SciDataCon 2016

Conference paper

Data Reuse Fitness Assessment Using Provenance

Summary

Assessing the fitness of data for reuse may require knowledge of how that data was produced. If knowledge of how data is produced can be represented using a standard data model, automated assessments of data fitness may take place, based on aspects of its production. In addition to knowledge of data's production, knowledge of how it has or hasn't been used can also be used to assess its fitness for further reuse.

Since 2014 we have had an international data model for representing data's production, namely the W3C's provenance data model, PROV-DM. It can also be used to represent how data has been used which is known as 'forward provenance'.

Here we present several types of provenance queries one may pose in order to assess data's fitness for reuse. These include discovering the methods used in data production; determining the reputation of ancestor data; determining the reputation of agents (human or machine) involved in data production; and assessing the social acceptance of data via its reported use which we believe to be the best form of social endorsement for data's utility.

Background

The provenance data model, PROV-DM

The World Wide Web Consortium's provenance data model, PROV-DM, is designed to record provenance: "information about entities, activities, and people involved in producing a piece of data or thing" (Moreau and Missier 2013) in a standardized way. It is the result of international working groups' developments of previous provenance models, e.g. the Open Provenance Model (Moreau *et al.* 2011). **Figure 1A** shows PROV-DM's basic components. Using PROV-DM's OWL[1] ontology formulation, PROV-O, we represent basic data production as per **Figure 1B**. Such a representation can also be used for 'forward provenance', indicating data's use, with the data in question considered as input data to usage process.

Associating provenance with real things

PROV-O information can be used for more than just provenance representation: the identifiers for class objects in OWL can be URIs which, like URLs, can resolve to representations of the objects on the Internet. Using Linked Data[2] principles, these URIs can give access to different representations of objects such as their metadata or primary data in differing formats. In the case of non-information objects portrayed in provenance, such as a person acting as an *Agent* to whom the production of a dataset was attributed, the identifier could be a community-agreed identifier such as an ORCID[3] or a ResearcherID[4].

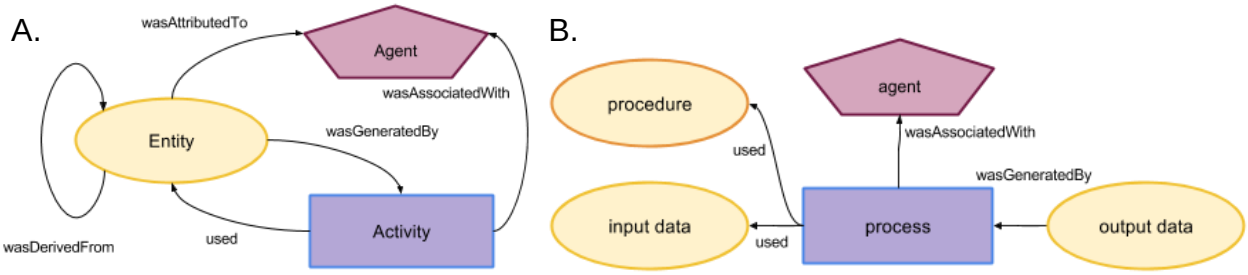

Figure 1 A: Basic classes of object and basic relationships in the PROV-DM data model (Moreau and Missier 2013), drawn according to an OWL ontology, PROV-O (Lebo *et al.* 2013). **B**: Basic PROV-O data processing model with the processing procedure – code or manual instructions – shown as an input dataset to the process.

Querying provenance information

Where provenance is stored according to PROV-O, we can query it using the SPARQL[5] query language. An example query finds the identifiers for all the ancestor datasets of a target dataset ("http://example.com/dataset/1"):

```
SELECT ?id WHERE {?id <http://www.w3.org/ns/prov#wasDerivedFrom>+
<http://example.com/dataset/1> .}
```

Other queries may find Agents associated with data' ancestors, procedures used or, if 'forward provenance' information is recorded, datasets derived from a target dataset.

Assessing fitness via provenance

Assembling more information than just provenance

Provenance, as recorded in conformance with PROV-DM, only associates *Entities*, *Activities* and *Agents* and doesn't, in and of itself, require metadata for items that may be useful in fitness assessments. Nevertheless, data recorded in the RDF[6] format within an OWL ontology is easily able to be joined to other metadata also recorded in RDF in accordance with other ontologies for metadata. There are plenty of demonstrations of the use of ontologies such as Dublin Core Terms (DCT) or DCAT[7], being used with PROV-O, even within the PROV-O specification document (Lebo *et al.* 2013). Combined PROV-O & DCAT or DCT or other ontology information can be stored in RDF databases and accessed together using SPARQL queries and used for fitness assessment.

