What's in a name? Exploiting URIs to enrich provenance explanations in plain English

Darren P. Richardson
Web and Internet Science,
University of Southampton,
Southampton, 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 allows decision-makers to evaluate the importance of pieces of data. PROV is the standardised model of provenance for use on the web, particularly suited for situations where data is generated by systems under distributed control, such as in coalition operations. If human decision-makers are to make effective use of provenance data, they need to understand it, and this work establishes techniques for explaining PROV graphs to human users in natural English.**

In this paper, we demonstrate the potential role of exploiting the linguistic information that is informally encoded in the URIs used to denote provenance data resources to generate these more natural English explanations of provenance. We show how this additional linguistic information allows us to generate richer, more readable explanation texts, thus enabling better decision-making and increasing the value of preexisting provenance data.

I. Introduction

In data-driven, network-centric operations, provenance is vital for helping analysts and decision-makers to evaluate the importance of a piece of data. This is particularly true in environments where data may have been generated by unfamiliar sensors, and may have passed through many hands before reaching a decision-maker. PROV [1] is the World Wide Web Consortium (W3C) standardised model of provenance, primarily intended for use on the open web, and it is consequently well-suited for situations where data is generated by systems under distributed control. As such, PROV has great potential for use in coalition or civil-military environments.

In order for decision-makers to be able to make effective use of provenance data, they first need to understand it. Previous work in the ITA ([2], [3]) has established techniques for explaining PROV graphs to human users in Controlled English, and our current work aims to extend this approach to generate explanations in less constrained, more natural English. In the remainder of this paper, we describe how we utilise the linguistic information informally encoded in the URIs used to denote the PROV resources to improve the readability of the explanations we generate.

II. Exploiting linguistic information in URIs

Existing template-based PROV explanation systems typically fall into one of two categories: either they generalise to all PROV, at the cost of readability [3], or they produce very good, clear, concise English, but only to describe very narrowly defined PROV constructs [4]. We have developed an approach that is able, in many cases, to bridge the gap and achieve both of these goals: plain English and generalisability.

Our approach is to use a number of templates to describe all the possible relations contained within the PROV data model. However, instead of the templates simply substituting variables into strings, as has until now been typical of such systems, we use templates to generate sentence plans, which can then be passed into a sophisticated off-the-shelf realisation engine [5], which in turn produces the natural language strings, orthographically correct with valid conjugations and number agreement.

However, in order to be able to perform this more sophisticated form of template-based generation, the system requires a greater amount of linguistic information to be available to it at runtime. Additionally, because we want our approach to explaining provenance to generalise to all PROV, without requiring system implementers to have to support any additional technologies, we are limited to using only the linguistic information available in the PROV data itself. As potentially useful features such as `prov:label` are optional, the only mandatory feature is the URI denoting each resource, and consequently we have focussed on them. URIs themselves are, per the RDF model, merely identifiers, and formally carry no information. Nevertheless, system developers often create meaningful URI schemas for a number of different reasons, such as increasing the maintainability of the system, or making it easier for human users to interact with the system on the occasions where it is necessary for them to see a URI.

Extracting this linguistic information from URIs for the purposes of natural language generation (NLG) is not completely novel in itself [6]. However, there are some interesting differences that should be noted in its application to provenance. Where the results of the study by Mellish and Sun [6] suggest that instances of the class `prov:Activity` are likely to be denoted by URIs containing a noun, our investigation into the way PROV generating systems are minting URIs shows that the results of this study may not hold entirely true for PROV. Our results show that it is not uncommon for instances of `prov:Activity` to be given a verb, or in cases where the developer felt the need to use a noun, to specifically use a gerund — the nominalised form of a verb. Both of these give us more linguistic information to use when generating explanations.

