OPEN DATA COMMONS, A LICENSE FOR OPEN DATA

Paul Miller
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
paul.miller@talis.com

Rob Styles
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
rob.styles@talis.com

Tom Heath
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
tom.heath@talis.com

ABSTRACT

Attendees at the WWW2007 panel session on Open Data [1, 2] will remember a wide-ranging discussion of the role that easily accessible data could play in endeavors from scholarly publishing [3] to the creation of canonical product catalogs [4]. The authors argued there and subsequently [5] that an effective and flexible licensing framework is needed in moving forward.

Paradoxically, we argue that you need to actively and consciously assert your desire that third parties be able to use data you place online in order for those 'visible' and 'accessible' data sets to be utilized most effectively.

Significant progress has been made in the past twelve months, with engagement [6] from Creative Commons [7, 8, 9] and others [10] resulting in a license [11] and notion of 'community norms' [12] upon which all can build.

Categories And Subject Descriptors

E.m [Data, Miscellaneous]

General Terms

Legal Aspects

Keywords

Open Data, Linked Open Data, Licensing, License, Creative Commons, Science Commons, Open Data Commons.

LDOW2008, April 22, 2008, Beijing, China.

1. INTRODUCTION

Much attention is currently being paid to the concept of Open Source [13], and to the value its adoption can bring to the development and dissemination of software within a vibrant mixed economy comprising traditionally commercial, open source, and hybrid solutions of various forms. In the academic sector, too, existing models of publication are being challenged by the rise of the philosophically related Open Access [14] movement. Here, as in the software world, the vehement polarization of early protagonists is increasingly giving way to a more pragmatic world view in which various models co-exist to meet a diverse set of requirements.

In scholarly publishing, there has tended to be an unfortunate presumption that rights in the raw data underpinning a paper's analyses and conclusions will be retained and enforced; that these data will not be shared in order to allow readers to test the results for themselves. More recently, some funders have begun to require that both reports *of* research and data produced *by* research be made easily available for re-examination, and organizations such as Creative Commons are taking a serious interest in this area with their Science Commons project.

However, beyond these scholarly disciplines far less attention has been paid to the manner in which data can be used and reused, with only a few projects such as OpenStreetMap [15] really challenging the traditional models of control over creating and accessing the underlying data upon which so many applications rely.

Almost everywhere one looks, now, increasing volumes of data are being published to the Web with the explicit aim of interoperability and a strong but often implicit commitment to openness. Despite this commitment in principle, data is rarely made available in a manner that makes it straightforward to ascertain the uses to which it may subsequently be put by a third party. In small, tightly-knit groups where interchange of data may be governed by existing social norms this may rarely present a problem. However, with data interchange and interoperability reaching Web scale, social norms alone cannot be relied upon to enforce fair and appropriate usage of data. Instead, licenses are required that make explicit the terms under which data can be used. By explicitly granting permissions, the grantor reassures those who may wish to use their data, and takes a conscious step to increase the pool of Open Data available to the web.

In this paper we will briefly outline and contextualize existing work in the field, highlighting the cases in which existing licenses are appropriate and those areas in which they can not be meaningfully applied. We will then present the work of the Open Data Commons, and describe the rationale behind the Open Data Commons Public Domain Dedication and License.

2. DATA IS NOT A CREATIVE WORK

Discussion of opening access to resources on the web often turns, sooner or later, to the laudable activities of Creative Commons, and we shall look at this effort in a little more detail shortly. It is important to understand at this point, however, that the legal protections upon which Creative Commons (and other similar) licenses rely depend upon national and international legislation around Copyright. Copyright protection applies to acts of creativity ('creative works'), and categorically does not extend either to databases nor to those non-creative parts of their content. Despite numerous cases in which well-meaning individuals or organizations release data onto the Web and apply a Creative Commons or similar license to this, there is no meaningful - or defensible - legal basis to this and they have in effect done little more than sow yet more confusion in this already complex space.

If we are to release large quantities of data onto the Web with the explicit intention that it be used and reused, then a different solution is required.

3. POLARIZING THE OPTIONS

Back in November of 2004 James Boyle published 'A Natural Experiment' in the *Financial Times* [16]. This piece saw him debating the merits of intellectual property rights over data with Thomas Hazlett and Richard Epstein. His primary thrust was that we should be making policy decisions in this area based on empirical data about the economic benefits one way or another. Something all three protagonists agree on.

Much has changed since 2004, not least our understanding of how the web can affect the way we collaborate, share, communicate; it fundamentally affects the way we live. We chat, we blog, we Twitter, we Flickr and we Joost. Content flows from person to person in unprecedented ways and at unprecedented speeds. This changes the nature of the experiment that Boyle talks about.

