Data Licensing on the Cloud - Empirical Insights and Implications for Linked Data

Ivan Ermilov
University of Leipzig, Institute of Computer Science, AKSW Group
Augustusplatz 10, D-04109 Leipzig, Germany
iermilov@informatik.uni-leipzig.de

Tassilo Pellegrini
UAS St. Pölten, Department of Media Economics, Matthias Corvinus Str. 15, 3100 St. Pölten, Austria
tassilo.pellegrini@fhstp.ac.at

ABSTRACT

This paper investigates necessities and pitfalls in existing data licensing practices on the World Wide Web. The authors analyzed four open data portals with respect to the available licenses and drew conclusions about the quantity and quality of available licensing information. Additionally the authors address reasoning issues with respect to the automatic detection and potential clearance of licensing conflicts when creating derivative works from multiple data sources. The issues raised in this paper should be taken into account when designing and implementing a Linked Data licensing policy.

Categories and Subject Descriptors

Computing / technology policy [**Intellectual property**]: Licensing; Computing / technology policy [**Intellectual property**]: Digital rights management

Keywords

Data Portals, RDF Datasets, Licensing

1. INTRODUCTION

Data-driven innovations are governed by technological (i.e. standards) and non-technological (i.e. norms) influences [14]. The first define the good characteristics of a digital artefact, while the latter set the boundaries in which technology can unfold. One of these non-technological components of innovation are licensing policies. They provide information on ownership, provenance and utilization of intangible artefacts that can be protected by intellectual property rights. Licenses are enablers and barriers for economic transactions. They define the legitimate or illegitimate usage of data for commercial and non-commercial purposes.

This is especially relevant for Linked Data, which alters the asset specificities of data into a network good. Network goods are characterized by specific externalities like positive feedbacks and economies of scale stimulating network effects around the production, provision and distribution of data-driven goods and services [14, 17].

Under such circumstances licensing information - ideally in machine-readable form - is a critical factor in the efficient and legally secure handling of data, especially when it comes to derivative works compiled from various datasets. The utilization of datasets licensed according to open licenses allows for greater freedom in the reuse of data. However, the users still have to take into consideration prohibitions, requirements and permissions when consuming these datasets. This can be a time- and cost-intensive undertaking if sufficient licensing information is missing, resulting in increasing transaction costs and decreasing the incentives to reuse existing data.

But data licensing is not a trivial issue given the fact that data as an economic asset is difficult to define and to protect. Various kinds of data assets can be subject to different legal protection instruments like copyright, database right, competition law and patent law [16, 11]. Alternative licensing instruments like Creative Commons, Data Commons as well as open source licenses and derivatives thereof complicate the matter further, bringing about new challenges like appropriate licensing policies, license compatibility and machine-processability of licensing information - especially under conditions of dual licensing.

The article addresses these and other issues as follows. Section 2 gives a brief overview over related work on the topic of data licensing and reasoning over licenses. Section 3 gives empirical insights into licensing practices on four open-data portals under special consideration of machine-readable licenses. Section 4 addresses the reutilization of data under special consideration of compatibility issues. The authors outline a reasoning framework that helps to detect compatibility conflicts. Section 5 closes the paper with a recapitulation and perspectives on future work.

2. (LINKED) DATA LICENSING - RELATED WORK

With the increasing proliferation of open data, i.e. as part of public sector information initiatives or open innovation policies, the issue of data licensing gained attention among governments, companies and non-governmental organisations [10, 9, 7, 3]. As a consequence a wide array of data publishing guidelines were established [8, 19, 4], giving expression to the fact that licensing of (semantic) data is a fairly new kind of economic practice and still subject to debate concerning the adequate design of licensing policies [16, 11].

