

Modeling and Reasoning over Data Licenses

Oleksandra Panasiuk¹, Simon Steyskal^{2,3}, Giray Havur^{2,3}, Anna Fensel¹, and Sabrina Kirrane²

¹ STI Innsbruck, Department of Computer Science, University of Innsbruck, Austria

² Vienna University of Economics and Business, Austria

³ Siemens AG Österreich, Austria

Abstract. In this paper, we propose an extension of the Open Digital Right Language for modeling well-known licenses and propose an approach to automatically check license compatibility.

1 Introduction

Copyright is a legal right, under Intellectual Property law, that enables creators of artistic works to specify how their work is used and distributed. When it comes to information available on the Internet there is often a misconception that public information can be freely copied and downloaded, however creative works available online are also protected via copyright, irrespective of whether a license is present or not. In order to support the automatic checking of licenses, it is necessary to model licenses in a manner such that it is possible to automatically verify if it is permissible to combine and reuse different datasets or software libraries. When it comes to machine readable licenses, there have been a number of Rights Expression Languages standardisation initiatives (e.g. the Open Digital Rights Language(ODRL)⁴ and the Creative Commons Rights Expression Language (ccREL)⁵). In addition, there have been a number of works that demonstrate how RDF can be used to represent and reason over licenses [2, 4, 7]. In this paper, we describe work conducted by the Data Licenses Clearance Center (DALICC) project⁶, which focuses on extending existing vocabularies to enable modeling and reasoning over well-known license texts. Herein we make the following contributions: (i) we extend ODRL so that it can be used to model several standard license families (CC, BSD, MIT, BSD, GPL); and (ii) we propose a system to automatically check license compatibility.

2 Related Work

Rights Expression Languages (RELs) are used to explicate machine-readable rights for purposes of Digital Asset Management. Among the most prominent REL vocabularies are ccREL (which is a W3C member submission) and ODRL (a W3C recommendation from February 2018), and a derivative RightsML⁷).

⁴ <https://www.w3.org/TR/odrl-model/>

⁵ <https://www.w3.org/Submission/ccREL/>

⁶ <https://www.dalicc.net/>

⁷ <https://iptc.org/standards/rightsml/>

Besides standardization other work includes: an OWL ontology that can be used to describe the copyright domain [2], a framework for adding licensing terms to web data [7] and a license composition tool for derivative works [4].

3 License Modeling

An example output from our modeling process, which is comprised of three parts: (i) analysis of license text; (ii) defining vocabularies to express licenses; and (iii) deriving modeling and mapping mechanisms, can be seen in *Listing 1.1*.

Analysis of the License Text Representation. For our analysis we selected 14 commonly used licenses, namely: CC BY, CC BY-SA, CC BY-NC, CC BY-ND, CC BY-NC-ND, CC BY-NC-SA, APACHE, BSD-2, BSD-3, GNU GPL-2, GNU GPL-3, APGL, LGPL and MIT, which can be applied to the different assets, such as creative works, software and datasets. From the text representation we identified important concepts, requirements, and conflicts between licenses.

Defining Vocabularies. Based on research conducted on the genealogy of RELs [5] we chose ODRL as it is particularly suitable for modeling licenses in the form of *policies*. The *policy* expresses *permissions*, *prohibitions* and *duties* related to the usage of *assets* (e.g. actions *odrl:reproduce*, *odrl:distribute* can be applied to the target "Image"). To represent the main asset targets we used the Dublin Core vocabulary⁸, which covers such concepts as: software, dataset, sound, text and image. Furthermore, the ODRL vocabulary includes terms that are deprecated by terms from CC REL (e.g. *odrl:commercialize* by *cc:CommercialUse*) or are supplemented by terms from CC REL (e.g. *cc:Notice* to capture copyright information). However, given that together the ODRL and CC REL vocabularies are not able to represent all of the necessary license concepts, we constructed a DALICC vocabulary⁹ in order to fill this gap (e.g. *dalicc:perpetual* as a validity period of the license, e.g. *dalicc:worldwide* as a jurisdictional property, *dalicc:modificationNotice* as an action to state changes, see in *Listing 1.1*).

Modeling and Mapping Mechanisms. When it comes to modeling licenses, we use provenance to model information about assets (e.g. *odrl:target dct:Software*) and additional information about the license (e.g. *cc:jurisdiction dalicc:worldwide*) and ODRL rules to represent common licensing conditions divided into three categories: permissions, duties and prohibitions. An RDF representation of the APACHE 2.0 license¹⁰ is shown *Listing 1.1*. The license permits redistribution, reproduction, modification, public presentation of the asset, commercial use, charging a distribution fee, creation of a new derivative, distribution and changing the license for a derivative work, but prohibits the charge of a licensing fee. The license requires the user to post a notice of the type of license, to give attribution to the creator and to state changes.