In Europe we have a right somewhat akin to Copyright, specifically intended to provide protection for aggregations of data; databases. If this European Database Right were working,

"we would expect positive answers to three crucial questions. First, has the European database industry's rate of growth increased since 1996, while the US database industry has languished? [...] Second, are the principal beneficiaries of the database right in Europe producing databases they would not have produced otherwise? [...] Third, [...] is the right promoting innovation and competition rather than stifling it?"

Boyle's first two questions centre around the creation of databases and his third, by his own admission, is difficult to measure. If one of our primary goals for the growth of the Internet is to have a web of data that can be linked and accessed across the globe we may be better served by assessing how companies might make data open rather than closed.

Boyle asks for, and discusses, the empirical evidence of databases being created in the EU and US. The differences in numbers should provide insight into the economic ups and downs as the EU adopted a robust database right in 1996 while the US ruled against such protection in 1991.

Boyle explains that the US Chamber of Commerce oppose the creation of a database right in the US;

"[The US Chamber of Commerce] believe that database providers can adequately protect themselves with contracts, technical means such as passwords, can rely on providing tied services and so on."

And therein lies the rub. Without appropriate protection of intellectual property we have only two extreme positions available: locked down with passwords and other technical means; or wide open and in the public-domain. Polarising the possibilities for data into these two extremes forces the creator of data toward one of two extremes, neither of which are likely to address the nuance of their own circumstances and desires.

With only technical and contractual mechanisms for protecting data, creators of databases can only publish them in situations where the technical barriers can be maintained and contractual obligations can be enforced.

We don't tolerate this with creative works, our photographs, our blog posts and so on. Why would we expect it to make sense for databases? Whether or not it makes sense comes down to whether or not it is beneficial to society. We allow Copyright in order to provide adequate remuneration to be collected by the creator of a work. We allow patents to allow the recovery of development costs for an invention. Which is database right more like?

The patent is a very broad monopoly. If one had a patent on the clock, a mechanical means of measuring the passage of time, nobody else would be able to make clocks without payment of some fee. Copyright on the other hand is much narrower, only allowing protection for the specific design of particular clocks. Database right in the EU is like Copyright. It is a monopoly, but only on that particular aggregation of the data. The underlying facts are still not protected and there is nothing to stop a second entrant from collecting them independently.

Richard Epstein points to this in his contribution to the *Financial Times*' discussion;

"The question is why do databases fall outside [the general principle of copyright], when the costs of compilation are in many cases substantial for the initial party and trivial for anyone who receives judicial blessing to copy the base? In answering this question, it will not do to say, as the Supreme Court said in the well known decision in Feist Publications v. Rural Telephone Service, (1991) that these compilations are not 'original' in the sense that it requires no thought to check the spelling of the entries and to put them all in alphabetical order. But that obvious point should be met with an equally obvious rejoinder. If it requires no thought or intelligence to put the information together, then why not ask the second entrant into the market to go through the same drudge work as the first."

This is exactly what we see happening with Open Street Map. The United Kingdom's national mapping agency, Ordnance Survey, have rights over the map data they have collected. The protection covers the collection of geospatial data that they have created. They are not granted a monopoly in geospatial data.

This leaves a special case of databases, those which are created at low cost as a by-product of normal business. Examples used in Boyle's article are telephone numbers, television schedules and concert times. Boyle gives us the answer directly;

"the [European] court ruled that the mere running of a business which generates data does not count as "substantial investment" enough to trigger the database right."

That a database right may not and should not apply in all cases, and that there is a requirement to restrict anti-competitive practices, does not necessarily extend to the conclusion that a right is not required.

It seems that much of the debate around intellectual property rights has focussed on how they are used to keep things closed. Having suggested earlier that we have only the abilities to keep databases locked away or in contrast open them completely, there is scope for considering - and defining - protections that lie somewhere between these two extremes.

4. EXISTING LICENSES

In response to Thomas Hazlett's contribution to the *Financial Times* debate, Boyle asks;

"How many databases are now created and maintained entirely 'free' and thus escape commercial directories altogether? There are obviously many, both in the scientific and the consumer realm. One can no more omit these from consideration, than one can omit free software from the software market."

This is an important point, and worthy of consideration. Taking one of the most prevalent free software licenses, the Gnu Public License [17], what might that look like for data?

One of the primary functions of the GPL is that it enforces Copyleft – the requirement to license derivative, and even complimentary, works under the same license. That is, any commercial software that makes use of GPL code must, under the terms of the license, also be released under the GPL. The viral nature of this license is possible only because of the legal backing of Copyright legislation.

Without a legally recognised Database right, communities have no mechanism to publish openly and still insist upon this kind of Share-Alike agreement for their data.

