There's No Money in Linked Data

Prateek Jain[1], Pascal Hitzler[2], Krzysztof Janowicz[3] and Chitra Venkatramani[1]

[1] IBM T.J. Watson Research Center, Yorktown, NY, USA
[2] Kno.e.sis Center, Wright State University, Dayton, OH, USA
[3] Geography Department, University of California, Santa Barbara, CA, USA

Abstract. Linked Data (LD) has been an active research area for more than 6 years and many aspects about publishing, retrieving, linking, and cleaning Linked Data have been investigated. There seems to be a broad and general agreement that *in principle* LD datasets can be very useful for solving a wide variety of problems ranging from practical industrial analytics to highly specific research problems. Having these notions in mind, we started exploring the use of notable LD datasets such as DBpedia, Freebase, Geonames and others for a commercial application. However, it turns out that using these datasets in realistic settings is not always easy. Surprisingly, in many cases the underlying issues are not technical but legal barriers erected by the LD data publishers. In this paper we argue that these barriers are often not justified, detrimental to both data publishers and users, and are often built without much consideration of their consequences.

1 Introduction

Linked Data is a research catalyst. And it has been so since it's inception over half a decade ago. It has played this role, in fact, despite many known deficiencies in design and practical realization [2–10]. In part, perhaps, it is such a catalyst *because* of the naivity with which Linked Data publishing has been conducted since—because it gave the research community ample reason for investigations how to recover from it.

Linked Data is also without doubt *useful*. Just think of Evan Sandhaus' keynote at ISWC 2010 on the use of RDF and Linked Data at nytimes.com[4] or its use for IBM's Watson system [11]. But where's the steamroller of commercial applications—those that are producing hard cash? Wouldn't we expect significant commercial uptake by now?

In this paper we discuss a show stopper for the commercialization of Linked Data. By no means do we want to claim that it is the key roadblock—but it seems to be one, and a significant one at least in some contexts. We also do not want to claim that we are the first ones having this insight. In fact the general observation may be an old hat for some, but we haven't seen it discussed anywhere in a structured manner so far. The issue has recently started gaining prominence and a very recent workshop *Open Data on the Web*[5] discussed it in some amount of detail. We think that keeping in theme with the workshop, it is an important issue and must be shared with the community, and must be discussed by the community if we hope to overcome this bottleneck.

[4] http://iswc2010.semanticweb.org/node/110
[5] http://www.w3.org/2013/04/odw/

Number of datasets	License
139	No license specified.
49	CC-BY.
31	CC-BY-SA
24	CC0
15	ODC-PDDL

Table 1. Top 5 licenses utilized for LOD datasets

The issue we discuss in this paper is in fact of legal nature. It concerns licensing practices regarding Linked Datasets. We are no lawyers, and as such cannot give legal advice, and understand intricate legal issues only to a certain extent. Nevertheless, we feel qualified to talk about the topic since we are users or potential users of Linked Datasets and thus have to concern ourselves with these legal issues to the best of our abilities, just like most users of Linked Data.

So in this paper, we will survey common licensing practices for Linked Datasets. Doing so, we will show that the current licensing situation is mostly prohibitive for commercial use of these datasets. This may be because such use is explicitly disallowed by the licenses, but in many cases the situation is much more tricky, e.g. because of missing licensing information, or because it would be necessary to trace the *sources* of a linked dataset and also take the licenses of these sources into account.

We will first discuss the licensing issues in Section 2, then talk about a way forward in Section 3, before we conclude in Section 4.

2 What are our complaints

Contrary to the name Linked 'Open' Data (LOD) we didn't find a whole lot in Linked Data which was 'open' and could foster a revolution like it was propounded in Tim Berners-Lee TED Talk.[6] We actually found the datasets to be interlinked silos like commercial websites as identified in [12]. You can do anything you want as long it is a proof of concept or a research paper. Try getting into the realm of actual commercial usage and you are in for a rude awakening.

You may be wondering how this is possible with the current LOD datasets, as pretty much all of them use open licenses, usually from the Creative Commons family of licenses. While you would be right about the statistics about the license family utilized for the datasets (as we show later), like all things in Computer Science, the devil is in the details. More specifically, the issues are as follows.

