Crowdsourcing data citation graphs using provenance[⋆]

Laura Dragan[1], Markus Luczak-Roesch[1],
Elena Simperl[1], Bettina Berendt[2], and Luc Moreau[1]

[1] University of Southampton, UK
l.dragan@soton.ac.uk, m.luczak-rosch@soton.ac.uk,
e.simperl@soton.ac.uk, l.moreau@ecs.soton.ac.uk
[2] KU Leuven, BE
bettina.berendt@cs.kuleuven.be

Abstract. In this paper we describe a tool designed to support crowdsourcing a-posteori provenance information about the datasets used in research publications. It generates PROV data both to capture the data citation graphs—via an extension to the PROV Data Model, and the crowdsourcing process—via prov:bundles.

1 Introduction

The reproducibility of research results, and in this context data citation, gains more and more importance, because of the uptake of data-intense experiments in many disciplines[3]. Openly available datasets like DBpedia for example become highly requested primary sources for experimental research but are not archived in any system, which would allow for referencing a particular version in a standardised way. This leads to references being made to the key papers of the DBpedia publishers instead of the actual version of the DBpedia dataset that was used in a particular study.

Since 2011 the series of USEWOD workshops[4] pushed forward the scientific discourse around Web Usage Mining in times of the rapidly growing Web of Data. At every edition of the workshop, a new USEWOD research dataset was released to be used for experiments which would be described in papers submitted in a special track called *USEWOD data challenge*[1]. The datasets contain server access logs from various well-known Linked Data sources like DBpedia, LinkedOpenGeoData, and BioPortal amongst others. What has started as a single data challenge at a scientific workshop has evolved into a reference dataset for query logs of Linked Data endpoints. Research based on this dataset are published across the boundaries of the workshop, so that we are now facing a

[⋆] This work was supported by the EPSRC Theory and Practice of Social Machines Programme Grant, EP/J017728/1.

[3] The current paper length does not permit us to discuss related research in the area of data citation standards. Such an overview will be part of an extended version.

[4] http://usewod.org/workshops.html

Table 1. Supported qualified derivations between resources.

	Publication	Dataset
Publication	Citation	Mention, Analysis, Comparison, Description, Evaluation
Dataset	(symmetric)	Extension, Inclusion, PartialOverlap, Transformation, Specialisation

complex infrastructure, consisting of a central root research dataset (and its versions) as well as the resulting publications, which may rely on various processed subsets of this primary source. Thus, the USEWOD research dataset is facing the same challenge as DBpedia but at a much smaller scale, which makes it an appropriate candidate for studying our approach to create data citation graphs.

In this paper we describe a crowdsourcing tool designed *to support gathering of a-posteori provenance information about the datasets used in research papers*. The tool was used and evaluated during this year's edition of the USEWOD workshop. The participants took part in a crowdsourcing experiment, available online at http://prov.usewod.org/. The experiment aimed to gather information about how the USEWOD datasets released over the years have been used in publications, and how they have been potentially transformed to derive new datasets. Auxiliary results include a citation network for the publications added in the system, collaboration network between authors, and more. Our tool is designed to be generic and supports the crowdsourcing of provenance information about any types of entities. The insights gained from the on-site experiment will be used to further develop the tool based on user feedback. The provenance data is generated according to recipes described in [2].

2 Data representation

The system works with two main types: *publications* and *datasets*. An auxiliary type of resource are people - which can be the authors of papers and the participants to the crowdsourcing experiment. We use `schema.org` to describe the datasets, research papers and people involved.

In the case of the USEWOD workshops, the organizers do not restrict the way the datasets are used for research, thus a varied spectrum of usage patterns emerge. We started by enumerating these patterns as verbs: a publication *mentions*, *describes*, *analyses*, *evaluates*, *compares* datasets; a publication *cites* another; a dataset *is a transformation of*, *includes*, *extends*, *overlaps*, or *specialises* another. To formally describe these patterns – which datasets were used by which publications, and *how* – we extend the W3C `PROV` Data Model[5] with subclasses of `prov:Derivation` corresponding to the verbs we identified. We

[5] http://www.w3.org/TR/prov-overview/

Fig. 1. Analysis paper about a dataset.