Consider the impact of this for situations where you you might use the idea of promiscuous copying to maintain the availability of data. Promiscuous copying relies on two things; lots of copies being made and lots of copies being available. Without the necessary licensing in place there is no mechanism with which to compel those who have copies to make those available. Public Domain means, by definition, no restriction. There is nothing to prevent someone from taking data released into the public domain and locking it away behind a pay wall or similar restrictive mechanism.

Copyleft is just one position along a spectrum where 'locked away' and 'free as a bird' sit at each end. What the web shows us is that other business models form crucial parts of the eco-system. Epstein picks up on the controlling aspect of Boyle's argument:

"They can control their list of subscribers; give them each passwords; charge them based on the amount of the information that is used, or some other agreed-upon formula; and require them not to sell or otherwise transfer the information to third parties without the consent of the data base owner."

Imagine if this were true of Copyright material on the web? It has been, and still is on the occasional site. But mostly copyright owners are starting to see the value of publishing content online and they are underpinning the delivery of that content to consumers with other business models. Without Copyright the types of business that could participate would be reduced.

Epstein goes on to say:

"The contractual solution is surely preferable, because general publication will allow for use by others that may not offend the copyright law, but which will block the possibility of payment for the costly information that is supplied."

And again, the very heart of the matter. If we are to encourage those who have large databases to make them open, to post them on the Semantic Web, we must provide them with models and solutions that are preferable to technical barriers and restrictive contracts. Allowing them to pick their own position on the spectrum seems a necessary part of that. You can see any form of protection in two lights. When Boyle says;

"They make inventors disclose their inventions when they might otherwise have kept them secret."

we say;

"They allow inventors to disclose their inventions when they might otherwise have had to keep them secret."

In the world of creative works, notions espoused by Lawrence Lessig and others over a number of years are becoming increasingly well understood. A Creative Commons license, for example, is recognized as giving the holder of rights an ability to prospectively grant certain permissions rather than limit use of their work by expecting all comers to request these permissions, again and again. Those rights are not cast aside, removing all opportunities to protect your work, your name, or your potential revenue stream. Rather, you are provided with a means to explicitly declare that your work may be used and reused by others in certain ways without their needing to request permission. Any other use is not forbidden; those uses must simply be negotiated in the 'normal' way... a normal way that also applied to those uses covered by Creative Commons licenses before the advent of those licenses.

Creative Commons licenses are an extension of copyright law, as enshrined in the legal frameworks of various jurisdictions internationally. As such, it doesn't really work terribly well for a

lot of (scientific, business, whatever) data... but the absence of anything better has led people to apply Creative Commons licenses of various types on data that they wish to share. It will be interesting to see what happens, the first time someone seeks redress in a court, citing the Creative Commons license that they selected as an appropriate protection against abuses of their data.

5. A LICENSE FOR OPEN DATA

Back in 2006, Talis released a first public attempt at an open data license, the Talis Community License [17], and began to use it for some early submissions to the Talis Platform [18]. In building a Platform, we understood from the outset the importance of recognizing - and celebrating - the rights of those contributing their data to the shared pool. The Talis Community License allowed us to do that.

Not long after, Tim O'Reilly wrote;

"One day soon, tomorrow's Richard Stallman will wake up and realize that all the software distributed in the world is free and open source, but that he still has no control to improve or change the computer tools that he relies on every day. They are services backed by collective databases too large (and controlled by their service providers) to be easily modified. Even data portability initiatives such as those starting today merely scratch the surface, because taking your own data out of the pool may let you move it somewhere else, but much of its value depends on its original context, now lost." [19]

At Talis, we have an interest in seeing large bodies of structured data available for use. Through the Talis Platform, we offer one means whereby such data may be stored, used, aggregated and mined, although we clearly recognize that similar data may very well also be required in diverse contexts.

Recognizing that contributors of such data need to be reassured as to the uses to which we - and others - may put their hard work, we spent some time drafting what was then called the Talis Community License. This draft license is based upon protections enshrined in European Law, and has been used 'in anger' for a while to cover contributions of millions of records to one particular application on the Talis Platform.

Despite interest in open (or 'linked') data, licenses to provide protection (and, of course, to explicitly encourage reuse) are few and far between. Amongst zealous early adopters, there does seem to be a tendency to either (mis)use a Creative Commons license, to say nothing whatsoever, or to cast their data into the public domain. None of these strategies are fit for application to business-critical data.

Building upon our original work on the TCL, we provided funding to lawyers Jordan Hatcher and Charlotte Waelde [10]. They were tasked with validating the principles behind the original license, developing an effective expression of those principles that could be applied beyond the database-aware shores of Europe, and working with us to identify a suitable home in which this new license could be hosted, nurtured, and carried forward for the benefit of stakeholders far outside Talis.