This is important, because if the only linguistic information we could garner from the source data were the nouns, then it

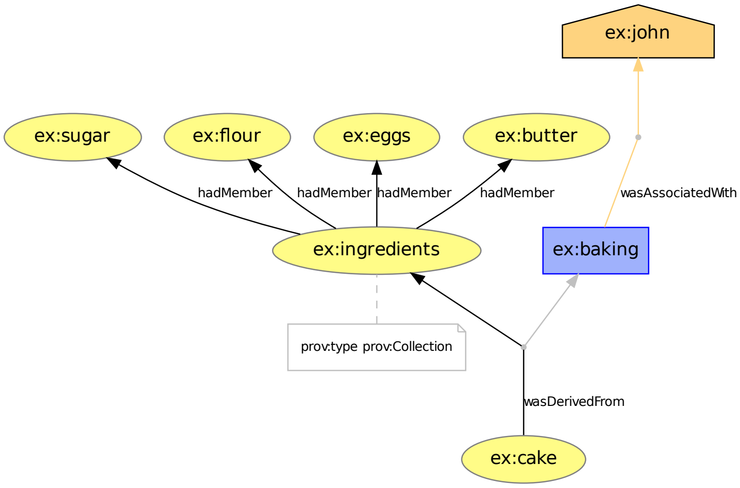

Fig. 1. A simple example PROV graph, consisting of 18 triples, showing the derivation of an entity, from a collection of entities, by an activity associated with an agent. In plain English: *John baked a cake from some ingredients. The ingredients were eggs, flour, butter, and sugar.*

would be necessary to use templates that were constrained to using the verbs built into the PROV data model, such as *to derive*, *to attribute*, and *to invalidate*. On the other hand, by being able to extract verbs from URIs we are able to generate a much wider variety of sentences, whilst still using only a small number of templates. It is these features that allow us to create templates that are collectively able to generalise to being able to explain all instances of PROV, without compromising on the readability of the sentences generated.

III. Example generation

To illustrate that this approach leads to more readable sentences than previous systems, we present the provenance graph shown in Figure 1. Without taking advantage of the linguistic information in the URIs, we are only able to generate the following explanation, similar to those described in [3]: **ex:ingredients was a collection that had ex:sugar, ex:flour, ex:eggs, and ex:butter as members. ex:cake was derived from the collection ex:ingredients, by the activity ex:baking. ex:baking was associated with the agent ex:john.**

In contrast, by using our approach to extracting linguistic information from URI, we can use an off-the-shelf realisation engine to generate the much more readable explanation: **John baked a cake from some ingredients. The ingredients were eggs, flour, butter, and sugar.**

There are a number of heuristics at work here. Of particular interest, we firstly note that the gerund *baking* can be converted back into the verb *to bake*, which the realisation engine can then conjugate. Secondly, the verb which adequately describes the relationship between a collection — in this case, *ingredients* — and its members appears to be related to the number of the head noun in its URI. If the head noun is plural, then the members of the collection tend *to be* instances of the singular form of the collection URI, whereas when the head noun in the collection URI is singular, the verb *to contain* seems to be more often appropriate. We have been able to incorporate a number of such heuristics into our template set, increasing the overall readability.

IV. Conclusion

In this paper we have shown how exploiting the linguistic information informally encoded in URIs can lead to more readable explanations of provenance. Moreover, our approach is domain-agnostic, and does not require implementers to support any particular technologies other than PROV. In future work, we intend to conduct a user-evaluation to establish what impact such provenance explanation technology might have on the way analysts and decision-makers process data.

Acknowledgement

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] L. Moreau and P. Missier, "PROV-DM: The PROV Data Model." World Wide Web Consortium Recommendation, Apr. 2013.

[2] 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)*, (Southampton, UK), Sept. 2012.

[3] D. P. Richardson, L. Moreau, and D. Mott, "Beyond the graph: Telling the story with PROV and Controlled English," in *2014 Annual Fall Meeting of the International Technology Alliance*, (Cardiff, UK), Sept. 2014.

[4] H. Packer and L. Moreau, "Sentence Templating for Explaining Provenance," in *Provenance Week 2014, 5th International Provenance and Annotation Workshop (IPAW'14)*, (Cologne, DE), June 2014.

[5] A. Gatt and E. Reiter, "SimpleNLG: A realisation engine for practical applications," in *12th European Workshop on Natural Language Generation*, (Athens, Greece), pp. 90–93, Mar. 2009.

[6] C. Mellish and X. Sun, "The semantic web as a Linguistic resource: Opportunities for natural language generation," *Knowledge-Based Systems*, vol. 19, pp. 298–303, Sept. 2006.