Fitness assessment methodologies

The range of possible methodologies for data reuse fitness assessment using provenance and metadata information is open ended, limited only by what information can be collected and stored. We discuss and demonstrate three methods, some of which are entirely automatable and some of which require manual work. Fitness for reuse could be based on:

1. **Properties of ancestor data**, e.g.: all ancestor data licensed in certain ways
2. **Properties of agents involved in data production**, e.g.: only using data whose ancestors' creators have a certain reputation
3. **The methods used in data production**, e.g.: data made with viewable code

For the three methodologies and examples listed above, we provide example provenance data and method implementations in the code repository associated with this paper: http://promsns.org/repo/prov-data-fitness. The method implementations are realized in Python code that processes example RDF graphs in files *eg1.ttl* and *eg2.ttl*. Graphical representations of the RDF data, drawn as OWL diagrams, are given in *eg1.png* and *eg2.png*. In addition to provenance data recorded according to PROV-O, the example data includes data metadata according to the DCAT and Dublin Core Terms ontologies.

For 1, the function `assess_license_3_or_4()` determines ancestor data for target data by examining the target data's provenance. It then checks ancestor data's licenses and returns `True`, i.e. fit for reuse according to this test, if all the licenses are either Creative Commons v3.0 or v4.0. For the *eg1.ttl* data file it passes and for *eg2.ttl* it fails as in the latter case, an ancestor dataset, *Dataset 3*, has a Creative Common 2.5 license.

For 2, the function `assess_min_drep_points()` finds target data's ancestor data and then finds the agents (people) associated with those ancestors. It then determines fitness based on owners' "dataOwnerRepPoints", an invented measure of a data owner's reputation. It passes using data from file *eg1.ttl* when the minimum dataOwnerRepPoints for an ancestor is set to 3 and fails when set to 6 as the ancestors have either 5 or 10 points.

For 3, the function `assess_find_method_code()` inspects data's provenance to locate methods, in this example computer code, used in its production. An assessment of fitness based on that method would be a further manula step however the provenance has been used to find the method in a systematic way. The function passes using example data in *eg1.ttl* and fails for *eg2.ttl* which doesn't including links to code.

Fitness assessment methodologies using 'forward provenance'

For assessments based on 'forward provenance' we could measure fitness for reuse by:

4. Data esteem measured by reported reuse
5. Properties of derived data
6. Properties of agents involved in derived data production
7. The methods used in derived data production

The code repository for this paper at http://promsns.org/repo/prov-data-fitness provides example data that could be used to test methodologies 4, 5, 6 and 7 in *eg3-forward.ttl* and its graphical representation, *eg3-forward.png*,. We provide explanations of these methodologies, not working code, as they strongly echo patterns seen in methodologies 1, 2 & 3.

For 4 it contains an example of data reuse recorded according to PROV-O (Datasets 2, 3, 5 and the Journal Article A all derive from Dataset 1) with which a reuse count could be generated. For 5 it contains derived data (Dataset 3 and Journal Paper A) that have attributes that may be used for a fitness for reuse assessment of the target data. In this case, Dataset 3 is housed in a certain repository and Journal Paper A is presented in a journal with an impact factor of 4.3. For 6 the file *eg3-forward.ttl* contains derived data, Dataset 2, that is associated with an Agent, Agent J, that has a certain reputation indicated by the imaginary property of "dataOwnerRepPoints" equal to 10. For 7 the computer code used to generated the derived data, Dataset 5, can be discovered querying the target data forward provenance.

We believe an indication of any form of data reuse in formalised provenance information is far better for assessing its utility than simplistic social media voting or tagging given both the detailed information given in such reports and also the effort a reporter must have gone to in order to generate such information which indicates high esteem for the original data.

Conclusion

We have demonstrated potential methodologies for assessing data reuse fitness based on standardized provenance and 'forward provenance. We have used the PROV-DM for provenance representation and other well-known ontologies for metadata representation meaning queries used for assessment can used the standardized SPARQL language. Together this means fitness assessment queries, or queries that gather data for manual fitness assessment are could be used widely and in a predictable manner.

Future work could see data profiles and standard fitness assessment queries published for community use. An OWL ontology could be delivered to assist with modelling common metadata used to assess fitness, much as the repository associated with this paper references an imagined 'eg' ontology.

Competing Interests

The authors declare that they have no competing interests.

Notes

1 https://www.w3.org/standards/techs/owl - the W3C's Web Ontology Language.

2 https://www.w3.org/standards/semanticweb/data - W3C's Linked Data specification.

3 http://orcid.org – persistent digital identifier for researchers

4 http://www.researcherid.com – "a solution to the author ambiguity problem"

5 https://www.w3.org/TR/rdf-sparql-query/

6 Resource Description Framework: https://www.w3.org/RDF/

7 http://dublincore.org/documents/dcmi-terms/ & https://www.w3.org/TR/vocab-dcat/

References

Lebo, T., **Sahoo, S.**, & **McGuinness, D.** (eds.) 2013 PROV-O: The PROV Ontology. W3C recommendation. Online http://www.w3.org/TR/prov-o/ [Last accessed 18 May 2016].

Moreau, L., and **Missier, P.** (eds.) 2013 PROV-DM: The PROV Data Model. *W3C recommendation*. Online https://www.w3.org/TR/prov-dm/ [Last accessed 18 May 2016].

Moreau, L., **Clifford, B.**, **Freire, J.**, **Futrelle, J.**, **Gil, Y.**, **Groth, P.** & **den Bussche, V.** 2011. The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, 27(6), 743–756. http://doi.org/doi:10.1016/j.future.2010.07.005