A Research Agenda for Linked Closed Data
(vision paper)

Marcus Cobden, Jennifer Black, Nicholas Gibbins, Les Carr, and Nigel Shadbolt

{mc08r,jlb08r,nmg,lac,nrs}@ecs.soton.ac.uk
University of Southampton, UK

Abstract. While it is preferable that Linked Data is published without access or licence restrictions, there will always remain certain datasets which, perhaps due to financial considerations, cannot be published as Linked Open Data. If these valuable datasets do join the Web of Linked Data, it will be as Linked Closed Data – Linked Data with access and license restrictions. In this paper, we outline a research agenda for Linked Closed Data that considers the effects that access and license restrictions may have on the Web of Linked Data. If implemented poorly, access restrictions have the potential to break URI resolvability, but even when implemented well, we can expect them to affect dataset selection processes and inter-dataset link creation rates. Additionally, there remains the technical challenge of developing and standardising access restriction and automated payment techniques for the Web of Linked Data.

1 Introduction

To date, most research from the Linked Data community has focussed on Linked Open Data and failed to consider its logical counterpart *Linked Closed Data*. By Linked Closed Data, we refer to datasets which adhere to the principles of Linked Data publishing, but for which access to and use of the data is subject to legal or technical restrictions which go beyond attribution and share-alike obligations.

Linked Closed Data may seem at odds with the aims of the Linked Data community, however, this is not necessarily the case. Undoubtedly it would be a shame if existing 'open' datasets changed to a 'closed' publishing model, but it would still be preferable to the datasets becoming completely unavailable due to financial pressures on their publishers.

High quality data has value. If we wish to see more high quality datasets published as Linked Data, we must provide incentives for publishers to do so. Profit is arguably the most significant publishing incentive, and so, revenue models, and how they interact with Linked Data publishing patterns, are of great importance to the future of the Web of Linked Data.

While publishing 'open' datasets is to be commended, we anticipate that few companies will be prepared, or able, to make their data available without some access or licence restrictions. Ordnance Survey, Great Britain's national mapping agency, is in this position; they are required to operate as a self-funded organisation through the commercial sale of mapping data. Offering all their data for free would undermine their primary source of income.

The Times and The New York Times newspapers both used to offer free access to articles on their Web site, recently, however, they both adopted a subscription-based access model for their Web content. Crucially, unlike The Times, The New York Times allows non-subscribers free access to view up to 20 articles per calendar month. Despite the difference in markets, the introduction of paid access may be comparable to the addition of restrictions to Linked Data publishing.

It is imperative that we fully understand the implications of Linked Closed Data so that we can be sure that desirable properties of the Web of Data will not be lost in the face of access and license restrictions. Against this background, in this paper we identify the key research challenges posed by Linked Closed Data.

2 Access and Licensing

While we describe datasets as 'open' and 'closed', reality is less clear-cut; in practice we observe a spectrum of 'openness', which varies with the access restrictions a dataset is published with (who is permitted to 'see' the data), and the licence under which access is granted (what they are permitted to do with the data).

2.1 Access restrictions

Access to a dataset may be: completely open, restricted only by the resources of the dataset host; restricted, open to users who meet specific access criteria; or private, open only to its owner. If badly realised, access restrictions have the potential to undermine the resolvability of Linked Data URIs and to dis-incentivise the creation of inter-dataset links. The resolvability of a URI would be undermined if, due to access restrictions, it is no longer possible to resolve that URI to a description document. Access restrictions might also remove any useful information from a URI description document. Creating links to a restricted dataset is less worthwhile if URIs within it cannot be resolved to a useful description.

2.2 Dataset licences

Dataset licences range from public domain dedications (where all intellectual property rights are waived)[9], to permissive licences (which may impose anything from attribution, to more weighty obligations such as copyleft) to restrictive licences (which specify permitted uses of a dataset). Currently, 85% of the datasets in the Linked Data Cloud do not declare any license [2].

We are unlikely to see a restrictive license on a dataset without also encountering access restrictions; if access is open and practically anonymous, license breaches would be difficult to detect and punish. Since it is only worthwhile imposing license restrictions on valuable data, we expect that the publisher would also take the step of imposing access restrictions to protect its investments.

3 Business models

The 'openness' of a dataset is generally determined by the business model of the publisher. A range of business models have been proposed for Linked Data

publishers, including subsidised publishing, subscription or micropayment-based access, sponsorship/advertising funded publishing, and loss-leader models (to drive interest or sales in other products, or to shape markets) [3].

Broadly, we can classify these business models into those which offer only free access products, only paid access products, or some combination of both. The revenue models of the individual products these business models are build from can also be split between free and paid access. We can then further categorise them by how their costs are recovered:

Advertising supported Costs are covered by revenue from advertising within the content. We consider this to cover both per-view advertising payments, and per-sale commission through affiliate links.

Loss-leader Costs are written off as an investment. Loss-leader models may attempt to drive interest in other products, or perhaps to shape markets in the hope of future sales.

Subsidised Costs are covered by some form of sponsorship or subsidy. Common in public sector undertakings.

Free access models should, in theory, employ at least one of these strategies, though our list is not exhaustive. Equally, paid access models might employ any combination of these to complement revenue from access payments. Currently, all Linked Data publishers operate under loss-leader or subsidy-based revenue models – none have adopted paid access or advertising supported revenue models.

