealogist that clearly isn't in the source. How should that apparent contradiction be handled?

This paper defines a *metadata statement* as any statement whose subject is its containing bundle. Thus if the 'extracted from', 'creator' and 'created' statements in table 2 are in bundle `#B1`, then they are metadata statements. None of the other statements in either table are metadata statements. The *extracted from* predicate is now formally defined to apply only to non-metadata statements: that is, if `#B1` is extracted from `#S1`, then this implies that that every non-metadata statement in `#B1` is explicitly stated in `#S1`. This allows the three metadata statements from table 2 to be included safely in `#B1`.

What of the other three statements in table 2? The latter two are about the researcher who extracted the information. They could usefully be placed in a separate bundle containing just information about the researcher. In practice that bundle would likely contain contact details for the researcher (subject to the researcher's privacy concerns); perhaps it might also list the research interests or biographical details of the researcher.

Bundle	*Subject*	*Predicate*	*Object*
`#B2`	`#R1`	*type*	*agent*
`#B2`	`#R1`	*name*	Joe Genealogist
`#B2`	`#R1`	*email*	`joe@example.com`
`#B2`	`#B2`	*creator*	`#R1`
`#B2`	`#B2`	*created*	`2013-09-01`

Table 3: Information about a researcher

This leaves just the 'type' statement from table 2, saying that `#S1` is a source. It probably belongs in a separate bundle along with any other description of the source, such as where it can be accessed. (Another possibility, not explored here, is that the statement might be unnecessary if it is implied by `#S1` being the object in the 'extracted from' statement.)

5 Inferences

The lexicon of terminology in CFPS 74 defines *inference* as [9]:

> the process of deriving new claims from existing claims and other information. The derivation used is the *inference rule*; the pre-existing claims and information are the *antecedents* and the new claims are the *consequents* of the inference.

Inference is thus a wholly algorithmic process. (The lexicon recognises a second sense of the word in which the researcher's judgement plays a role. This paper avoids using the word in that sense.) As a simple example, if John is the grandson of George, then it can be inferred that George had a child, and that John is the son of that child. It can also be inferred that John is male. (This is only an example, and assumes that there exists a direct way of expressing the 'is grandson of' relationship. The FHISO may choose not to define such a relationship.)

Inferences are handled by CFPS 4 by introducing a new type of node called a inference node [1]. The inference node contains a description of a specific application of an inference rule. The consequents of that rule are a series of nodes, each of which links to the inference node instead of a source node. The inference node in turn links back to the antecedent nodes, of which there are potentially several.

This paper proposes something similar. The consequents of the inference are placed together in a bundle. Instead of using the 'extracted from' predicate to link a bundle to its source, this paper proposes a new *inferred from* predicate to link the consequent bundle directly to the antecedent bundle (and multiple such predicates can be used if the antecedents appear across multiple bundles). A new *inferred by* predicate is also used to link the consequents to the inference rule or rules that were used. The collection of 'inferred from' and 'inferred by' statements are this paper's equivalent to the inference node of CFPS 4, and are stored as metadata statements in the consequent bundle.

Like CFPS 4, this paper makes not attempt to define how inference rules should be stored. In the example in table 4, the label `#F1` is intended to reference the rule that a grandson implies a son of a child, and `#F2` refers to the rule that sons are male. This could be implemented by some general-purpose rules engine, perhaps one that supports a rules interchange language such as RIF [10]. But equally they might be two of a handful of rules hard-coded into the application. If an application supports the relevant rules, then it can verify that the inference does indeed follow from the specified antecedents; if not, it must take this on trust.

Bundle	*Subject*	*Predicate*	*Object*
#B3	#I3	*type*	*person*
#B3	#I3	*name*	John Smith
#B3	#I4	*type*	*person*
#B3	#I4	*name*	George Smith
#B3	#I3	*grandson of*	#I4
#B4	#I5	*type*	*person*
#B4	#I3	*son of*	#I5
#B4	#I5	*child of*	#I4
#B4	#I3	*sex*	male
#B4	#B4	*inferred from*	#B3
#B4	#B4	*inferred by*	#F1
#B4	#B4	*inferred by*	#F2

Table 4: The antecedent and consequent bundles of an inference

6 Hypotheses

This paper defines a *hypothesis* as a bundle of statements that is neither a direct extract from a source, nor an entirely algorithmic inference. It is essentially the same as the belief node of CFPS 4, however this paper prefers not to use the term 'belief' as that appears to imply that some particular researcher does believe it although CFPS 4 does not say this is the case. In this paper a hypothesis does not necessarily imply belief. A genealogist may wish to record two possible (and mutually-exclusive) explanations for some evidence without expressing belief in either one. Further discussion of belief is left to a future paper.