Most of work in the area of automatic processing of licens-

ing information is situated in the context of digital rights management systems. But so far little attention has been paid to the issue of license compatibility [12] and associated reasoning over machine-readable licensing information [5]. A logic for reasoning over the licenses was introduced by Pucella and Weissman [13], but their approach has not been implemented with semantic web standards. Therefore it is hard to operate on in the context of RDF licenses. Garcia and Gil [5] propose an OWL ontology to describe copyrights issues in closed datasets for rights clearance purposes. Their approach does not deal with alternative license models like Creative Commons or Open Source licenses. Thus it is not a viable solutions for problems arising from open data licensing. Villata and Gandon [18] describe the formalisation of a license composition tool for derivative works. They extend their research in [15, 6] by introducing deontic logic and heuristics for license composition. They use a subsumption approach for the comparison of the requirements, permissions and prohibitions of given licenses and derive new licenses out of them. Their work is an interesting approach to detect and potentially solve licensing conflicts by composing a new license. The pitfall of their approach lies in the circumstance, that an automatically composed license might result in logically correct but practically useless license. New licenses are created using machine-readable metadata, which are not necessarily in line with human-readable deeds and, what is more important, with lawyer-readable legal text. Additionally the new license might violate the intent of another rights holder who deliberatelly chose a more liberal license under the Share-Alike constraint. Thus, simply providing a deductive mechanism that chooses the most strict license for derivative works might not be sufficient.

Although progress has been made in the definition of machine-readable vocabularies for licensing like ccREL (Creative Commons Rights Expression Language) [2] or ODRL (Open Digital Rights Language) [1] empirical evidence presented in this paper reveals that the adoption rate of these standards is still very low.

3. ISSUES IN DATA LICENSING

3.1 Methodology

The authors chose three open government data portals and one open science data portal and aggregated the available licensing information of 441 315 individual datasets. Essentially, we selected the most popular data portals in the open data domain with respect to the available datasets.[1] For the government data portals we chose Publicdata.eu (EU), Data.gov (USA) and Open Data Canada (CAN). This allowed us to gain insight into regional differences in licensing practices. We additionally analyzed the portal Datahub.io, an open data repository provided by the Open Knowledge Foundation.

To collect the licensing information the authors developed an application[2] for crawling and aggregating data utilizing the CKAN API[3], caching the data in the local store for further processing by the application.

[1]The most up to date list of data portals can be found following the link: `http://dataportals.org/`

[2]Source code for the application is available on Github: `https://github.com/AKSW/ckan-aggregator-py`

[3]The description for CKAN API is available at: `http://docs.ckan.org/en/latest/api/index.html`

3.2 Results

In the following sections we will discuss major findings of our investigation. All in all the situation should be described as problematic with respect to the quality of licensing information and the quantity of individual licenses. It raises questions about the institutional viability of giant interlinked data clouds, given the fact, that a high degree of license heterogeneity requires technical means to effectively detect and resolve licensing conflicts. But for the time-being hardly any dataset provides licensing information in machine-readable form. Table 1 provides a brief overview of our findings.

Data Licenses on the Cloud				
.	Datagov	Open Canada	Public Data	Datahub
Datasets	132 206	244 257	55 481	9371
License Types	10	3	50	33
Not Specified	99.6 %	0.0 %	24.3 %	59.1 %
CC	0.4 %	0.0 %	35.3 %	17.1 %
ODC	0.0 %	0.0 %	0.5 %	4.8 %
Other	0.0 %	100.0 %	39.9 %	19.0 %
Deref. Link	0.4 %	100.0 %	43.2 %	23.1 %
Mach. Read.	0.0 %	0.0 %	2.6 %	2.2 %

Table 1: Data Licenses on the Cloud (as of May 2015)

3.2.1 Insufficient License Documentation

The first major finding was that a majority of the provided datasets lacks a sufficient amount of information about their licensing terms and conditions.

On Data.gov 99.6 % of all datasets lack explicit licensing information. Data from federal bodies are formally classified as "public" according to the US Open Data Policy[4] but this information is not explicated or referenced at the dataset level. Additionally this data policy is not binding for non-federal data providers, i.e. the City of New York or Cornell University. Approximately 25 % of all datasets on Data.gov are provided by non-federal organisations, who can define individual terms of use. Just 0.4 % of all datasets provide a de-referenceable link to their license.

