Towards Employing Semantic License Annotations for Sensor Data Profiling

Anna Fensel[1], Tassilo Pellegrini[2], Oleksandra Panasiuk[1]

[1] University of Innsbruck, Department of Computer Science,
Semantic Technology Institute (STI) Innsbruck
Technikerstr. 21a, A-6020 Innsbruck, Austria
{anna.fensel, oleksandra.panasiuk}@sti2.at

[2] Department of Media Economics,
University of Applied Sciences St. Pölten
Matthias Corvinus Strasse 15, A-3100 St. Pölten, Austria
tassilo.pellegrini@fhstp.ac.at

Abstract. The paper outlines the most up to date version of the semantic licenses library of Data Licenses Clearance Center (DALICC), and discusses the possibilities of employing it for data profiling. In particular, we outline possible real-life use case directions from the domain of the vehicle sensor data profiling, an approach for the evaluation of the DALICC system in use, as well as possible further directions for the settings requiring cooperation with the data owners, such as at digital workplaces.

Keywords: Data licensing, knowledge graph, semantic technology, sensor data, use case, evaluation.

1. Introduction

With large amounts of data being available, the profiling of data gets very important and has become an active research and development area [4]. The methods suggested up to now have focused mainly on the annotation of the contents of data, for example on datasets recommendation and linking [1], or vocabulary and vocabulary terms recommendations [11]. Licensing of data has been recognized as an important part of the data profiling [2]. Further, explicit license information in the data profiling will facilitate the implementation of laws such as General Data Protection Regulation (GDPR). To make the reuse of the data and content more efficient, such profiling should include semantic representations of the deontic conditions (permissions, prohibitions and duties) specified within licenses and provenance information about the associated data broadly in practice. The approaches and tools for such developments are still actively evolving.

The creation of derivative data works, i.e. for purposes like content creation, service delivery or process automation, is often accompanied by legal uncertainty about usage rights

and high costs in the clearance of licensing issues. The DALICC project [8] has developed a software framework[1] that supports the automated clearance of rights issues in the creation of derivative data and software works [7]. In essence, DALICC helps to determine which information can be shared with whom to what extent under which conditions, thus lowering the costs of rights clearance and stimulating the data economy.

Specifically, we present the current, most up-to-date version of the DALICC License Repository [10] containing the basic international and a large variety of the national licenses, with the semantic license models as well as the corresponding documentation. The repository's last extensions have significantly increased the number of licenses present in the platform, as well as substantially improved the documentation.

The linked data empowered repository for storing the structured semantic license data for specific license data is set up, and is currently the most or one of the most complete repositories of this kind. The data access is provided via appropriate interfaces using in particular REST/Web Service access, SPARQL endpoint (for semantic data). The repository is serving anybody who wants to address checking of the licenses' specificities and their compatibility, using reasoning engines or license design tools. The tools are suitable for license engineering in various scenarios, making them a very good foundation for use cases from different sectors.

One of the new scenarios related to the collection and processing of the sensor data includes enabling the data owners to give consent on how their data is used. This implies making the data sharing and usage policies explainable to the data owners, as well as giving the ability to the data owners to license their data to the service provider. In this paper, we elaborate how a semantic data licensing solution, such as DALICC, can be used for such sensor data sharing scenarios, particularly, for the vehicle sensor data. The data collected by various sensors in a modern vehicle are large in quantity and variety, record in detail various performance and usage aspects of the vehicle, and are broadly used in scenarios such as quality assurance and predictive maintenance of the vehicles, as well as increasingly in other scenarios, such as traffic flow optimisation or insurance policies. Such data has a sensitive character, e.g. it may characterize the driving style of the vehicle owner. As the vehicle user (as the data producer) is the owner of the generated sensor data, provisioning him/her a legally grounded data sharing or contracting solution is essential.

The paper is structured as follows. Section 2 introduces the semantic license library of DALICC. Section 3 describes examples of technical settings where the DALICC solution can be employed, particularly, in a scenario involving vehicle sensor data. Section 4 describes an evaluation approach for the DALICC system, and Section 5 concludes the paper and provides an outlook for future work.

