Towards Policy-aware Queries over Linked Data[⋆]

Sebastian Speiser

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
firstname.lastname@kit.edu

1 Introduction

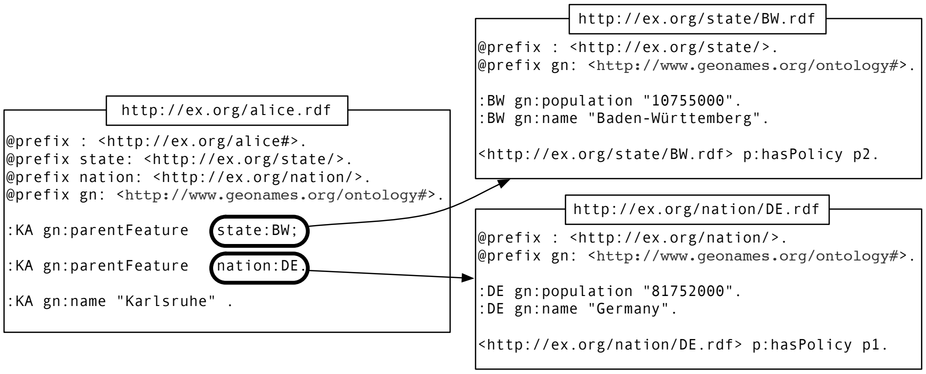
The Linked Data principles for publishing data on the Web enable the distributed evaluation of queries, where data sources are discovered during runtime. Data sources can have associated licenses that restrict allowed usages. Besides restrictions on the access to the data sources, the usage terms can also restrict the usage terms which can be assigned to derived data artefacts, e.g. query results depending on a data source. We propose to formalise such usage restrictions, in order to automate compliance checks, when query results are used or republished for a specific purpose. This work demonstrates a practical application of a suitable policy formalism published at this conference [1].

2 Scenario

Alice wants to sell a real estate dossier about the city of Karlsruhe, for which she needs an upper bound on the city's population. She has access to a description of Karlsruhe that links to the state Baden-Württemberg and the nation Germany, both geographical entities of which Karlsruhe is a part. Retrieving the linked entities results in descriptions of the state and the nation including their population numbers. The Linked Data graph is visualised in Figure 1. The corresponding data sources, however, are published under different licenses: the data source about Germany allows arbitrary usages, whereas the information about the state and any derived artefacts can only be used for non-commercial purposes.

In order to get the upper bound, Alice evaluates the following SPARQL query for the population numbers of entities of which Karlsruhe is a part:
`SELECT ?f ?p WHERE { ex:KA gn:parentFeature ?f . ?f gn:population ?p }`.
The query processor can first retrieve `ex:KA`, which gives two bindings for `?f` (`http://ex.org/state/BW` and `http://ex.org/nation/DE`). By retrieving the URIs of the bindings, she can retrieve both the population numbers of the state and the nation. She can however not use the tighter bound given by the state, as she specified that she wants to use the data for a commercial purpose (selling).

[⋆] This work was supported by the EU FP7 grant 257641 (PlanetData) and by the German Federal Ministry of Education and Research (BMBF) under the SAW-Project NanOn.



```
http://ex.org/alice.rdf
@prefix : <http://ex.org/alice#>.
@prefix state: <http://ex.org/state/>.
@prefix nation: <http://ex.org/nation/>.
@prefix gn: <http://www.geonames.org/ontology#>.

:KA gn:parentFeature  state:BW;

:KA gn:parentFeature  nation:DE.

:KA gn:name "Karlsruhe" .
```

```
http://ex.org/state/BW.rdf
@prefix : <http://ex.org/state/>.
@prefix gn: <http://www.geonames.org/ontology#>.

:BW gn:population "10755000".
:BW gn:name "Baden-Württemberg".

<http://ex.org/state/BW.rdf> p:hasPolicy p2.
```

```
http://ex.org/nation/DE.rdf
@prefix : <http://ex.org/nation/>.
@prefix gn: <http://www.geonames.org/ontology#>.

:DE gn:population "81752000".
:DE gn:name "Germany".

<http://ex.org/nation/DE.rdf> p:hasPolicy p1.
```

Fig. 1. Linked Data Graph of Scenario

3 Usage Policies

Usage policies are formal descriptions of the usages allowed to be performed on protected data artefacts. In contrast to traditional access control, usage policies still apply after initial access was granted. Specifically, usage policies can (i) impose obligations and (ii) restrict the policies of artefacts generated in dependency of the original protected artefacts.

Obligations are actions that a user is required to fulfill when executing an action allowed by the policy, e.g., a Creative Commons Attribution license allows the usage of an artefact but obliges the user to give credit to the original creator.

Restrictions of dependent artefacts' policies are in the easiest case just inherited from the original artefact. An example is a classified data set, which is downloaded by an admissible agent. The downloaded copy is still classified and thus inherits the same usage restrictions.

Depending on the actions performed on an artefact, the policy of a derived artefact can however differ in arbitrary ways. As an example, consider the Google Maps API. Accessing data from the API requires a user to be registered with a developer account. In contrast, the application using the API clearly does not require its users to have a Google developer account. Indeed, according to the terms and conditions of the Maps API, the application must be made available to users without requiring a payment[1].

Another popular example are share-alike clauses as found for example in some Creative Commons licenses. Share-alike licenses require that derived data artefacts are published under the same terms as the original artefact.