Open Canada should be considered as a good practice with respect to governmental data licensing. They provide solely Canadian governmental data under three licensing types to choose from. 100 % of all datasets provide a de-referenceable link to their license.

On the contrary the European open government data platform Publicdata.eu 35.3 % of all datasets come along with Creative Commons and 40 % provide an individual license. Open Data Commons are associated to 0.5 % of all datasets. 24.3 % provide no licensing information at all. The documentations of the individual licenses vary considerably in depth and ease of accessibility. Nevertheless 43.2 % of all datasets provide a de-referenceable link to their license.

In the case of Datahub.io 59.1 % of all datasets lack any kind of licensing information. 17.1 % are provided under Creative Commons, 4.8 % under Open Data Commons and 19 % make use of an individual license. 23.1 % of all datasets provide a de-referenceable link to their license.

[4]`http://www.data.gov/data-policy`

3.2.2 *Heterogeneity of Licenses*

We can observe a high heterogeneity of licenses, but with strong regional differences. Publicdata.eu, which provides 55k datasets, utilizes 50 license types. On Datahub, with 9k the smallest of the analyzed data repositories, we found 33 different license types. Data.gov, which lists 55k datasets, makes use of 10 different licenses. Canada, which provides more than 240k datasets, makes use of just three license types.

The degree of heterogeneity varies in accordance to the type of repository and the region it is located in. In Europe (39.9 %) and in Datahub.io (19 %) we can observe a high degree of individual licenses especially for governmental data. This draws from the fact that a lot of countries have created a standard license of their own. For the time being it is not possible to determine the degree of heterogeneity in the US portal Data.gov. This would require a separate analysis of all datasets that are provided by non-federal organisations.

3.2.3 *Compatibility Issues*

License heterogeneity raises the question about license compatibility. Terms and conditions of two or more different licenses might contradict each other. So a coupling of these datasets for the purpose to create derivative works might be prohibited.

Within the analyzed datasets, where licensing information was specified, we found various conflicting licenses. In Publicdata.eu at least 2 % of all datasets are not compatible with the open definition. In Datahub.io it is 7 % of all datasets and on Data.gov it is 3 %. This might not seem overwhelming, but given the high amount of unspecified datasets it is easy to conclude that the degree of potential conflicts might be significantly higher. The situation is further complicated by the fact that especially in Europe most countries have created a standard license of their own introducing slight semantic differences in the definition of licensing terms thus adding to the complexity of the subject matter and the correct fulfillment of licensing terms and conditions. Due to the lack of machine-readability these licenses need to be checked manually for compatibility thus adding significantly to the transaction costs associated with the re-purposing of existing datasets.

3.2.4 *Machine-readable Licenses*

The analysis of the four data portals revealed that for the time being hardly no machine-readable licensing information is being provided. The only exceptions can be found where links to CC licenses are being provided, which allows to automatically retrieve the licenses terms and conditions explicated in ccREL [2]. This applies to 2.6 % of the datasets on Publicdata.eu and 2.2 % of datasets on Datahub.io. The other portals did not provide a machine-readable licensing information at all.

4. OPEN DATA LICENSES COMPOSITOR AND RECOMMENDER

As a first step in mitigation of the problems outlined above, we propose an approach which combines the licensing information and provides recommendations on the data usage. The main difference from previously proposed approaches is that we do not provide a combined license, but a recommendation and, possibly, a reference to an existing