2.1 Highly restrictive licenses

We did a simple analysis of the different licenses utilized by the LOD datasets and the findings are reported in table 1.

Now let us explain why some of these licenses are problematic for using them in commercial settings:

[6] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

1. No license specified :- While, according to conventional wisdom, no news is good news, unfortunately the same is not true for dataset licenses. It makes it difficult or impossible for commercial organizations to use the datasets as they can be liable for damages resulting from any potential lawsuits or liabilities. More specifically, it makes it hard to gauge under what conditions it is ok to use them and for what purposes. It leaves a lot of questions unanswered such as: Are the analytics supposed to be released in the same condition? Is my organization responsible for answering to the original source from which the dataset was created? Can I bundle the datasets with my software? Besides these questions, some of these datasets have been created by one person, and if the person is not responding, there is no other way to obtain explicit permission to use the dataset. Thus, in the end a potentially useful dataset is on the pretty picture, but without any purpose.
2. CC-BY :- Used by roughly 49 datasets including Freebase, it is the second most popular license utilized by LD datasets, and it states "You are free to make commercial use of the work." A glance at this line will probably make you think, it is all clear and let's go ahead and use the datasets for commercial applications. But that is only the partial truth. These datasets are structured representations of underlying information gleaned from multiple sources. These entities include commercial entities like Hotels.com used by Geonames. Hotels.com clearly states in their terms of use[7] under PROHIBITED ACTIVITIES: '1. use this Website or its contents for any commercial purpose.' Whereas Geonames, by using CC-BY, is giving permission to use the dataset for commercial activities. Notice the inherent contradiction and confusion for anyone wanting to use these datasets. Geonames says they have permission from Hotels.com for using the information [8], but does the permissions cover commercial usage of the information? Which set of terms and conditions should be trusted? The source or the derivative?

   This kind of confusion has lead to lawsuits in the past against the companies using the data covered under CC-BY. For more details, please check the FAQ on the Creative Commons website.[9]
3. CC-BY-SA :- Used by roughly 31 datasets on LOD cloud, including some major ones such as DBpedia, it is one of the most prohibitive licenses regarding commercial utilization of the datasets. Before the obvious is asked "How did Watson manage to use it?" let us answer it: Watson used DBpedia in a research setting for Jeopardy. It needs to be seen how they will navigate the restrictions imposed by CC-BY-SA on commercial work. As from the text of the license, "If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one."

   This makes it extremely difficult for commercial companies to utilize these datasets as they will be forced to release their offerings under the same license. We can only guess that the underlying motivation behind this condition is to promote open source movement. Unfortunately, companies which want to utilize these datasets will usually not agree to these conditions as it may invalidate their business model.

[7] http://www.hotels.com/customer_care/terms_conditions.html

[8] https://groups.google.com/d/msg/geonames/ANfLA06MZ7E/vuw6pgjTas0J

[9] https://creativecommons.org/weblog/entry/7680

Surely all these issues can be resolved by backdoor negotiations and contracts, but imagine a team of young entrepreneurs with lots of energy, great ideas and visions. Unfortunately, none of the three qualities can pay the legal fees required to get a legal permission to use the datasets.

4. CC0 :- CC0 is one of the friendliest licenses for commercial usage by allowing the content creator to waive any copyright and related rights. In fact it has been utilized by 24 datasets which are part of LOD. Some of these datasets such as TWC: Linking Open Government Data have been created using data in the open domain, i.e., US Government datasets restricting their commercial usage in any way will not be appropriate. Unfortunately, in other cases, data publishers may shy away from CCO by asking *Why should we open it up?* Some of these issues have been raised recently in [13] regarding data published by the German National Library.

2.2 How we stumbled upon these issues

One of our industrial applications involving text analytics required dictionaries related to various entities such as names of people, cities, countries and populations. Having used LOD in the past for various research related tasks, we were very excited about the prospects of using LOD and applying the various datasets for an actual commercial application. We picked up *DBpedia* and started identifying what we can gain from the dataset.

