LIVE: a Tool for Checking Licenses Compatibility between Vocabularies and Data

Guido Governatori[1]*, Ho-Pun Lam[1], Antonino Rotolo[2], Serena Villata[3],
Ghislain Atemezing[4], and Fabien Gandon[3]

[1] NICTA Queensland Research Laboratory
firstname.lastname@nicta.com.au
[2] University of Bologna
antonino.rotolo@unibo.it
[3] INRIA Sophia Antipolis
firstname.lastname@inria.fr
[4] Eurecom
auguste.atemezing@eurecom.fr

Abstract In the Web of Data, licenses specifying the terms of use and reuse are associated not only to datasets but also to vocabularies. However, even less support is provided for taking the licenses of vocabularies into account than for datasets, which says it all. In this paper, we present a framework called LIVE able to support data publishers in verifying licenses compatibility, taking into account both the licenses associated to the vocabularies and those assigned to the data built using such vocabularies.

1 Introduction

The license of a dataset in the Web of Data can be specified within the data, or outside of it, for example in a separate document linking the data. In line with the Web of Data philosophy [3], licenses for such datasets should be specified in RDF, for instance through the Dublin Core vocabulary[1]. Despite such guidelines, still a lot of effort is needed to enhance the association of licenses to data on the Web, and to process licensed material in an automated way. The scenario becomes even more complex when another essential component in the Web of Data is taken into account: the vocabularies. Our goal is to support the data provider in assigning a license to her data, and verifying its compatibility with the licenses associated to the adopted vocabularies. We answer this question by proposing an online framework called LIVE[2] (LIcenses VErification) that exploits the formal approach to licenses composition proposed in [2] to verify the compatibility of a set of heterogeneous licenses. LIVE, after retrieving the licenses associated to the vocabularies used in the dataset under analysis, supports data providers in verifying whether the license assigned to the dataset is compatible with those of the vocabularies, and returns a warning when this is not the case.

* NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

[1] http://purl.org/dc/terms/license

[2] The online tool is available at http://www.eurecom.fr/~atemezin/licenseChecker/

2 The LIVE framework

The LIVE framework is a Javascript application, combining HTML and Bootstrap. Hence, installation has no prerequisite. Since the tool is written in Javascript, the best way to monitor the execution time is with the `performance.now()` function. We use the 10 LOD datasets with the highest number of links towards other LOD datasets available at `http://lod-cloud.net/state/#links`. For each of the URLs in Datahub, we retrieve the VoID[3] file in Turtle format, and we use the *voidChecker* function[4] of the LIVE tool to retrieve the associated license, if any. The input of the LIVE framework (Figure 1) consists in the dataset (URI or VOiD) whose license has to be verified. The framework is composed by two modules. The first module takes care of retrieving the vocabularies used in the dataset, and for each vocabulary, retrieves the associate license[5] (if any) querying the LOV repository. The second module takes as input the set of licenses (i.e., the licenses of the vocabularies used in the dataset as well as the license assigned to the dataset) to verify whether they are compatible with each others. The result returned by the module is a *yes/no* answer. In case of negative answer, the data provider is invited to change the license associated to the dataset and check back again with the LIVE framework whether further inconsistencies arise.

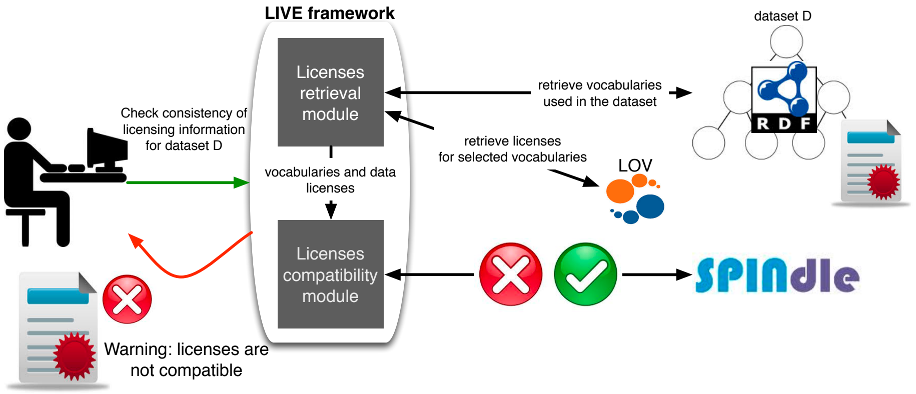
LIVE framework
Licenses retrieval module
vocabularies and data licenses
Licenses compatibility module
Check consistency of licensing information for dataset D
retrieve vocabularies used in the dataset
retrieve licenses for selected vocabularies
dataset D
R D F
LOV
SPINdle
Warning: licenses are not compatible

Figure 1. LIVE framework architecture.

Retrieving licensing information from vocabularies and datasets. Two use-cases are taken into account: a SPARQL endpoint, or a VoID file in Turtle syntax. In the first use case, the tool retrieves the named graphs present in the repository, and then the user is asked to select the URI of the graph that needs to be checked. Having that information, a SPARQL query is triggered, looking for entities declared as `owl:Ontology`,

[3] `http://www.w3.org/TR/void/`

[4] `http://www.eurecom.fr/~atemezin/licenseChecker/voidChecker.html`

[5] Note that the LIVE framework relies on the dataset of machine-readable licenses (RDF, Turtle syntax) presented in [1].

