Beyond the graph: Telling the story with PROV and Controlled English

Darren P. Richardson
Web and Internet Science,
University of Southampton,
Southampton, UK
dpr1g09@ecs.soton.ac.uk

Luc Moreau
Web and Internet Science,
University of Southampton,
Southampton, UK

David Mott
Emerging Technology Services,
IBM United Kingdom Ltd.,
Hursley Park, Winchester, UK

***Abstract*—Provenance is the information that represents the lifetime of a piece of data or an object, including how it is affected and changed by other objects, agents, or processes. PROV is a W3C standardised model of provenance designed for use in environments such as the Web. Previous work in the ITA has addressed mapping the logical structures of the PROV data model to Controlled English, allowing us to factor provenance into reasoning and rationale. We extend this existing work, tackling a challenge that arises when converting PROV from RDF to CE: the PROV data model is graph-based, whereas textual documents, including CE documents, are linear in structure. We describe an approach to serialising the provenance graph, that can be used to create not just CE, but additionally CE Gist, and natural language texts, with the aim of increasing the accessibility of the provenance data to human users.**

I. Introduction

Provenance is the information that represents the lifetime of a piece of data or an object, including how it is affected and changed by other objects, agents, or processes. This information is of great importance when choosing what significance to give data, particularly in environments where trust and information uncertainty are factors that must be considered. Due to the Web's decentralised, uncontrolled nature, the Linked Data community has sought an approach to representing provenance in such a way as to allow people to publish data alongside its provenance whilst maintaining interoperability. PROV[1] is a World Wide Web Consortium (W3C) standardised model of provenance designed to satisfy this need, with its own notation and an accompanying OWL ontology, allowing provenance to be expressed in RDF. Previous work in the ITA [1] has addressed mapping the logical structures of the PROV data model to Controlled English, which combined with Controlled English's rationale capability, offers a powerful tool for understanding how a state of affairs came to exist. Expanding upon this existing work, in this paper we describe a technique for transforming a PROV graph expressed in RDF into a textual document.

II. Using a controlled natural language (CNL)

Figure 1[2] shows an example of how complicated a provenance graph can become, even describing a relatively simple task conducted over a short timescale. When confronted with a graph like this, it is difficult to know where to begin, even with the convention that time flows from top to bottom and from left to right. On the other hand, humans are very familiar with written narratives, which are far more suited to our capabilities and limitations, with a flow clearly mapped out for us, and allowing us to focus on understanding what we read rather than trying to figure out what to focus on first.

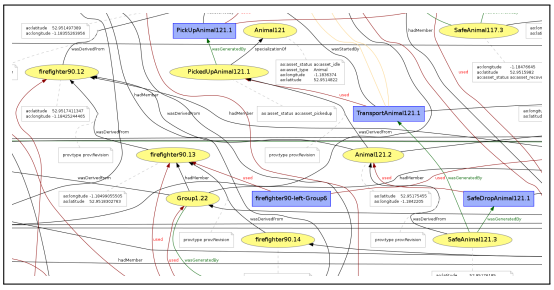

Fig. 1. This extract from a larger provenance graph shows that they can be highly-interconnected, and confusing to the uninitiated.

This, however, only justifies the use of a textual format, and not the use of a controlled natural language. Whilst there is likely to be little difference in comprehension between natural language and CNL, the main benefit one gains from the use of a CNL is that the transformation can be made bidirectional without the use of complex natural language processing techniques, by eliminating the potential for ambiguity. This means that textual interfaces expressing provenance in a CNL can be used to allow users to annotate, edit and create provenance data, in a way that an unconstrained natural language interface could not.

III. Serialising the graph

A more obvious approach to tackling the conversion from RDF to CE might involve grouping the triples by subject URIs, and then building a CE sentence to express each group. This approach has the advantage that it is fast and easy to implement, but it does not extend well to some of the more complicated features of CE, such as CE Gist [2] or extended CE. For example, it will be useful for us to create a number of PROV specific 'shortcuts' using the CE extensibility syntax that will allow for a more natural feeling expression of provenance. Furthermore, this textual conversion will be integrated as a service of the ProvStore [3], where it will, for

[1]Specification available: http://www.w3.org/TR/prov-overview/
[2]Derived from: https://provenance.ecs.soton.ac.uk/store/documents/10148/

applications that do not require bidirectional transformations, be more user-friendly to use unconstrained natural language constructs, and we were therefore motivated to explore another approach.