Needless to say, there was a plethora of information which we could have used for our application scenario. Using DBpedia we were able to create an extensive list for a number of requirements such as list of names of people (using type information) and assign gender to them as male or female. We were able to do similar things for names of places and synonyms for various organizations.

The benefits we could have gained from this would have been significant, and would have saved time for us, as is one of the rationales behind LOD. However, we were required to get approval for external datasets utilized for any commercial applications from the legal team to prevent violations of any intellectual property rights. After scrutinizing the license conditions of DBpedia, Freebase and Geonames, the legal team advised us about the issues mentioned above regarding CC-BY-SA, GNU Free Documentation License (GFDL) and issues regarding the use of commercial datasets by Geonames. Freebase allowed usage for a good chunk of attributes under CC-BY, but certain sections such as long descriptions and some images are licensed under GFDL and thus are tricky to utilize.

Eventually, we decided to not utilize any of the LOD datasets and try to make do with data provided by the US Government thru data.gov. At hindsight, it might have been possible to use datasets licensed under CC-BY without any issues related to the origin of data (such as in the case of Geonames). However, the amount of time required to do so and any possible impact of such matters on the project deadline made us settle and work on familiar territory.

2.3 Snag raw data now!

Referring back to the TED talk mentioned before and Tim Berners-Lee famous *raw data now!* slide, it is worth noting that he actually *asks* for raw data now. Why is this

important for our discussion on the commercial application of Linked Data in the future? For the initial creation of the Linked Data cloud, it was necessary to get some first key datasets out there. However, as a community, we are still in a phase where data from others, often not even involved providers, are taken and transferred to Linked Data. This is done with the argument that the used licenses allow for such actions. While this is true, it disconnects the original providers and domain experts that created the data and often also curate them from the Linked Data version. While we do not discuss the implications of this kind of data snagging on Linked Data quality here, this lack of connection has consequences for the licensing problems detailed above. To give a concrete example, raw data published under the CC-BY-SA license enforces that the derived Linked Data is published using the same license. As argued before, this and other licenses are not optimal for commercial reuse. If we would actively involve the original data creators in the triplification of their data, we could alert them about the resulting consequences for the global and interconnected reuse of their data and possible convince them to change their licensing strategy. Similarly, the snagging approach may not create sustainable and up-to-date LOD, e.g., if the raw data providers change their licenses by making them more restrictive.

3 A Way Forward

We are not lawyers, but rather are producers and consumers of this data. However, we believe that a few simple steps can already help significantly in alleviating these issues.

1. It's The License, stupid :- A very simple solution is to make it mandatory or give some details regarding the terms and conditions under which the data can be utilized when publishers submit their datasets for inclusion in LD listings. These details can be as simple as just saying whether commercial use is allowed or not, and under what circumstances. Whether these licenses or conditions are restrictive or not is an issue by itself which we will come to in a little bit. This condition will make sure that datasets in available listings are not just nodes in a cloud, but datasets which can be utilized for any practical and useful purpose such as research, commercial applications or non-profit purposes.
2. Changes in Grant Conditions :- Many of the datasets which are part of the LD cloud have been published by universities using grant money provided by public agencies such as NSF, NIH or EU bodies. These agencies are funded by tax payer money including taxes imposed on commercial entities. Research funded by these agencies in the past has lead to some important technological innovations like barcodes, cloud computing and DNA Fingerprinting [14]. These innovations were originally created in universities and then picked up by commercial organizations and now are ubiquitous. If we want something similar to happen with the use of structured data, we have to make it easier for them to be used, especially if the research has been funded by public money. By no means we are suggesting that universities should not be allowed to commercialize their research. A simple solution like the patent system, which gives rights to commercialize and profit for certain periods of time and makes it open after that period can be a possible solution to this issue.

