The effects of Licensing on Open Data: Computing a measure of health for our Scholarly Record

Richard Hosking, Mark Gahegan

University of Auckland, Department of Computer Science and Centre for eResearch,
Building 409, Rm G21, LG, 24 Symonds Street,
Auckland, New Zealand,
{r.hosking, m.gahegan}@auckland.ac.nz

Abstract. As data collections become established in key disciplines, some of the longstanding barriers to data sharing become to dissolve; yet others remain. While metadata and ontologies help overcome the problems of finding and interpreting data, the lack of clarity over licensing remains a real impediment to data reuse. Freedom from legal restriction and uncertainty is essential for the effective sharing, combining and deriving of data from these distributed collections. Reuse and recombination of data will be greatly facilitated by expanding the definition of the semantic web to include the semantics of data licensing. We aim to express licensing terms in a computable manner, within the context of research practice, enabling us to infer the resulting state of rights, obligations and conditions that are inherited by derived and recombined datasets, using a mixed bag of licenses. Building off this we aim to simulate the effects of varying licensing practices within communities, proposing a measure of health of our scholarly record based on compatibility and restrictiveness of the licenses contained therein.

1 Introduction

The semantic web has brought untold opportunities to share and reuse data. In this research we address an uncertainty in data reuse and recombination. Expressing the terms and conditions of use through data licenses guides allowable usage, but even the use of prominent licenses, such as Creative Commons [1] and GPL [2] can surface incompatibilities that would hamper certain data combinations. We develop a semantic model to describe data licenses, allowing us to compute the impacts of licensing decisions on integrating the scientific record, especially on the derivation of new data from existing datasets. For instance, what would be the resulting state of rights and obligations from combining two related data sets with different underlying licenses, or deriving new data by querying an existing distributed collection? The resulting restrictions stem from two related but distinguishable forces: (i) The explicit terms or reserved rights placed on data content, and (ii) incompatible licensing conditions across content. Automated computational tools are bringing this problem into sharp focus; for example, data mining a corpus [3] of heterogeneously licensed work, or a large data integration effort such as creating global land cover maps from many national maps [4, 5]. We currently have limited means of measuring our collective freedom to integrate and republish our scientific record [6]; Lacking an effective means of expressing licensing terms that are semantically computable hinders our ability to efficiently and effectively reason over the legality of our research practices and products.

2 Relevancy

We require new measures of the health of our research data, both as individual entities, and as an interdependent ecosystem. The culture shift towards open access and open licensing provide momentum and direction towards a more integrated scientific record. However, there has been little attention paid so far to the effect that data licensing plays in complicating or inhibiting the reuse and recombination of data; the work of Wilbanks [6,7] being a notable exception. Providing useful answers presents many challenges, such as classifying data by permissible uses, through to determining the resulting legality of deriving new data from our current records, and subsequently the limitations that the new dataset will carry. Data portals, repositories and collections play an essential role disseminating scholarly data, but to what extent are enforcing 'open' licensing ideals on data ingress ostracizing valuable data? To aid with these questions, we require better measurements: Firstly to gain greater understanding of the challenge, and secondly to act as an essential resource guiding collective governance over these issues.

In this research we focus on the human imposed restrictions and obligations placed on scientific data in its contemporary digital form. We present this control as inertia, reducing or even preventing reuse. The prevailing position of Copyright Law, and the individual and combined effects of licensing provide the source of restrictions. With uncertainty stemming from both ambiguities in these terms and the novel usage demanded from scientific research. To bring about effective sharing and reuse of the scholarly record we need to address this burden of control and restrictions. From the standpoint of the semantic web, this research contributes a novel application of ontological reasoning, provides simulation tools to explore the effects of combining data that uses a variety of license types, and therefore supports the notion of an open web of linked data that is free from legal risk.