The result of this effort was the Open Data Commons Public Domain Dedication and License [11], itself a fusion of ideas from the Talis Community License, an initial phase of redrafting from Hatcher and Waelde, and a focussed piece of activity to align with a related framework developed within the Science Commons project of Creative Commons at the same time.

The current iteration of the license asks licensors to waive various local protections in order to create a level playing field upon which a set of 'community norms' may be documented in order to define a set of shared expectations as to the ways in which the data may subsequently be reused. The first of those community norms is defined on the Open Data Commons site [12], and all concerned expect compatible sets of norms to be created elsewhere in time.

The Public Domain Dedication and License is now available for use, following a period of consultation. At the time of writing, all those concerned in getting to this stage are engaged in the process of placing the wider Open Data Commons initiative itself on a sound footing, creating a safe place in which this license and those to follow it may be maintained and evolved. The Open Knowledge Foundation (OKF) in Cambridge, UK, is to lead by providing that neutral new home, and funders, drafters and other interested parties are united in supporting this move to a sound and sustainable footing [20].

6. CONCLUSIONS AND OUTLOOK

There is a lot still to do, but the interdisciplinary collaboration we're already seeing with respect to permissive licensing of data for the web means that we can all begin to move forward in lowering the walls of our silos, releasing data to play its part in the Data Web. All of us invest heavily in collecting and curating data, which is traditionally locked away and left to atrophy, failing to achieve anything like its true potential. Appropriately released and sensibly licensed, data held by every one of us can contribute hugely to the promise of the Semantic Web. Here, the whole really is far greater than the sum of its parts.

The current license is available for use. It provides us with the capability to build upon the efforts of those philanthropic contributors to the existing Linking Open Data project [21], and to take the linked data proposition to that broader market of data curators who need more persuasion and reassurance. The opportunity is immense, as is the benefit to the Semantic Web itself.

7. REFERENCES

[1] Building a Semantic Web in which our Data can Participate, WWW2007 Panel Session (May 2007). http://www2007.org/panel7.php

[2] Miller, P. 2007 Presentations from WWW2007 Open Data panel now online. In Nodalities weblog. http://blogs.talis.com/nodalities/2007/05/presentations_from_www2007_ope.php

[3] Suber, P. 2007 Peter Murray-Rust on open access and open data. In Open Access News. http://www.earlham.edu/~peters/fos/2007/05/peter-murray-rust-on-open-access-and.html

[4] Miller, P. 2007 Jamie Taylor Talks with Talis about Metaweb and Freebase. In Nodalities weblog. http://blogs.talis.com/nodalities/2007/05/jamie_taylor_talks_with_talis.php

[5] Styles, R. 2007 Open Data Licensing, an unnatural thought. In Nodalities weblog. http://blogs.talis.com/nodalities/2007/07/open_data_licensing_an_unnatur.php

[6] Miller, P. 2007 Licensing open data - Creative Commons and Talis have something to say. In Nodalities weblog. http://blogs.talis.com/nodalities/2007/12/licensing_open_data_creative_c.php

[7] Creative Commons. http://creativecommons.org/

[8] Wilbanks, J. 2007 Announcing the Protocol for Implementing Open Access Data. In Science Commons weblog. http://sciencecommons.org/weblog/archives/2007/12/16/announcing-protocol-for-oa-data/

[9] Steuer, E. 2007 Creative Commons launches CC0 and CC+ Programs. Creative Commons media release. http://creativecommons.org/press-releases/entry/7919

[10] Miller, P. 2007 Seeking a license for open data. In Nodalities weblog. http://blogs.talis.com/nodalities/2007/09/seeking_a_licence_for_open_dat.php

[11] ODC Public Domain Dedication and License. http://www.opendatacommons.org/odc-public-domain-dedication-and-licence/

[12] ODC Community Norms. http://www.opendatacommons.org/odc-community-norms/

[13] Open Data Commons. http://www.opendatacommons.org/

[14] Linked Data definition from Wikipedia. http://en.wikipedia.org/wiki/Linked_Data

[15] OpenStreetMap. http://www.openstreetmap.org/

[16] Boyle, J. 2004 James Boyle: a natural experiment. *Financial Times* 22 November. http://www.ft.com/cms/s/2/4cd4941e-3cab-11d9-bb7b-00000e2511c8.html

[17] The Talis Community License. http://www.talis.com/tdn/tcl/

[18] The Talis Platform. http://www.talis.com/platform/

[19] O'Reilly, T. 2006 Four Big Ideas About Open Source. In O'Reilly Radar weblog. http://radar.oreilly.com/archives/2006/07/four-big-ideas-about-open-sour.html

[20] Open Knowledge Foundation. http://www.okfn.org/

[21] Linked Data Project. http://www.linkeddata.org/