2. Semantic License Library

During the DALICC project runtime, we have performed an in-depth Rights Expression Languages (RELs) evaluation that laid the foundation for compiling the relevant set of machine-processable RELs and complementing vocabularies [9]. Based on the research of existing

[1] www.dalicc.net

RELs, we chose the Open Digital Rights Language (ODRL)[2] as it is particularly suitable for modeling licenses. Furthermore, the ODRL vocabulary includes terms that are deprecated or supplemented by terms from Creative Commons (CC) REL. However, we discovered that even a combination of various vocabularies is not sufficient to represent all of the necessary license concepts. To fill this gap, we constructed a DALICC vocabulary and introduced additional terms. Moreover, DALICC utilizes a dependency graph for representing the semantic relationships between defined concepts (see Figure 1). The interdependencies here are crucial for the license modelling, and particularly for interoperation and reasoning scenarios, where e.g. the data comes from various platforms and is being profiled using various license semantic annotations.

We elaborated a modelling workflow whose purpose is to govern a user through the process of a license composition. This stream of activity was accompanied by a steep learning curve, especially as we detected multiple complex interdependencies between the domain ontology, the rights expression languages and the dependency graph (see Figure 2) representing the logical relationships between domain concepts. In DALICC, the license engineering is performed by letting the user to answer basic questions on what he/she needs from the license, and the answers are utilized in the workflows for the license selection or creation. The resulting licensing outcomes are semantically interoperable with existing semantic license semantic models, in particular employing the developed dependency graph.

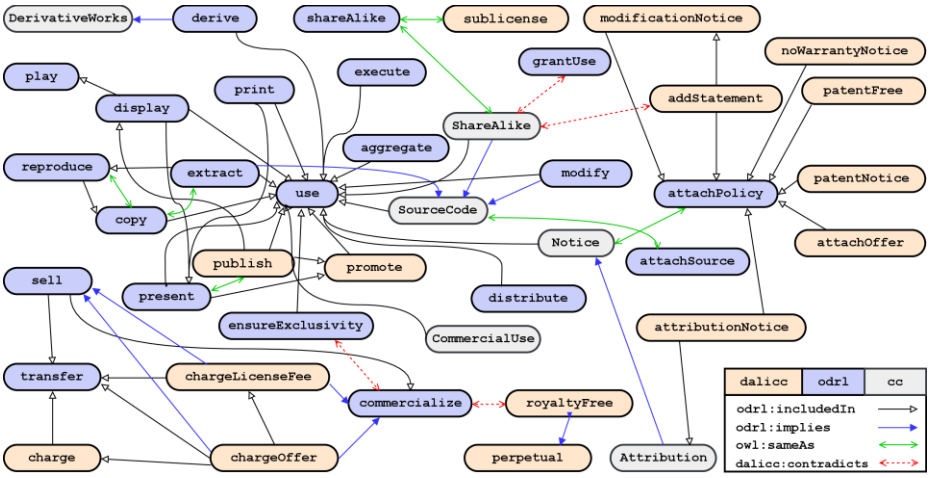

Fig. 1: Semantic Model for the Knowledge Graph of DALICC

[2] https://www.w3.org/TR/odrl-model

Finally, we set up the License Library [10], a repository containing currently 117 legally valid and approved licenses in human-readable and machine-readable form relevant for the licensing of digital assets. The data can be accessed via the publicly available demonstrator, and it can be retrieved via a REST/Web Service[3] and via the SPARQL endpoints: for the licenses[4] and for the licenses' metadata[5].

3. Employing the Semantic License Library for Data Profiling

The semantic licenses are intended to be used within the use cases of the CampaNeo project and are described below. In the CampaNeo project[6] (according to its proposal), "an open platform will be developed on which private and public institutions can create campaigns and collect and analyze vehicle data in real time. The goal is to set up a prototype platform for secure campaign-based data collection in Hannover, Wolfsburg and in cross-regional scenarios, as well as the implementation of the first smart use cases based on the campaign data. The focus is in particular on the data ownership of vehicle owners and the traceability of data processing". The campaigns will be run to collect and process the vehicle data to improve certain real-life situations e.g. in the city and regional traffic, insurance, etc., and ensure that the vehicle and the data owners are active first class participants of these campaigns.

When approaching the implementation of this goal, there are two challenges or use cases, where data licensing is of relevance. First, it should be communicated to the user which of his/her data may be used and in which manner, and an option to authorize the usage should be available; second, the usage contract has to be formed for the data, so that the data can be used at the platform.