In both examples policies of dependent artefacts are restricted, by requiring that the policies must allow at least certain actions (e.g., access without payment) or at most certain actions (e.g., reuse only in combination with attribution). Such content-based policy restrictions stand in contrast to name-based restrictions, which specify a hard coded list of policies which are allowed for

[1] Sec 9.1 of the Google Maps API Terms http://code.google.com/apis/maps/terms.html

dependent artefacts. In our scenario of Linked Data query processing, data is integrated from different sources with potentially different usage policies. With content-based restrictions, interoperability is increased as policies with different names can be inferred to be compatible if they have the same intention. Lessig proposes content-based restrictions for Creative Commons licenses [2] [2].

Usage policies with content-based policy restrictions can be expressed using the formalism proposed in [1]. The formalism models policies as formulae in decidable fragments of first-order logic (FOL) with one free variable. Possible bindings for the free variable are the usages compliant to a policy. Furthermore, a $\mathsf{containedIn}$ predicate with the following extension is introduced: a policy p_1 is contained in a policy p_2, if every usage compliant to p_1 is also compliant to p_2.

In the following we present the policies of the data sources containing population information as introduced in Section 2. We choose a description logic as base formalism for the policies. The policy of the artefact containing the nation's population simply allows every usage or derivation: $p_1 : \ \mathsf{Usage} \sqcup \mathsf{Derivation}$.
The policy of the data source about the state's population requires a non-commercial purpose for usages and the same terms for derived artefacts:

$$\begin{aligned}p_2 :&(\mathsf{Usage} \sqcap \exists \mathsf{hasPurpose}.\mathsf{NonCommercial}) \sqcup \\ &(\mathsf{Derivation} \sqcap \forall \mathsf{wasGeneratedBy}^{-1}.\forall \mathsf{hasPolicy}.\exists \mathsf{containedIn}.\{p_2\}).\end{aligned}$$

For a more complete and formal treatment of the policy formalism, its semantics, and the used vocabulary, we refer the reader to [1].

4 Policy-aware Linked Data Query Processing

Linked Data refers to four principles for publishing data on the Web, which essentially allow, given an entity, to discover more and more related information by following links and dereferencing HTTP URIs[3]. The availability of large amounts of Linked Data has spurred the development of query processors, which dynamically discover relevant data sources during query evaluation by retrieving the URIs of intermediate query results [3–5].

The documents obtained by retrieving an URI can be further described in RDF, for example specifying the license of the document. The result of a query is not only a set of bindings for the queried variables, but for each binding also a list of triples that where used to produce the binding, which in turn can be mapped to the containing documents and the corresponding usage restrictions.

Each result of a SPARQL query can be regarded as a data artefact that was derived from the documents containing the triples that produced the binding. If policies of the data sources are given in the proposed policy language, then several policy tasks can be performed automatically:

[2] Unfortunately, the real licenses represented by their legal code still lack behind and are formulated with name-based restrictions.

[3] The principles can be found in http://www.w3.org/DesignIssues/LinkedData

1. Check compliance of derivations in general. Some data sources may have restrictions on the derived artefacts that are contradicting each other, thus the binding cannot be used at all.
2. Check compliance of derivation for a desired target policy. The user can specify a policy for the derived artefact and check if it is compatible to the policies of the used data sources.
3. Determine the target policy. Given the policy restrictions of the used data sources, a possible policy for the produced artefact can be generated automatically (cf. [6]).

The situation where Alice derives population upper bounds for Karlsruhe from the data source about the nation (derivation d_1) and from the data source about the state (derivation d_2) is formally described in the following:

$$\mathsf{Derivation}(d_1).\mathsf{wasGeneratedBy}(a_1, d_1).\mathsf{hasPolicy}(a_1, p_{alice}).\mathsf{used}(d_1, \mathsf{NationData}).$$

$$\mathsf{Derivation}(d_2).\mathsf{wasGeneratedBy}(a_2, d_2).\mathsf{hasPolicy}(a_2, p_{alice}).\mathsf{used}(d_2, \mathsf{StateData}).$$

The policy p_{alice} models the usages that Alice desires to perform on the derived data: $p_{alice} : \mathsf{Usage} \sqcap \exists \mathsf{hasPurpose}.\mathsf{Commercial}$.

The policy p_{alice} allows commercial usages and is thus not contained in p_2, which in turn makes the derivation d_2 non-compliant to the policy p_2 of the used data artefact StateData. Therefore, the derivation and its produced data cannot be used. Alice is stuck with the upper bound given by the nation's population.

5 Conclusions

Linked Data enables the dynamic discovery and integration of new data sources in the process of satisfying information needs. Information is however in general needed for specific usages and purposes, which not always comply with the usage restrictions imposed by data owners. We outlined how such usage restrictions can be modeled with an appropriate formalism in order to automate checking the compliance of a data usage. In future work, we plan to build a system integrating Linked Data query answering and policy compliance checking.

References

1. Krötzsch, M., Speiser, S.: ShareAlike your data: Self-referential usage policies for the Semantic Web. In: International Semantic Web Conference (ISWC). (2011)
2. Lessig, L.: CC in Review: Lawrence Lessig on Compatibility (2005) available at: http://creativecommons.org/weblog/entry/5709.
3. Hartig, O., Bizer, C., Freytag, J.C.: Executing SPARQL Queries over the Web of Linked Data. In: International Semantic Web Conference (ISWC). (2009)
4. Ladwig, G., Tran, T.: Linked Data Query Processing Strategies. In: International Semantic Web Conference (ISWC). (2010)
5. Harth, A., Hose, K., Karnstedt, M., Polleres, A., Sattler, K.U., Umbrich, J.: Data summaries for on-demand queries over linked data. In: International Conference on World Wide Web (WWW). (2010)
6. Speiser, S.: Policy of Composition $\neq$ Composition of Policies. In: IEEE Symposium on Policies for Distributed Systems and Networks (POLICY). (2011)