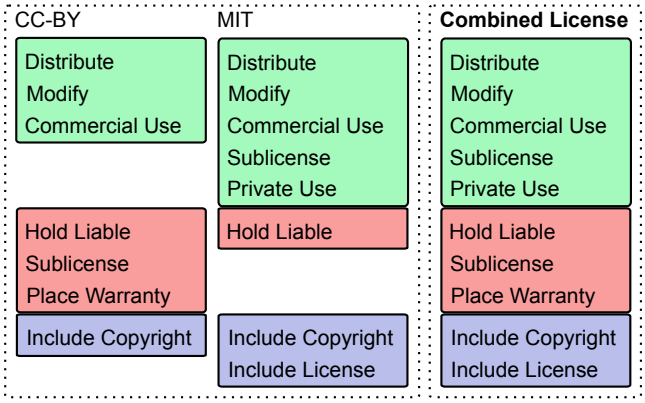

Figure 1: Composition problem on example of CC-BY and MIT licenses.

license compatible with all of the input licenses. In this section we describe an application for license composition and recommendation. The application processes the RDF representations of licenses as input data. To show case our application we utilize Creative Commons (CC) licenses.

CC licenses consist of three components: human-readable deeds, machine-readable metadata, and lawyer-readable licenses. Although, the license composition problem was previously addressed in [18, 15, 6], the combined licenses were created using machine-readable metadata, which leads to a new license that do not necessarily comply with human-readable deeds and, what is more important, with lawyer-readable legal text. This occurs when combining, for example, CC-BY (Creative Commons) and MIT (Open Source) licenses (see fig. 1)[5]. The combination in our example is achieved by a simple combination of all the features of each licence (i.e. permissions, prohibitions and requirements). The output license contains a set of permissions and prohibitions, which does not match CC-BY or MIT. Therefore the combined license does not match any existing legal text and can not be utilized as is.

The *Licenses Compositor and Recommender*[6] application combines open data licenses, inspects their compatibility and gives a single existing license with requirements (e.g. attribution) as a recommendation to the user. In fig. 2 we show a scenario, where a user chooses three datasets that come along with three different licenses. The chosen CC-BY, CC-BY-NC-SA and CC-BY-SA licenses are queried for requirements, permissions and prohibitions using SPARQL. After this step license features are combined and a recommendation is displayed: "These three licenses are not compatible. However, you can use the first two licenses together and license your data with CC-BY-NC-SA. Attribution for both licensors are required, notices should be kept."

The main purpose of the License Compositor is to provide a user with information, explaining the conditions for using licenses in combination. To accomplish this task, the License Compositor analyzes the features of input licenses such as permissions, requirements and prohibitions. After the analysis a decision is made on the licenses compatibility,

[5]The data about licenses is taken from `https://tldrlegal.com/`

[6]The application is available at: https://github.com/AKSW/LicenseCompositor

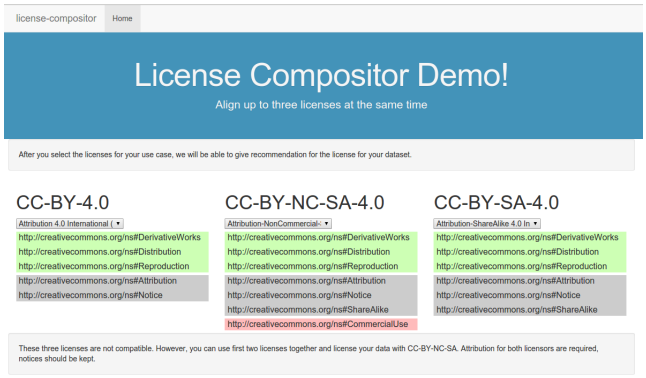

Figure 2: License Compositor and Recommender.

which clusters the licenses into sets, presenting to the user which licenses can be utilized in combination and on which conditions. Each set corresponds to a single existing license, we therefore avoid the problem of combining license features into a non-existing or non-appropriate license.

We foresee two usage scenarios for the License Compositor: **i)** a data consumer wants to exploit existing datasets from the Web to use inside her application, in this scenario user utilizes the License Compositor to check the datasets compatibility from the licensing perspective; and **ii)** a data provider chooses a dataset license in accordance with other existing datasets, in this scenario the License Compositor enables flexible license choice for a use case at hand.