3.1 Use Case 1: Transparent and explainable data sharing

A concept and a light-weight prototype for an approach for transparent and explainable sharing of the data will be designed. It will facilitate the understanding of the data sharing obligations and permissions, both for the data owners as well as for the data users. The actual data sharing workflows and the usages are also to be made traceable and displayable for the data owners, giving them a better understanding of the actual value of their data.

When the visualized data comes from a knowledge graph that has been built with machine learning (e.g. such as in Google's Knowledge Vault [5], or in the scenarios employing aggregated sensor data), the probabilities of the correctness of certain knowledge graph constructs will also be taken into account when visualizing the data. For example, the probability that the data will have an impact on one or another geographical region will be displayed. The latter can be implemented by analyzing the specifics of the structured and non-structured data of the geographical regions with similar characteristics, as well as taking of the known or expected trends into account. The approach and the solution will also contribute to the

[3] https://dalicc.net/license-library
[4] https://dalicc-virtuoso.poolparty.biz/sparql
[5] https://dalicc.net/license-library-meta
[6] https://www.sti-innsbruck.at/research/projects/platform-real-time-vehicle-data-campaigns

field of Explainable Artificial Intelligence. The works in the latter field up to now mainly focus on explaining the machine learning (in particular, deep learning) outcomes to the users, but little has been done so far in explaining the data sharing practices (especially the ones that are of larger scale and not easily comprehensible for the users) employing knowledge graphs, and especially in this project's domain.

The prototype (proof of concept) will be based on a web and mobile framework (such as Angular), which will enable it to be deployed in various settings and be independent from the specificities of proprietary app platforms.

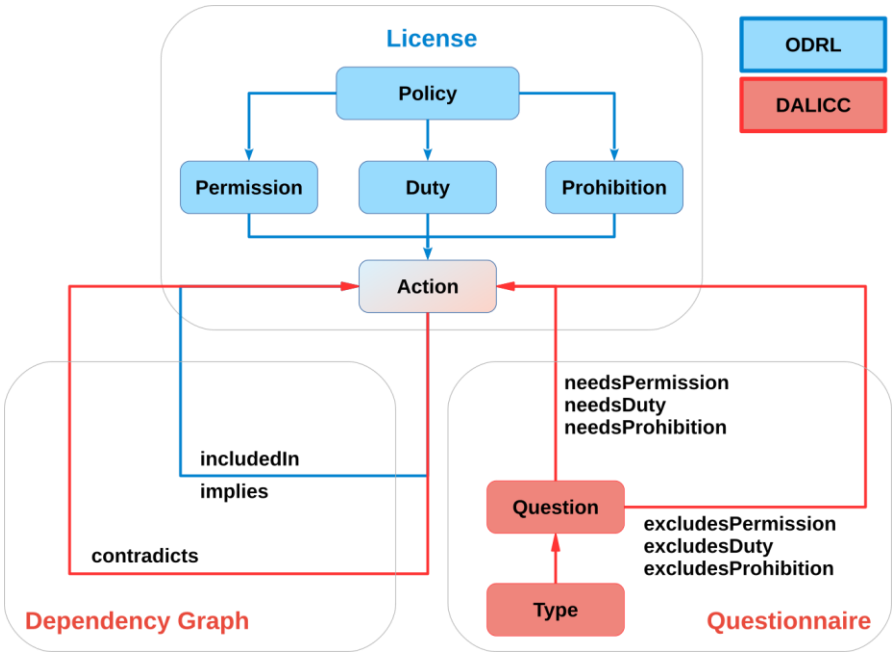

Fig. 2: Interplay of License Ontology, Dependency Graph and Modelling Workflow

3.2 Use Case 2: Knowledge graph based models for smart contracts

We will conceptualize the models needed for explaining the data sharing practices to the user, as well as for the formation of the smart contracts i.e. formalizations and protocols that are to be used for defining, controlling and executing the agreements comprising the data sharing rights and obligations. Technically, knowledge graphs, applying semantic modelling techniques, will be utilized. The modeled concepts will comprise the information needed for the representation of a smart contract, and will take into account the information about the relevant context (e.g. such information may comprise the records about the use of different parts

of the vehicle, geo data) and the users/user groups – typical data providers and consumers (e.g. drivers, manufacturer companies, public authorities). The knowledge graph based models will be used to comprehensively explicate the parts of the semantic models relevant for the transparency and the explanations. The CampaNeo data analytics module will be one of the sources of the raw data for the construction of the knowledge graph, and a part of the graph will be built employing machine learning techniques. The task's resulting models will be technically embedded in the project's smart contracts on a blockchain infrastructure.