3 Related Work

The moral and pragmatic imperatives for fostering a more open and reusable scientific record have been well made [6–8]; our legal tools are also advancing to meet our evolving conceptualization of openness. The Free Software Movement's [2] virally 'open' licenses—challenging the closed and proprietary model of software development—includes explicit provisions that derived works must also carry similar affordances [9]. This particular approach of propagating values brought with it a cultural movement. With the advent of the Creative Commons family of licenses [1] we observe a shift towards the public domain, bringing a much-simplified licensing model while also employing the use of graphical notations to improve clarity. Some affordances of control remain for authors, such as preventing commercial use, and—borrowing from the Free Software Movement—the share-alike clause, to allow propagating of intent to derivative works. Most recently the legal commitment to the commons was strengthened with the creation of waivers [10, 11] designed to release (to the fullest extent possible under the law) all conditions and reserved rights on content. This provides a useful baseline and it's difficult to envision a situation that provides more freedom at scale.

Rights Expression Languages provide formal machine-readable expressions of licenses, MPEG-21[12] and the Open Digital Rights Language (ODRL)[13] being the most prominent. In creating these representations, we have learnt the importance of understanding content lifecycles, granularity, ambiguity, extensibility and choosing an appropriate formal language for these representations [3, 14, 15]. We observe a trend towards standardized formats. For example the original right expression language, DPRL [16], was written in LISP [17] but subsequently (following its standardization by the W3C) migrated to XML DTD [18], then XML Schema [19]. Then Creative Commons developed its own machine-readable expression language (ccREL) [20] opting to use RDF [21]. A distinguishing point in these expression languages is the role of machine-actionable control over use of the content; ccREL relies on the existing copyright law to protect digital content, MPEG and ODRL providing integration to enforcement systems. In order to connect the CC licenses to copyright law, the Creative Commons initiative created a set of human-readable "classical" licenses. These licenses are also available as summarized, and graphical notation, for users not interested in the legal text.

Aside from Creative Commons (ccREL), there is a distinct lack of support for Copyright law in these representations [14, 22]. Soft rights such as 'fair use' are difficult to encode in rights languages due to their ambiguous nature. Most existing Rights Expression Languages take the "Everything not permitted is forbidden" approach, while ccREL uses the Berne convention [23] as a baseline, permitting any action this is not explicitly stated, such as 'NonCommercial' or 'NoDerivative'. Interoperability concerns, and the lack of support for Copyright law drove the development of Copyright Ontologies: The Semantic Copyright project [24], and the Copyright Ontology [15] respectively. Spurred by the goals of supporting an Intellectual Property Registry, the Semantic Copyright projects provides rich representation for the attributes of work, and even supports basic reasoning over allowable uses, limited by only capturing basic rights. The Copyright Ontology was designed to extend notions of Copyright Law to existing rights expression languages, providing a layer of interoperability and the promise of web standards. The approach relied heavily on the subsumption capabilities of OWL-DL, but could only capture the expression and transfer of rights, neglecting obligations and conditions. Additionally, its strict interpretation of Copyright law does not address the diversity and potential misalignment between terms of various licenses. Both these efforts explicitly exclude cross-jurisdictional issues, yet data licenses often suffer from these ambiguities. Lastly neither of these efforts has attempted to compute the resulting state of rights and obligations on derivative works.

4 Research Questions

There exists a rich discussion on the role licensing plays on our research data [3,6,7] (and even richer examples of it being reused [8]), however there exists a disconnection between these discussions, the current means of expressing licensing terms and measuring the impacts on research practice. Without this our tools will remain ignorant of the social context in which they exist. We aim to answer the following:

1) How can data licensing be expressed semantically to include the rich interpretative discussion, and its relationship to research practice?

Data-driven research [8] is premised on the availability of quality data, with combing and deriving this data a fundamental activity of successful research. But faced with interpreting licensing constraints on performing these common tasks, it is often unclear in terms of both the permissibility of undertaking a given action, and the propagating state of rights and obligations to the result. Faced with this, we aim to answer:

2) How can we compute the resulting state of rights, obligations and conditions associated with derived data, particularly when multiple license types are implied?

Individual actions do not occur within a vacuum, nor in the case of licensing are they immutable. The move towards open access means practices and community norms need to change. Effective change relies on constructive discussions based on the availability of knowledge. We aim to address the lack of congruency within and between data hosting communities.

3) Can we compute a health measure based on #1 and #2 to describe how combinable (distributed) community datasets might be, based on their licensing?

