Efficient Use of DALICC in Data Processing Pipelines with Fuzzy License Information

Kurt Junghanns[1], Michael Martin[1], Norman Radtke[1] and Sabine Gründer-Fahrer[1]

[1]*Institute of Applied Informatics (InfAI), Leipzig*

Abstract

Integration of huge amounts of data from various sources forms the basis for many of today's (web) applications and use cases. In order to be able to reuse, transform, process, analyze, aggregate and republish such data, terms of use and license information are particularly important. DALICC provides a very comprehensive database for licenses and offers viable services for license handling that have been made available as open source. In this paper, we outline how DALICC can be used in (web) applications (represented by the project COYPU) in which heterogeneous data from various data sources with different usage conditions are processed. The paper aims to provide feedback on DALICC, to discuss necessary adjustments and present extensions that have been made by us. The overarching objective is to further activate and cooperatively develop the DALICC ecosystem.

Keywords

License clearance, DALICC, Dataset license processing, RDF Knowledge Graphs, Constraints, Feedback

1. Introduction

In many (web) applications, data processing happens, originated from Open Data portals, repositories and web services of various organizations. To reuse, transform, process, analyze, aggregate and republish such data, terms of use and license information are particularly important. When designing processes for searching and finding, updating, synchronizing and integrating data efficiently and programmatically, general metadata (e.g. authorship, update information) and license information are to be made available machine-readable, for example via Data Catalog Vocabulary (DCAT). Terms of use that are declared by specifying a license can be retrieved by querying Data License Clearance Center (DALICC) [1] using its license name. Afterwards such retrieved terms of use can e.g. be made available to application user group on the one hand, but can also be used to control the application logic on the other hand.

The Cognitive Economy Intelligence Platform for Resilience of Economic Ecosystems[1] (COYPU) is one of those applications that very intensively processes heterogeneous open and closed datasets (partially in combination), transforms them into knowledge graphs, aggregates and analyzes them and republishes parts of the resulting datasets (raw and through visualizations). Based on the technical COYPU ecosystem, scientific, governmental and industrial

SemTech4STLD: Semantic Technologies for Scientific, Technical and Legal Data, May 28, 2023, Hersonissos
junghanns@infai.org (K. Junghanns); martin@infai.org (M. Martin); radtke@infai.org (N. Radtke); gruender-fahrer@infai.org (S. Gründer-Fahrer)
0000-0003-1337-2770 (K. Junghanns)

[1]https://coypu.org/en/

(commercial) usage scenarios will be supported which aim at the evaluation and improvement of the resilience of value creation networks (companies, markets and regions) and decisions for action. Information about global crises, sanctions, environmental disasters, scarcity of resources, geopolitical conditions and information about markets, companies, products, logistics is included. It comes from various sources (i.e. *ACLED*[2] and *Weltrisikoindex* [3]), in different formats and is very heterogeneously linked to metadata and license information[4]. The following types of data can be distinguished on basis of their of the terms of use and the processing possible.

- **Data processable -** License is interlinked; terms of use can be dereferenced (results are machine-readable).
- **Data processable with additional effort -** terms of use are given, but have to be manually converted in a machine-readable form (license not given or not dereferenceable)
- **Data not processable -** Neither license nor terms of use are given.

Due to the massive amount of data, it is particularly important to acquire the terms of use (permissions, prohibitions, obligations) as efficiently as possible. A further challenge is the aggregation of datasets for which the terms of use are not subsumed by license URIs. The Data Licenses Clearance Center (DALICC) [1] has a very comprehensive database for licenses and offers viable services for license handling that have been made available as open source. In this paper, we outline how DALICC can be used in large-scale (web) applications (represented by COYPU) in which heterogeneous data from various data sources with different usage conditions are processed. The aim of the paper is to provide feedback on DALICC, to discuss necessary adjustments and announce extensions that have been made by us.

2. Approach

DALICC[5] offers, among other things, services for efficient handling of licenses to providers of digital assets (commercial and open source) and interested parties. The publicly available software framework consists of three main components: license library, license search and license composer. DALICC is using RDF and SPARQL to provide processable license details, facet based search, conflict detection and license resolving published via Github[6]. The DALICC services are highly valuable and thus used in COYPU for (a) dereferencing the terms of use based on existing license URIs (forward search), (b) determining license compatibility (c) determining possible licenses based on given permissions, prohibitions and obligations (reverse search).

The Data Catalog Vocabulary (DCAT)[7] is the standard RDF vocabulary to describe data catalogs, data services and datasets with their metadata, esp. its sources. In this paper, DCAT is used to describe data sources and link to its licenses. The Open Digital Rights Language[8]

[2] https://acleddata.com/
[3] https://weltrisikobericht.de
[4] A full list of public data sources can be found at https://datasets.coypu.org/
[5] https://www.dalicc.net/
[6] https://github.com/dalicc/dalicc
[7] https://www.w3.org/TR/vocab-dcat-2/
[8] https://www.w3.org/TR/odrl-vocab/