Various techniques may be used to enable and facilitate the attachment of the semantic licenses to the data and content. The possibilities include such options as:

- employing meta tags: RDF file attached or link to the file. In particular, the Extensible Metadata Platform (XMP) ISO standard can be used to include links to the specific data licenses. XMP can be used in several file formats such as PDF, JPEG, JPEG 2000, JPEG XR, GIF, PNG, WebP, HTML, TIFF, Adobe Illustrator, PSD, MP3, MP4, Audio Video Interleave, WAV, RF64, Audio Interchange File Format, PostScript, Encapsulated PostScript, and proposed for DjVu,
- introducing hashtags determining the authors, timestamps and applied licenses: this technique would be useful for systems such as blockchains. It remains to be clarified whether the use cases will require storing of the whole license history and its evolution, in the way the blockchain systems typically enable it.

The corresponding tool support, such as a web application accessible with an API (that could work together with some platforms like GitHub, Facebook,...) may be realised. In particular, in Github or in Zip archives, the license can be inserted as the RDF/XML-file, instead of the usual text-file. A part of the solution may also include a Python library (e.g. employing an XMP Python toolkit) for connecting data and content files and licenses, or a stand-alone web service attaching the licenses to the data and content.

4. A Use Case-based Approach to DALICC Evaluation

While the main experimental contribution of DALICC is the design and implementation of a platform that facilitates correct usage of the semantically defined licenses, it is important to systematically approach the evaluation of the DALICC solution in the use cases.

In this way, we raise four hypotheses regarding the effectiveness of the proposed solution:

- H1: The use of the proposed platform facilitates correct selection and/or creation of the semantic licenses.
- H2: The solutions suggested by the proposed platform are clear and explainable to the users.
- H3: The use of the proposed system advances the content and data sharing economy.
- H4: The use of the system increases users' satisfaction / meets the users' goals.

For realizing the platform's components, we have been adopting parts of the well-known design science paradigm for information systems proposed by Hevner et al. [6]. All developed artefacts have been evaluated using representative sample scenarios (e.g. "create a new license", "assign a license to a dataset", etc.), and investigated with the help of well-defined case studies (up to now, the initial case studies from the DALICC project, and from now on

also the presented use cases of the CampaNeo project). These case studies allow us to draw conclusions about the general applicability of the developed artefacts and provide feedback for potential further refinement. Both development and evaluation of the platform's components have been carried out in close collaboration with legal experts from a law firm[7]. They ensured that the platform not only functions correctly technically, but also delivers correct results from the legal point of view. H1 and H2 have been already initially tested within the scope of the DALICC project development. The evaluations have been taking place involving the project experts and network (5 organisations from the project, and 5 organisations in addition, ca. 10 people in total) to exploit the platform with the aim to check if the basic scenarios are working correctly. For further use case driven evaluations, a similar approach is being followed.

In order to evaluate the proposed system with respect to all the hypotheses (and especially, H3 and H4), user studies are being performed. For this, a prototype of the proposed system has been deployed on the Web, and is available via the DALICC's website. We are facilitating further users, who have not been involved in the project, with targeted hands-on workshops, to get feedback on all the hypotheses. Further, we communicate our results to relevant bodies that can have a multiplier effect on the application of our solution, such as political bodies (e.g. the EC), recommendation/standardization bodies (e.g. W3C). The success of the work with them, in particular, impacts the outcomes for H3 and H4, as these depend on the level of priority set for such solutions by the regulators. Here, we are however optimistic, as the solutions for making the data usage practices more transparent and interoperable with the semantic technologies, as well as for making the data adding more value to the data owners are in the highest demand, as revealed in an EU Big Data research roadmap that takes into consideration the societal impact of the data [3].