5 Hypotheses

Licensing norms will differ between communities, with some practices being more advantageous. Thus the impact of licensing on research practices will vary. Towards the aim of computing a measure of health we develop measures that allow us to assess legal aspects of openness both within and between communities. We put forward the following to serve as a basis for creating these measurements:

- An increased number of licensing conditions will decrease reusability and / or increase legal uncertainty
- A greater diversity of licensing terms across a corpus will decrease the aggregate level of reuse
- Increased use of non-standard terms will increase legal uncertainty
- Increased restrictions and conditions on a dataset will decrease its ability to be integrated into existing workflows and collections

Thus, given our ability to compute over individual licensing implications, we put forward the following dimensions as meaningful measures to evaluate the health of a collection of scholarly data:

- Uniformity: *A measure of the degree and distribution of different legal affordances of a collection*
- Combinability: *The extent of interoperability between the licensing conditions placed across a collection*
- Level of clarity: *A measure of the legal uncertainty that underlies the use and combination of the collection*

The quality of this model will be judged on its ability, given a collection with differing underlying licenses, to generate these meaningful measures that facilitate understanding and enable comparison between licensing practices.

6 Approach

Collectively this research contributes to two aims, firstly to capture, compute and usefully convey the practical effects of licensing on our research data, and secondly demonstrate the value of this by developing a measure of health for our scholarly record that will, we believe, help in the pursuit of best practice policies to encourage data reuse.

6.1 Conceptualization

In creating a rich representation of this domain, we must first ask:

i. How diverse is the terminology commonly used to license data, and how can these semantics be formally captured?
ii. What are the relationships between the terms used in data licensing and undertaking common research activities that operate on the data?
iii. How in practice do the conditions, obligations and rights combine when we derive new data?
iv. How can we compare licensing practices between communities?

1. We will firstly undertake a review of the discussions and observations around data licensing. This will provide a rich comprehension of the effects of various licenses aiding our interpretation of the meaning and impact of these terms.

2. Next we conceptualize a model of data usage in the sciences, primarily focusing on commonly used verbs. Understanding will be drawn from existing workflow and lifecycle models, along with analytical methods and phrases in common use.

3. We will then model the most prominently used data licenses and a selection of commonly applied custom terms. Undertaking a matching exercise[25] we will then relate the terms and conditions of these licenses to each other and research actions.

4. Following, we will model licensing practices across several communities within the Earth and Environmental Sciences, and in parallel generate synthetic data collections that caricature various licensing strategies. Together this will enable us to simulate the resulting effects of combining and deriving data.

6.2 Formalization

In creating a computable formalism we aim to ascertain:

i. the legality of performing research actions;
ii. the compounding affects of related licensing terms, and the union of multiple licenses and
iii. the sets of data we are able to effectively integrate or work with for a specified purpose.

1. We will formalize our conceptual model into the Web Ontology Language (OWL). The level of expressivity will be determined by the requirement to reason over the legality of research actions that involve data under license.

2. Particular focus will then be placed on research actions that result in the creation of new data, especially derived works resulting from multiple sources. Here we aim to compute the resulting state of: rights, restrictions and obligations.

3. We will then develop an agent-based model. This will allow us to simulate the emergent affects of licensing over time, with a focus on understanding the propagating rights on derived data. We chose an agent-based approach to extend the metaphor of a data-ecosystem allowing us to explicitly study generations of simulated data use and recombination, and understand how changes in the underlying licensing practices impact the health of the ecosystem.

The aim of these tasks is to develop a reusable and extendable formalization that answers the stated questions of this research to provide ongoing value in understanding the impacts licensing has on data reuse. Additionally by developing a machine-readable representation of licensing, this allows us to both capture the observed practices within real communities, and also affords us the ability to generate synthetic collections characterizing various licensing strategies. By removing the dependency to gather real data on licensing practices we are able to develop and test our model in parallel to solving the additional challenge of gathering sample data.

7 Evaluation

In addition to successfully demonstrating the value of each step of our approach, we conduct an evaluation of this research, broken into two components: Firstly a review of the ontological reasoning we have developed, and secondly an analysis of the simulations we run. The areas on which we focus for our review of the developed reasoning are:

- Coverage of licensing terms
- Extensibility of our representation to incorporate additional licensing terms
- The level of expressivity and the reasoning facilitated.