3. Organized Community Discussion :- The LOD community within W3C can get more actively involved around these issues to increase the LD utilization. The community has spent a good chunk of the past 6-7 years in expanding the number of datasets available on the cloud, developing tools for interlinking data and efficient retrieval and storage of data. We believe issues regarding data access, utilization and license conditions should be more actively discussed to find a common middle ground for both data publishers and consumers. Recent discussions around these issues at venues like Open Data on the Web[10] are a welcome step in this direction. There is a greater need for forums like these to encourage the exchange of ideas in the community.
4. Mark out Commercially Usable Datasets :- In the 400+ datasets available on the LD cloud, it is quite tricky to identify what can and cannot be used easily for commercial purposes. Helpful would be, for example, an easy way to clearly mark these datasets which can be used commercially without any constraints by a special tag to create a cluster on CKAN. This will allow the community to identify how many of the publicly available datasets can be used commercially and what can be done to increase the commercialization potential.
5. Greater Discussion regarding What Licenses Mean :- Creative Commons and the GNU family of licenses provide an easy framework to use and apply licenses to your work. The flip side of this ease is the lack of any requirement to understand what the terms and conditions of these licenses entail. This is somewhat similar to accepting the end user agreements for softwares without reading the lines and between the lines. This licensing framework might also be leading to a situation where developers and data publishers are utilizing them without realizing the pros and cons of these conditions or if anything different can be done. It is extremely important that greater discussion and dissemination happens around these issues.

4 Conclusions

We find that currently licensing practice for linked datasets is a severe bottleneck for commercialization. We believe that this issue can be addressed successfully, however a community discussion is required. In the wake of such discussions, community measures and organizational measures by funding agencies would promise to alleviate the issue.

Acknowledgements. Pascal Hitzler acknowledges support by the National Science Foundation under award 1017225 "III: Small: TROn – Tractable Reasoning with Ontologies." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or International Business Machine(IBM).

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22

[10] http://www.w3.org/2013/04/odw/

2. Auer, S., Lehmann, J.: Creating knowledge out of interlinked data. Semantic Web **1** (2010) 97–104
3. Ding, L., Shinavier, J., Finin, T., McGuinness, D.L.: owl:sameAs and Linked Data: An empirical study. In: Proceedings of the Second Web Science Conference, Raleigh, NC, April 2010. (2010)
4. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameAs isn't the same: An analysis of identity in linked data. In Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B., eds.: The Semantic Web – ISWC 2010 – 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. Volume 6496 of Lecture Notes in Computer Science., Springer (2010) 305–320
5. Hitzler, P.: What's wrong with Linked Data? http://blog.semantic-web.at/2012/08/09/whats-wrong-with-linked-data/ (2013)
6. Hitzler, P., van Harmelen, F.: A reasonable semantic web. Semantic Web **1**(1–2) (2010) 39–44
7. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: 3rd International Workshop on Linked Data on the Web (LDOW2010) at WWW2010, Raleigh, USA, April 2010. (2010) Available from http://events.linkeddata.org/ldow2010/.
8. Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P.: Linked Data is Merely More Data. In Brickley, D., Chaudhri, V.K., Halpin, H., McGuinness, D., eds.: Linked Data Meets Artificial Intelligence, AAAI Press, Menlo Park, CA (2010) 82–86
9. Janowicz, K.: Place and location on the web of linked data. http://stko.geog.ucsb.edu/location_linked_data (2013)
10. Polleres, A., Hogan, A., Harth, A., Decker, S.: Can we ever catch up with the Web? Semantic Web **1** (2010) 53–59
11. Kalyanpur, A., Murdock, J.W., Fan, J., Welty, C.A.: Leveraging community-built knowledge for type coercion in question answering. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N.F., Blomqvist, E., eds.: The Semantic Web – ISWC 2011 – 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part II. Volume 7032 of Lecture Notes in Computer Science., Springer (2011) 144–156
12. Berners-Lee, T.: Long live the web: A call for continued open standards and neutrality. Scientific American (December 2010)
13. Svensson, L.G.: Licensing library and authority data under CC0: The DNB experience. Available at http://www.w3.org/2013/04/odw/odw13_submission_57.pdf (2013)
14. National Science Foundation: NSF Sensational 60. Available at http://www.nsf.gov/about/history/sensational60.pdf (2010)