5. Conclusion and Future Work

We envision approaches such as DALICC to change particularly the digital workplaces of the future, making the data profiling and sharing policies even more accessible. As defined by Gartner "The Digital Workplace enables new, more effective ways of working; raises employee engagement and agility; and exploits consumer-oriented styles and technologies."[8]. Ontological, or semantic, sharing of meaning is essential for the state of the art work scenarios, applicable to knowledge intensive labor, where also customers become collaborators. For example, the vehicle owners may choose to contribute their vehicle sensor data for one or another purpose (e.g. choosing to contribute the data to the city authorities, insurance companies, etc.). With development and application of new data and content licensing semantic techniques, we aim to bring the area of Digital Workplace to the new level, by assisting humans in highly intellectual tasks, that so far are barely being delegated to the machines:

[7] https://h-i-p.at/en

[8] https://www.gartner.com/it-glossary/digital-workplace/

namely, in decision making, content and data selection, creation and distribution, and management activity.

This will be achieved in the chosen application domains going beyond the current state of the art of ontology-based service interfacing, integration and participation involvement. The potential further work directions are as follows:

- Enabling easier license modeling of the data and content in both design time and the run time of the digital workplace scenario – and eventually the organisations creating their own applications and workflows based on these models,
- License-relying schemes rewarding and motivating data, content and service providers, that can be deployed in transparent infrastructures, such as blockchains; advancing the design and implementation of the data and content value chain and economy,
- Speeding up the velocity of the data and content flow in information systems (e.g. in scenarios connected with content generation, reporting),
- Making the decision processes transparent, traceable, and easier to optimize (e.g. it can be easily established which nodes are causing delays, inconsistencies), integrate new techniques facilitating easier data use in decision making, particularly, with the semantic information on how the data and content can be licensed,
- Visualisation of the data and workflows in a form that is actionable to humans in a digital workplace scenario, taking into account the license and provenance information.

Acknowledgements. The work is partly funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) - DALICC and CampaNeo projects.

References

1. M. Achichi, M. Ben Ellefi, Z. Bellahsene, and K. Todorov. Doing Web Data: from Dataset Recommendation to Data Linking. *NoSQL Data Models: Trends and Challenges*, *1*, 57-91, 2018.
2. M. Ben Ellefi, Z. Bellahsene, J. G. Breslin, E. Demidova, S. Dietze, J. Szymański, and K. Todorov. RDF dataset profiling–a survey of features, methods, vocabularies and applications. *Semantic Web*, (Preprint), 1-29, 2018.
3. M. Cuquet and A. Fensel. The societal impact of big data: A research roadmap for Europe, *Technology in Society*, Elsevier, 2018. DOI: https://doi.org/10.1016/j.techsoc.2018.03.005
4. S. Dietze, E. Demidova, and K. Todorov. RDF Dataset Profiling. *Encyclopedia of Big Data Technologies*, Springer International Publishing, pp.1378-1385, 2019.
5. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601-610, 2014.
6. A. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly, 28(1)*, 75-105, 2004.
7. T. Pellegrini, G. Havur, S. Steyskal, O. Panasiuk, A. Fensel, V. Mireles-Chavez, T. Thurner, A. Polleres, S. Kirrane, and A. Schönhofer. DALICC: A License Management Framework for Digital Assets, in: *Data Protection / LegalTech - Proceedings of the 22nd International Legal Informatics Symposium IRIS 2019, Colloquium. Presented at the IRIS 2019 - 21st International Legal Informatics Symposium*, Salzburg, Austria, 2019.

8. T. Pellegrini, V. Mireles, S. Steyskal, O. Panasiuk, A. Fensel, and S. Kirrane. Automated Rights Clearance Using Semantic Web Technologies: The DALICC Framework. In *Semantic Applications*, pp. 203-218, 2018.
9. T. Pellegrini, A. Schönhofer, S. Kirrane, S. Steyskal, A. Fensel, O. Panasiuk, V. Mireles-Chavez, T. Thurner, M. Dörfler, 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*, Salzburg, Austria, pp. 243–250, 2018.
10. O. Panasiuk, S. Steyskal, G. Havur, A. Fensel, and S. Kirrane. Modeling and Reasoning over Data Licenses. In: Gangemi A. et al. (eds) *The Semantic Web: ESWC 2018 Satellite Events*. Lecture Notes in Computer Science, vol 11155. Springer, pp.218-222, 2018.
11. I. Stavrakantonakis, A. Fensel, and D. Fensel. Linked Open Vocabulary ranking and terms discovery. In *Proceedings of the 12th International Conference on Semantic Systems*, pp. 1-8, 2016.