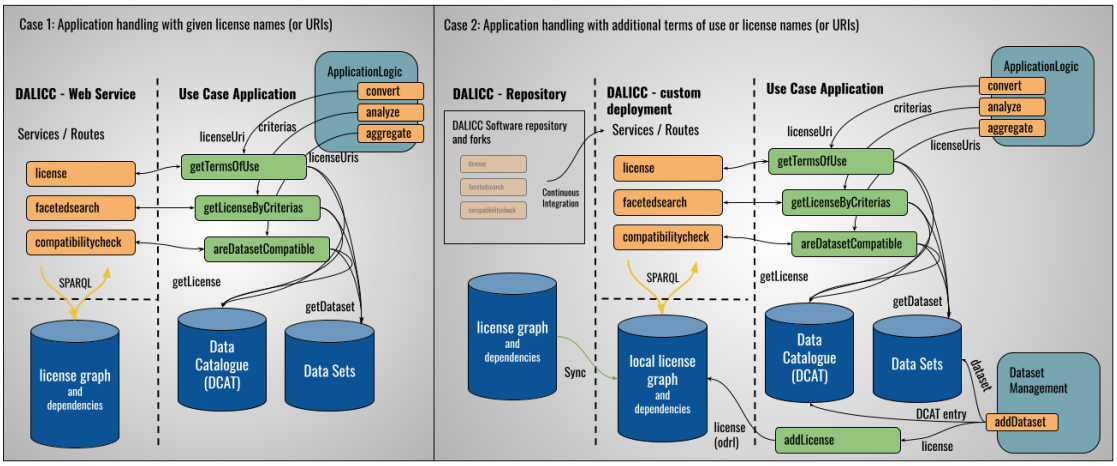

Figure 1: Draft of the common DALICC use case (case 1) and the COYPU use case (case 2)

(ODRL) is a policy expression language including a flexible and interoperable information model, vocabulary, and encoding mechanisms and providing a basis for representing statements about usage of content and services. Its concepts are used in DALICC and in our use case alike.

The COYPU approach has been created as an extension of the common DALICC use case. Figure 1 drafts the two use cases. In case 1, there is a license graph with a set number of licenses, available as a SPARQL endpoint and used by different API routes to support applications using processable datasets which have their licenses linked. This already allows for retrieving details of licenses, facet based search and compatibility checking.

In case 2, the terms of use of various data sources do not match any of the licenses currently available in DALICC, or even not fit with any existing license at all. For instance, in COYPU, this affects 13 out of of 40 data sources of our test sample. Hence, users need more support with license information and - if necessary - to add it. Moreover, as the use case includes commercial as well as scientific application scenarios, an applicable solution should represent relevant distinctions very prominently and accessible in their respective license information and the DALICC API routes. Last but not least, whenever datasets from different sources are to be aggregated, combined, published and used together, users are dependent on the tools used and their terms of use. In a large-scale application scenario, such as COYPU, this functionality should be considered the most wanted as well as most complex requirement, as it builds and depends on the availability and quality of processes for the simple cases and a non-trivial combinatorial logic. Using DALICC as our basis, we have so far implemented[9] the following extensions to tackle the issues just mentioned. To the list of licenses available in DALICC we added RDF-representations of *Datenlizenz Deutschland – Namensnennung – Version 2.0* (DLDEBY20) as well as a license placeholder including prohibitions, duties and permissions for the *Weltrisikoindex*. At the same time, we modified the existing API as to deliver turtle instead of JSON as its output format and supplemented it by a flag reflecting whether commercial use of a dataset is permitted or not. Taking into account the sovereignty of DALICC with respect to their license resources and

[9]Gitlab repository: https://gitlab.com/coypu-project/dalicc

the COYPU tooling, which is performing IRI resolving only with respect to internal graphs, internal IRIs for all licenses are created, enriched by further internal information according to our project needs and represented in a Coypu license KG.

Future work will focus on adding more licenses and license details (e.g., regarding permission for data analysis and entanglement). To tackle the problem of derived terms of use for aggregated dataset, we are currently cooperating with legal experts to work out a basic combinatorial logic. Furthermore, we plan to implement new API routes for the validation and upload of new licenses as well as for tracking changes in terms of use or data licenses via semantic versioning. Most of our contributions will be made available via pull requests on Github to give back contributions to DALICC.

3. Conclusion and Open Questions

In application scenarios with well-defined and well-known data ecosystems, DALICC can be used for data license processing directly. Thereby it closes a large gap in getting data efficiently into applications. However, if license information and terms of use are not available or provided only rudimentarily, the efforts are significantly increasing and direct application of the approach may fail. At this point, DALICC still provides valuable support by offering its resources open source, thereby enabling further development, as has been outlined in this paper.

As alternative applications presumably have similar requirements for the efficient handling of license data and, in particular, for the addition of further license resources, the use of cross-project synergies would be a worthwhile goal. At this point, urgent questions appear and need to be discussed within the research community. For instance, are project-specific terms of use and licenses to be published in open repositories or are they to be committed via a central web service with a subsequent review? How could consistency of different additions to an assumed common resource be ensured? As a step into this direction, we envisage to evaluate to what extent additional terms of use can be reused by applying similarity measures. It would be interesting and helpful for us to have available other experience reports, in order to learn which extensions got implemented for application of DALICC in alternative project contexts.

With this feedback and our outlined approach on extending the currently available functionalities, we hope to contribute to the activation and further cooperative development of the DALICC ecosystem.

Acknowledgements The authors acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy of Germany in the project COYPU (project number 01MK21007[A]).

References

[1] T. Pellegrini, V. Mireles, S. Steyskal, O. Panasiuk, A. Fensel, S. Kirrane, Automated rights clearance using semantic web technologies: The DALICC framework, in: T. Hoppe, B. Humm, A. Reibold (Eds.), Semantic Applications, Methodology, Technology, Corporate Use, Springer, 2018, pp. 203–218. doi:10.1007/978-3-662-55433-3_14.