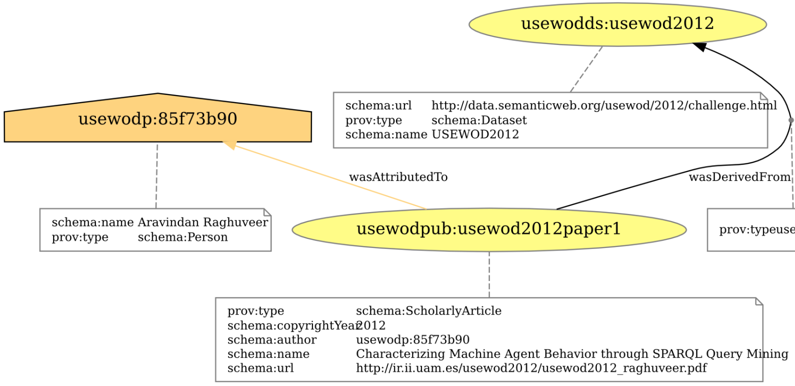
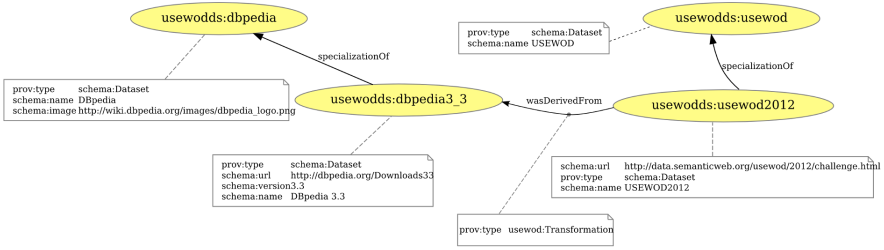

usewodds:usewod2012
schema:url http://data.semanticweb.org/usewod/2012/challenge.html
prov:type schema:Dataset
schema:name USEWOD2012
usewodp:85f73b90
wasAttributedTo
wasDerivedFrom
schema:name Aravindan Raghuveer
prov:type schema:Person
usewodpub:usewod2012paper1
prov:typeusewod:Analysis
prov:type schema:ScholarlyArticle
schema:copyrightYear2012
schema:author usewodp:85f73b90
schema:name Characterizing Machine Agent Behavior through SPARQL Query Mining
schema:url http://ir.ii.uam.es/usewod2012/usewod2012_raghuveer.pdf

Fig. 2. Transformation of datasets.



created a simple model in which the relations between resources are represented as qualified derivations, as shown in Table 1. The model allows us to elicit more precise information than just which dataset is used in which publication. The list of subtypes can be extended as required with new values, if other patterns are recognised.

An alternative representation would have employed a combination of `prov:Generation` and `prov:Usage`. We could intuitively say "the dataset D was used in the creation of publication P" – and represent this as a qualified generation which includes a "write-up" `prov:Activity`, which in turn `prov:used` the dataset. While the representation is not incorrect, it involves the creation of additional activities which are weakly specified. Thus, we chose to extend `prov:Derivation` for our model, which has the additional advantage of being more general and usable for any types of entities.

Figure 1 shows that the winning paper from the USEWOD 2012 data challenge was an analysis of the corresponding year's USEWOD dataset. Figure 2 shows that the dataset USEWOD2012 contained a transformation of DBpedia3.3, which is a specialization of the generic DBpedia dataset.

We use `prov:bundles` to also capture the provenance of the crowdsourced information. The participants to the experiment choose nicknames under which they contribute, and the data they create is attributed to them.

3 Analysis and validation of generated provenance

As with most crowdsourcing systems, we use measures to determine the accuracy of the generated information. By allowing participants to create connections between publications and datasets independently from each other, we can observe duplication of relations as reinforcement.

Participants see the list of publications and datasets created by others, but they are not shown the existing links between them. This limits the creation of duplicate resources (e.g. same paper added by different users), while supporting the duplication of links between resources, which in turn allows a measure of validity – the more people claim the existence of a relationship, the more likely it is that it exists.

We plan to extend the crowdsourcing experiment with a voting feature which would allow participants to validate or invalidate (via up and down votes) the links created by other users. In this scenario, the links between resources are created only once, and are visible to all participants. We would like to study how the results of the voting system compare to those yielded by the original duplication-as-reinforcement system.

Using a reputation system and domain knowledge we can give different weights to information provided by participants. For example if a participant is also author, their contribution about a publication they wrote might be more accurate. Domain knowledge can also be used to automate partial validation of the data – in our case, a publication cannot logically use a dataset which was released after the date of publication. If we bring in additional data sources, like information about the submission deadline of the event where the paper was published, we can further restrict the domain. However, such validation mechanisms are not yet included in our tool, but will be applied to the resulting data after the first crowdsourcing experiment.

When relationships are validated, we include them in a "curated" citation graph which will be exposed as Linked Data, with references to additional bibliographic data. This graph will be available for query and further analysis.

4 Conclusion

In this paper we describe a tool designed to support crowdsourcing data citation graphs. We defined an extension to the `PROV` Data Model to allow the capture of more detailed information about the use of datasets in research publications. The extension consists of a set of sub-types of `prov:Derivation`. We also capture provenance information about the crowdsourcing process with `prov:bundle`s.

References

1. Berendt, B., Hollink, L., Hollink, V., Luczak-Rösch, M., Möller, K., Vallet, D.: Usage analysis and the web of data. SIGIR Forum 45(1), 63–69 (May 2011)
2. Moreau, L., Groth, P.: Provenance: An Introduction to PROV. Morgan and Claypool (September 2013)