A dataset of RDF licenses[1]

Víctor Rodríguez-Doncel[a], Serena Villata[b], Asunción Gómez-Pérez[a]
[a]*Ontology Engineering Group, Universidad Politécnica de Madrid*
[b]*INRIA Sophia Antipolis*

Abstract. This paper describes a dataset of licenses expressed as RDF. The most important rights and conditions present in licenses for software, data and general works are expressed with the Open Digital Rights Language (ODRL) 2.0 vocabulary and extensions thereof. The dataset contains licenses identified by a dereferenceable URI, which are served with content negotiation providing a double representation for humans and machines alike. This feature enables a generalized machine-to-machine commerce if generally adopted.

Keywords. ODRL, licenses, policies, rights, intellectual property rights, RDF

1. Introduction

Computer science, software engineering and the Web technologies themselves have accomplished remarkable achievements in a very short period of time. This is mostly due to the massive reutilization of algorithms, source code and data in its broadest sense. Some of these contributions have been free willingly released as 'open', intended for its reuse, but sometimes with some limitations.

Source code published in the web is usually distributed along with a license declaring precisely what is allowed and what is not. A license may be defined in this domain as a document by which the rightsholder of a copyright protected resource waives some of his or her rights, possibly conditioned to the satisfaction of some requirements. These licenses have been written in a natural language, most of the times in plain English, and they have been addressed to other software developers to read and understand. Similarly, many photos and media resources have been published in the Web with a clear rights declaration. Further, the domain of licensed resources is extending its realm on the Web and other online assets such as raw data is being now also licensed.

Yet, machines have usually difficulties in locating and understanding the license in published goods and they can not retrieve and integrate resources automatically without a human supervision which approves or prevents the use of others' resources. Indeed, there are policy languages and rights expression languages able to express the rights information in a structured, machine-readable form. However, they have not gained much widespread. The postulate of this paper is that one of the factors that has prevented this to happen, is that there were computer languages to build the rights digital expressions, but not a collection of already built templates or pre-defined licenses. Given that the set of licenses is limited in practice to a reduced group of

[1] This work has been partially supported by a Juan de la Cierva fellowship and the EU FP7 LIDER project (FP7 – 610782).

documents used once and again (Creative Commons, Apache, MIT, etc.), considering that one of the best represention with a clear semantics is the one based on RDF, and making use of the best publishing practices as Linked Data, in this paper we present RDFLicense[2], a RDF dataset of licenses for its use with resources in the Web.

2. Related work

Diverse collections of licenses have been gathered in some websites. These licenses are typically summarized, categorized and often attributed a distinction label. For example, the TDRLegal [3] initiative gathers license summaries from the community, where permissions, prohibitions and obligations can be taken from a pre-defined list of common terms. Also, the Open Knowledge Foundation presents an extensive list[4] of licenses compliant to their definition of 'open'. Yet, there is not a downloadable dataset with well structured licenses for a machine to parse; rather a better arrangement for humans to understand.

The only relevant application capable of automatically tracking licensed resources is Apache Rat[5]. This tool is intended for auditing the distributions of source code, scanning for the text of common licenses in the headers of the source code files, providing license reports and facilitating the license addition. However, there is no more fine-grained description for a license beyond an identifier.

However the languages to digitally represent the key information in licenses have existed for at least a decade. XML-based Rights Expression Languages like MPEG-21 REL [1] or ODRL [2] included all the elements for such licenses exist, but no effort was made to systematically map existing licenses to these languages. Only Creative Commons defined and used their RDF-based language ccREL [3] to describe digitally their licenses --but only those.

3. Model and dataset description

The RDFLicense dataset contains over 100 licenses written in RDF extensively using ODRL 2.0[6]. They include licenses for data (like Open Data Commons'), software (like Apache, MIT or BSD licenses) and general works (like Creative Commons licenses). ODRL 2.0 is a language to express rights and policies, specified by the W3C ODRL Community Group [7] and based on an abstract model. This model accepts serializations as XML, JSON and RDF, the latter being based on the ODRL 2.0 Ontology[8]. Typical ODRL expressions allow declaring sentences like this: «*An action is (permitted /prohibited / obliged) to be acted by a party over an asset, provided that the constraints hold*». A policy is a collection of such rules (permissions, prohibitions or duties), whose structure is shown in Figure 1.

[2] http://datahub.io/es/dataset/rdflicense
[3] https://tldrlegal.com/
[4] http://opendefinition.org/licenses/
[5] http://creadur.apache.org/rat/
[6] http://www.w3.org/community/odrl/two/model/
[7] http://www.w3.org/community/odrl/
[8] http://www.w3.org/ns/odrl/2/

ODRL specifies a common vocabulary of general terms (such as "read", "write", etc.) which does not include all the rights or conditions that could be deemed as relevant in a license for Linked Data. For this regard, terms from other vocabularies have been used, like the Linked Data Rights 2.0 (LDR) ontology[9], which provide the vocabulary for defining rights expressions for Linked Data resources.