⁸ <http://dublincore.org/documents/dcmi-type-vocabulary/>

⁹ <https://dalicc.poolparty.biz/DALICCVocabulary>

¹⁰ <http://www.apache.org/licenses/LICENSE-2.0>

Listing 1.1. Excerpt RDF serialisation of the APACHE 2.0 license

```
[...]
:APACHE_v2 a odr1:Set ;
  odr1:permission [
    a odr1:Permission ;
    odr1:target dct:Software ;
    odr1:action odr1:distribute, odr1:reproduce, odr1:modify, odr1:present,
      cc:CommercialUse, dalicc:chargeDistributionFee, odr1:derive,
      cc:DerivativeWorks, odr1:grantUse ;
    odr1:duty [
      a odr1:Duty ;
      odr1:action cc:Notice, cc:Attribution, dalicc:modificationNotice ] ] ;
  odr1:prohibition [
    a odr1:Prohibition ;
    odr1:target <http://purl.org/dc/dcmitype/Software> ;
    odr1:action dalicc:chargeLicenseFee ] ;
  dalicc:validityType dalicc:perpetual ;
  cc:jurisdiction dalicc:worldwide ;
  dct:title "Apache License , Version 2.0"@en ;
[...]
```

4 Verifying License Compatibility

The license compatibility check is performed by a reasoning engine, which uses Answer Set Programming (ASP) [1], a declarative knowledge representation and reasoning formalism that is supported by a wide range of efficient solvers. An ASP program consists of rules: $Head \leftarrow A_1, \dots, A_m, not A_{m+1}, \dots, not A_n$ where $m, n \geq 0$, $Head$ and each A_i are atoms. A rule is called a *fact* if $m = n = 0$. Sets of rules are evaluated in ASP under the stable-model semantics which allows several models, i.e. “answer sets” [1]. We use the CLINGO [3] ASP solver for our experiments, as it is one of the most efficient implementations available.

Licences should be understood as a set of rules derived from the RDF graphs of the licenses. Herein, a rule that permits or prohibits the execution of an action on certain assets does not only affect other rules that govern the execution of the same action on the same asset(s) but also those permitting or prohibiting related actions on the same asset(s). DALICC utilises a dependency graph for representing the semantic relationship between defined actions (cf., Listing 1.2). The function of this graph is to encode expert knowledge on the implicit and explicit dependencies between actions. Following the work of Steyskal and Polleres [6], the corresponding dependency graph represents hierarchical relationships (e.g., present *includes* display), implications derived from a specific action (e.g., share *implies* distribute), equalities (e.g., copy *equals* reproduce), and contradictions between specific actions (e.g., non-derivative *contradicts* derivative).

Listing 1.2. ASP encoding of an excerpt of DALICC’s dependency graph

```
[...]
action(odr1_display;odr1_present;odr1_reproduce;odr1_derive).
includedIn(odr1_display,odr1_present).
[...]
```

In order to verify license compatibility, the RDF representation of the licenses are first translated into an ASP program as follows: (i) **rule**(l, c, i, α, t), a rule in a licence l of category c (i.e. permission, prohibition or duty) is granted to an

assignee i for executing an action α on the asset t ; (ii) **action**(α), α is an action; (iii) **sameAs**(α_1, α_2), α_1 and α_2 are the same action; (iv) **includedIn**(α_1, α_2), action α_1 is included in action α_2 ; (v) **implies**(α_1, α_2), action α_1 implies action α_2 .

Our ASP program returns an answer set that consists of the predicate **conflict**($rule_1(l_1, c_1, i_1, \alpha_1, t_1), rule_2(l_2, c_2, i_2, \alpha_2, t_2)$) which means $rule_1$ is in conflict with $rule_2$ (i.e., l_1 does not comply with l_2). In ODRL, if an action α_1 is included in or equal to another action α_2 (α_1 **odrl:includedIn** | **owl:sameAs** α_2), all the rules defined for α_2 must also hold for α_1 and vice versa. Moreover, if an action α_1 implies another action α_2 (α_1 **odrl:implies** α_2), a prohibition of α_2 conflicts a permission of α_1 (but not necessarily vice versa).

An extended version of this program is – given multiple licenses as input – capable of finding all non conflicting sets of permissions, prohibitions, and duties of those licenses. These reasoning functionalities are accessed via an UI in a web service.

5 Conclusion

In this paper, we discussed how well-know licenses can be modeled using ODRL. We analyzed 14 licenses in total and extended existing vocabularies so that we can both model and check the compatibility of licenses automatically.

Acknowledgments. Funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) DALICC project <https://www.dalicc.net>.

References

1. G. Brewka, T. Eiter, and M. Truszczynski. Answer set programming at a glance. *Communications of the ACM*, 54(12), 2011.
2. R. García and R. Gil. Copyright licenses reasoning using an owl-dl ontology. *Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, 188, 2009.
3. M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. Clingo= asp+ control: Extended report, 2014.
4. G. Governatori, H.-P. Lam, A. Rotolo, S. Villata, G. A. Atemezing, and F. L. Gandon. Live: a tool for checking licenses compatibility between vocabularies and data. In *International Semantic Web Conference*, 2014.
5. T. Pellegrini, A. Schoenhofer, S. Kirrane, S. Steyskal, A. Fensel, O. Panasiuk, V. Mireles-Chavez, T. Thurner, M. Doerfler, and A. Polleres. A Genealogy and Classification of Rights Expression Languages - Preliminary Results. In *Data Protection/LegalTech - Proceedings of the 21st International Legal Informatics Symposium IRIS 2018*, pages 243–250, 2018.
6. S. Steyskal and A. Polleres. Towards formal semantics for ODRL policies. In *9th International Symposium RuleML*, pages 360–375, 2015.
7. S. Villata and F. Gandon. Licenses compatibility and composition in the web of data. In *Third International Workshop on Consuming Linked Data*, 2012.