By using synthetic data sets of caricatured strategies as well as harvesting sample data from the web we will evaluate our measures of data collection health along with our model that generates them. We use the following set of criteria:

- Explanatory adequacy – *Does it help make sense of the observed data*
- Interpretability – *Are the component of the model understandable and linked to known processes*
- Descriptive adequacy – *Does the model fit the observed data*
- Principal of Simplicity – *is the model overly complex for the task*
- Generalizability – *is the model a good predictor of future observations*

Descriptive accuracy will be particularly difficult to measure for two reasons. Firstly, the predicted effects on the reuse of data are only one element of a much larger ecosystem thus any correlation to observed data must take into account a diversity of additional processes. Secondly, the impacts of licensing are largely felt privately and dispersed globally; without richer publication of both data provenance and data citations it is very difficult to gather evidence.

8 Reflections

Relative to existing attempts to capture rights, or ontologies to represent Copyright we aim to make misalignment, uncertainty and ambiguity first class citizens, due to their implications on research practice. We build on existing knowledge that discusses the effects of licensing, and expand on the current means and medium of sharing and applying this knowledge. We place importance on the role of governance in making progress in this area, thus this work is strongly guided by fostering effective governance strategies for data. We aim to constructively contribute to the discussion of utilizing semantic tools to measure and improve our understanding of such an important aspect of contemporary science. We aim to not replace the role of legal advice, nor directly guide the application of licensing; as we do not venture into the larger social and economic norms that form an essential component of these choices. What we aim is to facilitate the discussion, and ensure that we are all asking the right questions.

References

1. Creative Commons: About The Licenses, http://creativecommons.org/licenses/.
2. Free Software Foundation, I.: GNU GENERAL PUBLIC LICENSE, http://www.gnu.org/licenses/gpl.html.
3. Korn, N., Oppenheim, C., Duncan, C.: IPR and Licensing issues in Derived Data. Report submitted to the JISC. 1–12 (2007).
4. Gobal Land Cover Network, http://www.glcn.org/index_en.jsp.
5. Global Observation for Forest Cover and Land Dynamics, http://www.fao.org/gtos/gofc-gold/ http://www.fao.org/gtos/gofc-gold.
6. Wilbanks, J.: Public domain, copyright licenses and the freedom to integrate science. JCOM 7, (2008).
7. Wilbanks, J.T., Wilbanks, T.J.: Science, Open Communication and Sustainable Development. Sustainability. 2, 993–1015 (2010).
8. Hey, T., Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft, (2009).
9. Stallman, R.M.: Free Software, Free Society.
10. Commons, C.: CC Zero, http://creativecommons.org/publicdomain/zero/1.0/legalcode.
11. Open Knowledge Foundation: Open Data Commons Public Domain Dedication and License (PDDL), http://opendatacommons.org/licenses/pddl/.
12. Paper, A.W.: The MPEG-21 Rights Expression Language. 1–16 (2003).
13. W3C: ODRL Community Group, http://www.w3.org/community/odrl/.
14. Coyle, K.: Rights Expression Languages A Report for the Library of Congress. 1–53 (2004).
15. García, R.: A semantic web approach to digital rights management. Doctorate in computer science and digital communication, Department of Technologies, Universitat Pompeu Fabra, Barcelona. (2006).
16. Digital Property Rights Language, http://xml.coverpages.org/dprl.html.
17. Lisp Programming Language, https://en.wikipedia.org/wiki/Lisp_(programming_language).
18. W3C: Extensible Markup Language (XML), http://www.w3.org/XML/.
19. W3C: XML Schema, http://www.w3.org/XML/Schema.
20. Abelson, H., Adida, B., Linksvayer, M., Yergler, N.: ccREL : The Creative Commons Rights Expression Language. 1–32 (2008).
21. W3C: Resource Description Framework (RDF) , http://www.w3c.org/RDF/.
22. Barlas, C.: Digital Rights Expression Languages (DRELs). JISC Technology and Standards Watch. (2006).
23. Act, P.: Berne Convention for the Protection of Literary and Artistic Works Berne Convention for the Protection of Literary and Artistic Works. 1–29 (1979).
24. Semantic Copyright Project, http://semanticcopyright.org/index.php/ontology.
25. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering. 25, 158–176 (2012).