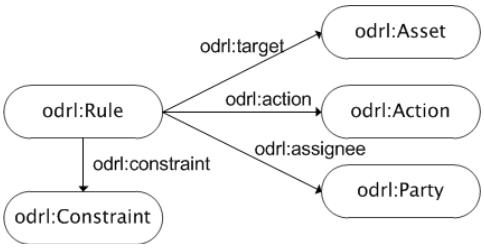

Figure 1. Main classes and object properties in the ODRL 2.0 Ontology

The following excerpt represents the UK Non-commercial Government License as RDF (in Turtle syntax) published by UK National Archives. It declares that copies, distributions and derivative works are allowed as long as the license text is attached and the work duly attributed. In any case, it is strictly forbidden to commercialize with the licensed resource. Two Creative Commons ccREL properties are used to declare the jurisdiction as long as the actual legal code reference. The license is linked to resources from other datasets (DBPedia, Lexvo).

```
<http://purl.org/NET/rdflicense/licOGL>
  a odrl:Set;
    cc:jurisdiction < http://dbpedia.org/page/United_Kingdom>;
    rdfs:label   "UK NONCOMMERCIAL GOVERNMENT LICENSE";
    dct:language <http://www.lexvo.org/page/iso639-3/eng>;
    cc:legalcode <http://www.nationalarchives.gov.uk/..ommited..> ;
       odrl:permission [
              a odrl:Permission;
              odrl:action odrl:copy, odrl:distribute, odrl:derive;
               odrl:duty [
                   a odrl:Duty;
                    odrl:action odrl:attribute, odrl:attachPolicy;
               ]
       ] ;
       odrl:prohibition [
               a odrl:Prohibition;
              odrl:action odrl:commercialize
       ] .
```

The dataset on licenses has been published as Linked Data. Each license is identified by a URI which resolves when browsed via HTTP. The served content is presented in two different forms, addressing machines or humans, and depending on whether HTML was requested (typical case of a human with a web browser) or RDF (expected to be the case for machines). Licenses can be attributed to a web resource with the standard Dublin Core license metadata element. This, along with the content negotiation, is depicted in Figure 2.

[9] http://purl.oclc.org/NET/ldr/ns#

The RDFLicense dataset, which is described in VOiD [10], is accesible via SPARQL[11] and its licenses are expected to be referenced by other resources in the Linked Data Cloud[12]: the Semantic Web is the first and most natural domain for semantic licenses to be used.

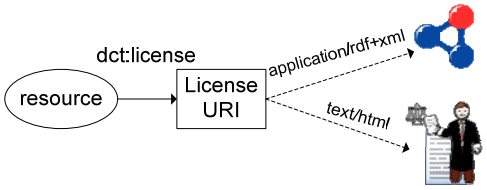

Figure 2. A resource licensed with an RDF graph and a legal document

4. Relevance of RDFLicense

As of today, RDF licenses do not substitute in any manner the legal texts. But they enable a number of applications whenever the simplification of using a reduced set of generally defined terms suffice. First, having an identifier for each license makes references unequivocal. Second, having a structured description of licenses allows the advanced search of resources by its license ("*find resources whose price is below $10*"). Third, the aggregation of differently licensed resources is made easier, as license compatibility calculations can be computed [4]. Fourth, the mapping of natural language licenses to its condensed RDF version is a rich training information for NLP algorithms working with licensing texts [5]. Fifth, access control based on RDFLicenses is possible, as it has been recently shown for the case of Linked Data[13].

The RDFLicense dataset is expected to be enlarged with non-free license templates, where the precise price can be simply specified within the URI and the served document (HTML or RDF) has the customary clauses but with a different price. These templates will allow the easy publication of non-free offers in a simple manner.

We believe that the generalized use of the RDFLicense licenses will favour the unambiguous understanding of rights and conditions, and it will create the conditions for automated markets to appear, with machines negotiating and re-utilizing resources massively beyond the open paradigms.

References

[1] ISO/IEC 21000-5:2004, Information technology — Multimedia framework (MPEG-21) — Part 5: Rights Expression Language, 2004
[2] Ianella, R. (ed.): Open Digital Rights Language v.1.1 http://www.w3.org/TR/odrl/
[3] Abelson, H., Adida, B., Linksvayer, M., & Yergler, N. (2008). ccREL: The creative commons rights expression language. Technical report, Creative Commons, 2008.
[4] G. Governatori, A. Rotolo, S. Villata, and F. Gandon. One license to compose them all - a deontic logic approach to data licensing on the web of data. In Proc. of ISWC, LNCS 8218, pp 151–166, 2013.
[5] Cabrio E., Palmero A., and Villata, S. These are your rights: a natural language processing approach to automated RDF licenses generation. In Proc. of ESWC-2014, pp. 255-269, 2014.

[10] http://oeg-dev.dia.fi.upm.es/licensius/rdflicense/void.ttl
[11] http://linkeddata4.dia.fi.upm.es:8907/sparql
[12] http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/
[13] http://conditional.linkeddata.es