It is anticipated that one of the more common uses of hypothesis will be to express *same as* statements. These were introduced in CFPS 77 as a way of representing CFPS 4's match node in a statement. They say that the subject and object refer to the same entity. For example, a 'same as' statement could express the fact that the John Smith in table 4 is the same person as the John Smith of table 1. This paper introduces a complementary *different from* predicate for expressing the opposite hypothesis: that the two refer to different people.

Grouping several 'same as' and 'different from' statements into a bundle more accurately reflects the genealogical process. Sometimes a genealogist will find just two references to a person and then conclude they are the same (or different) people. But more frequently there will be many references to people of the same name, and the researcher will consider them together and come to a conclusion on which refer to the same individual, and which are separate people.

This paper defines a fairly loose *derived from* predicate to link a bundle to those

other bundles that informed its creation. It is intentionally loosely defined as it is intended for use when no more specific predicate is suitable. If the FHISO data model allows one predicate to be defined as subtype of another, then the 'extracted from' and 'inferred from' predicates would be subtypes of 'derived from'. They are collectively referred to a *derivations*. Following the chain of derivations from bundle to bundle eventually gets to the relevant sources.

Bundle	*Subject*	*Predicate*	*Object*
#B5	#I1	*same as*	#I3
#B5	#B5	*derived from*	#B1
#B5	#B5	*derived from*	#B3
#B5	#B5	*creator*	#R1
#B5	#B5	*created*	2013-11-14

Table 5: A simple hypothesis: the John Smiths in tables 1 & 4 are the same person

7 Concluding remarks

This mechanism for grouping statements into bundles and making them the subject of statements has been used in several other recent Semantic Web technologies. They are all based in RDF which they extend by storing not only the three parts of the RDF statement, but also the URL of the RDF document that contains them. The name given to that URL varies between specifications: 'context' (in the original N-Quads specification [11]), 'subgraph' (in TriG [12]) and 'bundle' (in PROV [5]); but notwithstanding the different nomenclature, they all refer to the same underlying data model. It seems quite possible that the next version of the base RDF specification will include this. To a significant degree, GEDCOM X has also adopted the same data model [13].

In the proposal developed here, statements form the basic atoms with which data and metadata are expressed. They are somewhat lower level than the nodes of CFPS 4, and are about the simplest pieces of knowledge that have independent meaning. By contrast the bundles of this proposal are at times higher level than the nodes of CFPS 4, but serve a similar role to CFPS 4's nodes. It is bundles that are sourced, and bundles that must be preserved intact when copied by applications.

The proposals in this paper and in CFPS 77 are not driving towards the adoption of RDF by the FHISO. At the moment it remains unclear whether the requirements for a genealogical data model can be accommodated in RDF. Even if RDF's adoption is desirable, it would be premature to propose it; instead a policy of sensible harmonisation with the RDF technologies is recommended, much as in the current GEDCOM X draft. For one thing, this makes it easier to leverage existing RDF data.

References

[1] Luther Tychonievich, 2013, *Modeling Research, not Conclusions* (CFPS 4), http://fhiso.org/files/cfp/cfps4.pdf

[2] Richard Smith, 2013, *A unified formalism for genealogical statements* (CFPS 77), http://fhiso.org/files/cfp/cfps77.pdf

[3] Luther Tychonievich, 2013, *The Principle of Sensible Disbelief* (CFPS 79), http://fhiso.org/files/cfp/cfps79.pdf

[4] World Wide Web Consortium, 2012, *RDF Primer*, http://www.w3.org/TR/rdf-primer/

[5] World Wide Web Consortium, 2013, *PROV Model Primer*, http://www.w3.org/TR/prov-primer/

[6] Church of Jesus Christ of Latter-day Saints, 1996, *The GEDCOM Standard (Release 5.5)*, https://devnet.familysearch.org/docs/gedcom/gedcom55.pdf

[7] Dublin Core Metadata Initiative, 2012, *DCMI Metadata Terms*, http://dublincore.org/documents/dcmi-tersm/

[8] Intellectual Reserve Inc., 2013, *The GEDCOM X Standard Header Set*, http://gedcomx.org/headers/v1

[9] Luther Tychonievich, 2013, *Family History Lexicon* (CFPS 74), http://fhiso.org/files/cfp/cfps4.pdf

[10] World Wide Web Consortium, 2013, *RIF Primer (Second Edition)*, http://www.w3.org/TR/rif-primer/

[11] Richard Cyganiak, Andreas Harth & Aidan Hogan, 2008, *N-Quads: Extending N-Triples with Context*, http://sw.deri.org/2008/07/n-quads/

[12] World Wide Web Consortium, 2013, *RDF 1.1 TriG*, http://www.w3.org/TR/trig/

[13] Intellectual Reserve Inc., 2013, *The GEDCOM X Conceptual Model*, http://gedcomx.org/conceptual-model/v1