We have decided to use a template based method, where templates can be used to generate sentences from which a document can be derived. Each template comprises three functions:

a) Bindings function: A function that takes the graph and returns a list of sets of bindings, e.g.: `[{"thing":"ex:007", "type":"prov:Agent"}, ...]`. Each set of bindings corresponds to a possible expression of a part of the graph in textual form — a possible *sentence*. Because of the many possible combinations of sentences that could be used to express a graph, many more sets of bindings are generated than are actually necessary to do this.

b) Coverage function: A function that returns the set of triples in the graph that can be inferred from, or are expressed by a set of bindings, though these triples may not be explicitly represented in the sentence, e.g.: `[(ex:007, rdf:type, prov:Agent), ...]`. For example, if a sentence refers to some thing as a ***collection***, it is possible to infer that the thing is also an ***entity***[3]. Consequently the triple stating that the thing is a ***prov:Entity*** would be included in the coverage set as well as that stating ***prov:Collection***.

c) String function: A function that takes a set of bindings and returns the sentence as a string. This is done by simple template string substitution, and consequently it imposes no restriction in terms of syntax or grammar. This allows the same system to be used to convert a provenance graph from RDF to CE, CE gist, or natural language texts.

In order to convert a provenance graph using these templates, we first generate all the possible sets of bindings for each of the templates, along with the coverage sets for those bindings. Each coverage set is a subset of the triples in the provenance graph as a whole, and therefore — as we are trying to find a set of sentences (that is to say, a document) that expresses the whole provenance graph — this is a variant of the set-cover optimisation problem. To find the minimal solution to the set-cover problem is an NP-hard task, and consequently we use a greedy algorithm instead, which is likely to find a sub-optimal solution, but does so in a reasonable timescale:

- There is a set of all triples that remain to be covered.
- There is a set of remaining candidate sentences — initially the set of sentences generated by all the templates — where each sentence has a set of triples that it would cover. Conversely, for each triple remaining to be covered, there exists a set of candidate sentences that cover it.
- If there are any remaining triples for which that set contains only one candidate sentence, that sentence must be made a part of the document. If there are no triples for which this is the case, then the sentence is chosen that would cover the most triples.
- When a sentence is chosen, for each triple in its set of covered triples, that triple is removed from the set of remaining triples. Additionally, if it is not an ***rdf:type*** triple, all sentences that would also cover that triple are removed from the set of candidate sentences.
- Sentences are chosen until there are no triples remaining to be covered.

Using this technique, we can guarantee universal coverage of the provenance graph with two CE templates. The first covers the classes of PROV things, i.e.: ***ex:007 rdf:type prov:Agent*** becomes ***there is an agent named ex:007.***

The second covers PROV relationships between two things, i.e.: ***ex:wiretapping prov:wasAssociatedWith ex:007*** becomes ***the activity ex:wiretapping was associated with the agent ex:007.***

It should be noted that the latter case actually covers more than just one triple, because it also covers the two triples that express that ***ex:007*** and ***ex:wiretap*** are instances of the classes ***prov:Agent*** and ***prov:Activity*** respectively, and consequently whilst the sentences ***there is an agent named ex:007.*** and ***there is an activity named ex:wiretapping.*** would be generated, they would not be included in the document in this instance.

In addition to these two templates, others have been created that, for example, make use of conjunctions, and that are therefore able to have larger coverage sets associated with them. With the design of suitable CE extensions and the creation of templates to map onto these extensions, we will be able to express large numbers of triples with relatively compact sentences that feel natural to the user, but still support the bidirectional transformation permitted through the use of a controlled natural language.

IV. Conclusion

In this paper, we have presented a novel, versatile, template-based approach to converting an RDF provenance graph into a textual document, in both controlled and unconstrained natural languages. We expect this approach to allow us to create bidirectional CNL interfaces to help users understand, annotate, and create provenance data. This, in turn, will allow for the creation of systems that depend on rich provenance information in order to support better situational awareness and decision-making.

Acknowledgment

Research was sponsored by US Army Research Laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defence, or the UK Government. The US and UK Governments are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] J. Ibbotson, D. Braines, D. Mott, S. Arunkumar, and M. Srivatsa, "Documenting provenance with a controlled natural language," in *2012 Annual Conference of the International Technology Alliance (ACITA'12)*, Sep. 2012.

[2] A. Preece, D. Braines, D. Pizzocaro, and C. Parizas, "Human-machine conversations to support multi-agency missions," *Mobile Computing and Communications Review*, vol. 18, no. 1, Jan. 2014.

[3] T. D. Huynh and L. Moreau, "ProvStore: a public provenance repository," in *5th International Provenance and Annotation Workshop (IPAW'14)*, Cologne, Germany, Jun. 2014.

[3] ***prov:Entity*** is a superclass of ***prov:Collection***