`voaf:Vocabulary` or object of the `void:vocabulary` property. The final step is to look up the LOV catalogue to check whether they declare any license. There are two options for checking the license: *(i)* a *"strict checking"* where the `FILTER` clause contains exactly the namespace of the submitted vocabulary, or *(ii)* a *"domain checking"*, where only the domain of the vocabulary is used in the `FILTER` clause. This latter option is recommended in case only one vocabulary has to be checked for the license. In the second use case, the module parses a VoID file using a N3 parser for Javascript[6], and then collects the declared vocabularies in the file, querying again LOV[7] to check their licensing information. When the URIs of the licenses associated to the vocabularies and the dataset are retrieved, the module retrieves the machine-readable description of the licenses in the dataset of licenses [1].

Licenses compatibility verification. The logic proposed in [2] and the licenses compatibility verification process has been implemented using SPINdle [4] – a defeasible logic reasoner capable of inferencing defeasible theories with hundredth of thousand rules.

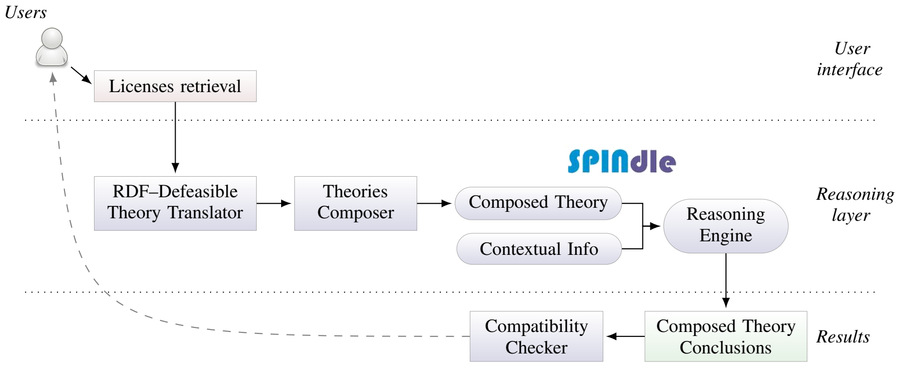

Figure 2. Licenses compatibility module.

As depicted in Figure 2, after receiving queries from users, the selected licenses (represented using RDF) will be translated into the DFL formalism supported by SPINdle using the *RDF-Defeasible Theory Translator*. That is, each RDF-triple will be translated into a defeasible rule based on the subsumption relation between the *subject* and *object* of a RDF-triples. In our case, we can use the subject and object of the RDF-triples as the antecedent and head of a defeasible rule, respectively. Besides, the translator also supports direct import from the Web and processing of RDF data into SPINdle theories.

The translated defeasible theories will then be composed into a single defeasible theory based on the logic proposed in [2], using the *Theories Composer*. Afterwards, the

[6] `https://github.com/RubenVerborgh/N3.js`

[7] Since LOV endpoint does not support the JSON format in the results, we have uploaded the data in `eventmedia.eurecom.fr/sparql`.

composed theory, together with other contextual information (as defined by user), will be loaded into the SPINdle reasoner to perform a compatibility check before returning the results to the users.

We have evaluated the time performances of the LIVE framework in two steps. First, we evaluate the time performances of the licenses compatibility module: it needs about 6ms to compute the compatibility of two licenses. Second, we evaluate time performances (Chrome v. 34) of the whole LIVE framework for the 10 LOD datasets with the highest number of links towards other LOD datasets, considering both the licenses retrieval module and the licenses compatibility one. The results show that LIVE provides the compatibility evaluation in less than 5 seconds for 7 of the selected datasets. Time performances of LIVE are mostly affected by the first module while the compatibility module does not produce a significant overhead. For instance, consider Linked Dataspaces[8], a dataset where we retrieve the licensing information in both the dataset and the adopted vocabularies. In this case, LIVE retrieves in 13.20s 48 vocabularies, the license for the dataset is CC-BY, and the PDDL license is attached one of the vocabularies[9]. The time for verifying the compatibility is 8ms, leading to a total of 13.208s.

3 Future perspectives

We have introduced the LIVE framework for licenses compatibility. The goal of the framework is to verify the compatibility of the licenses associated to the vocabularies exploited to create a RDF dataset and the license associated to the dataset itself. Several points have to be taken into account as future work. More precisely, in the present paper we consider vocabularies as data but this is not the only possible interpretation. For instance, we may see vocabularies as a kind of compiler, such that, after the creation of the dataset then the external vocabularies are no more used. In this case, what is a suitable way of defining a compatibility verification? We will investigate this issue as well as we will evaluate the usability of the online LIVE tool to subsequently improve the user interface.

References

1. Cabrio, E., Aprosio, A.P., Villata, S.: These are your rights: A natural language processing approach to automated rdf licenses generation. In: ESWC2014, LNCS (2014)
2. Governatori, G., Rotolo, A., Villata, S., Gandon, F.: One license to compose them all - a deontic logic approach to data licensing on the web of data. In: International Semantic Web Conference (1). Lecture Notes in Computer Science, vol. 8218, pp. 151–166. Springer (2013)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
4. Lam, H.P., Governatori, G.: The making of SPINdle. In: Proceedings of RuleML, LNCS 5858. pp. 315–322. Springer (2009)

[8] http://270a.info/

[9] http://purl.org/linked-data/cube