5. LIMITATIONS AND FUTURE WORK

This paper presented empirical insight into the licensing practices on four data clouds. The findings reveal characteristic patterns and pitfalls of licensing practices with respect to region and domain specificity of the data portal. Further research is necessary to investigate compatibility issues under special consideration of the large amount of individual licenses. This raises the question how machine-readability of licensing information can be improved to enable automatic detection of conflicts and - in the long run - provide a infrastructure for automatic negotiation and clearance. At the time of writing the License Compositor exists only as a proof of concept with the interface limited to three licenses. Extending the License Compositor to represent sets of compatible licenses as well as implementing features such as automatic retrieval of licenses from the datasets is a subject of future work. To evaluate License Compositor we plan to use licenses from the Creative Commons and tldrlegal portals.

6. ACKNOWLEDGEMENTS

This work has been carried out within the project NoLDE (Network of Linked Data Excellence) funded by the Austrian Promotion Research Agency under grant number 3592880. We acknowledge support from GeoKnow project, GA number no. 318159, as well as BMWi project SAKE.

7. REFERENCES

[1] ODRL Community Group. https://www.w3.org/community/odrl/. Accessed: 2015-05-29.

[2] H. Abelson, B. Adida, M. Linksvayer, and N. Yergler. ccREL: The Creative Commons Rights Expression Language. W3C Member Submission, 2008.

[3] P. Archer et al. *Study on business models for Linked Open Government Data*. ISA programme by PwC EU Services. European Union, 2013.

[4] M. Frosterus, E. Hyvönen, and J. Laitio. Creating and publishing semantic metadata about linked and open datasets. In *Linking Government Data*, pages 95–112. Springer, 2011.

[5] R. García and R. Gil. Copyright licenses reasoning using an owl-dl ontology. *Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, 188:145–162, 2009.

[6] G. Governatori, H.-P. Lam, A. Rotolo, S. Villata, and F. Gandon. Heuristics for licenses composition. In *JURIX*, pages 77–86, 2013.

[7] L. Guibault and C. Angelopoulos. *Open Content Licensing: From Theory to Practice*. G - Reference,Information and Interdisciplinary Subjects Series. Amsterdam University Press, 2011.

[8] B. Hyland and D. Wood. The joy of data-a cookbook for publishing linked government data on the web. In *Linking government data*, pages 3–26. Springer, 2011.

[9] E. Hyvonen. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2012.

[10] P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's no money in linked data. 2013.

[11] T. Pellegrini. Linked data licensing – datenlizenzierung unter netzökonomischen bedingungen. In *Transparenz. Tagungsband des 17. Internationalen Rechtsinformatik Symposium IRIS 2014*, pages 159–168. Verlag der Österreichischen Computergesellschaft, 2014.

[12] J. Prenafeta. Protecting copyright through semantic technology. *Publishing research quarterly*, 26(4):249–254, 2010.

[13] R. Pucella and V. Weissman. A logic for reasoning about digital rights. In *Computer Security Foundations Workshop, 2002. Proceedings. 15th IEEE*, pages 282–294. IEEE, 2002.

[14] E. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 2003.

[15] A. Rotolo, S. Villata, and F. Gandon. A deontic logic semantics for licenses composition in the web of data. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 111–120. ACM, 2013.

[16] M. Sonntag. Rechtsschutz für ontologien. In *e-Staat und e-Wirtschaft aus rechtlicher Sicht*, pages 418–425. Stuttgart: Richard Boorberg Verlag, 2006.

[17] H. Varian, J. Farrell, J. Farrell, and C. Shapiro. *The Economics of Information Technology: An Introduction*. Raffaele Mattioli Lectures. Cambridge University Press, 2004.

[18] S. Villata and F. Gandon. Licenses compatibility and composition in the web of data. *COLD*, 905, 2012.

[19] D. Wood. *Linking government data*. Springer Science & Business Media, 2011.