In the last year, Ordnance Survey have begun to offer free access to some of their data. These free datasets include postcode location data, electoral and administrative boundaries, and gazetteer (at 1:50000 scale), though only some of these are published as Linked Data. Currently, as they offer no paid access Linked Data, this is likely to be a loss-leader exercise, perhaps to demonstrate what is possible with their premium subscriptions, or to shape the market in anticipation of a paid offering. Coupling a premium Linked Data product with a free version may be an effective way of mitigating the effects of access restrictions on phenomena such as link creation, providing they use the same URIs, and URI resolvability is maintained.

4 Research Challenges

The addition of access restrictions to Linked Data publishing systems and the emergence of datasets published under proprietary licenses will be an inevitable fact of commercial Linked Data. These changes will alter our expectations of URI resolvability and change the ways in which we use Linked Data.

We have identified six key challenges which are important in a Linked Data ecosystem where not all data is free – they are: i) building dataset reputation, ii) developing access, authentication and payment protocols, iii) fostering links between datasets, iv) managing confidential data, v) respecting license terms, and vi) validating business models. We elaborate further on these in the remainder of this section.

4.1 Building reputation

Access restrictions make it all the more important for a dataset to have a good reputation. Without access, prospective users will be unable to evaluate whether the dataset meets their needs before they commit to a purchase. Reputation and word of mouth are a common means through which users judge whether a product meets their requirements when considering a purchase.

Unfortunately there are few incentives to risk purchasing access to a dataset of unknown quality, so new and untried datasets are unlikely to create a reputation for themselves. To combat this bootstrapping problem, publishers will need to improve their reputation through other means; perhaps by (temporarily) removing access restrictions, or by seeking endorsements from trusted authorities. Freemium revenue models [1] and free time- or extent-limited licences are a common means of providing access to prospective customers. Future research is needed to explore this area in more detail, and to identify other dataset reputation bootstrapping techniques.

4.2 Access, authentication and payment

The issue of restricting access to closed datasets poses clear technical challenges. HTTP content negotiation and redirection allows different documents to be served in response to a URI resolution attempt. Similar methods are needed for restricted-access datasets in order to field requests between free and paid content. Additionally, a new vocabulary is needed with which one can declare the presence of related documents which require payment to access, otherwise premium content may not be discoverable.

Authentication is a requirement of any access restriction techniques. The foundations for authentication are already being laid; the W3C WebID incubator group is standardising TLS client certificate based authentication [4].

Finally, in order for automated access to premium content to succeed beyond isolated individual publishers, we must standardise the means by which we indicate that payment is needed, and the methods by which payment can be made. The HTTP '402 Payment Required' response code has long been reserved, but no standards have yet specified how it should be used [6].

4.3 Fostering links

Restricting access to a dataset may also negatively impact how likely external sources are to create links to that dataset. Inter-dataset links are said to add value to datasets, and incoming links provide a form of advertisement and endorsement for the target dataset.

Publishers of Linked Closed Data will need to take into account the effect that their particular access and licensing schemes may have on incoming link creation rates. Datasets with schemes which permit some degree of free access may maintain a higher rate of link creation than those under more restrictive schemes. The different paid access models adopted by The Times and The New York Times may allow us to examine how access restriction affects link creation on the Web, although whether this translates to a Semantic Web context remains an open research question. Further research is needed to examine this behaviour in detail, and to explore other means of encouraging link creation.

4.4 Managing confidentiality

It is not inconceivable that linked datasets might contain sensitive information. While authentication and authorisation schemes can be used to limit access to restricted information, the challenges posed by confidential data go beyond access, particularly in systems containing mixed confidentiality-level data.

Care must be taken that sensitive information cannot be inferred from non-sensitive data. For example, unique identification number ranges with unexplained gaps may suggest hidden information. Further, a Linked Data system, which adheres to current best practices, might inadvertently admit the existence of a sensitive URI by responding with an 'HTTP 403: Forbidden' response code instead of an 'HTTP 404: Not Found'. While this alone means little, if the URI contains embedded information, or can be correlated with other available data, it may constitute a breach of confidentiality.

Examining the issues in detail, and publishing amended best practices for publishing Linked Data in the presence of confidential information, remains an opportunity for future research.

4.5 Respecting licenses

As we mentioned in Section 2.2, there are a wide range of common licences which datasets might be licensed under. Standardised licences, such as the GNU Public Licences (GPL), Creative Commons (CC) and Open Database Licences (ODBL), often offer variants with additional restrictions such as attribution, non-commercial use, and copyleft sharing requirements. Some of these licences, such as the GPL and CC, may eventually fall out of favour for Linked Data as they were not designed to be used in this context [9]. Governments often have their own licenses for data, for example the United Kingdom's Open Government Licence (OGL).

Dataset licenses present two main challenges: tracking the licences under which data was received, and respecting the license conditions. In order to be able to honour license restrictions, Linked Data systems need to maintain appropriate provenance records of license conditions. Provenance is an active area of research for Linked Data [8], and we expect that once provenance-aware systems reach maturity much of this will be automated.

Ensuring adherence to dataset licenses is a more difficult research challenge. Interpreting licenses confidently and accurately requires legal training. Ideally standard licenses would include an approved machine-understandable description of the license, describing how a dataset might be used. Existing work has explored the potential of this approach [7]; however, further work is needed to apply this to common license conditions.

4.6 Validating business models

While many business models have been proposed for the Semantic Web and Linked Data publishers [3], it remains to be seen how many of them will prove viable. Some, such as advertising-supported Linked Data publishing, have been

called into question somewhat [5]; however, ultimately, the success of any business model will depend upon the market conditions in which is employed. Nonetheless, until there exist successful businesses built around these models, we must consider them unsolved research challenges.

5 Conclusions

In this paper, we introduced the notion of Linked Closed Data: Semantic Web datasets which are published in accordance with Linked Data principles, but which include access and licence restrictions. While we described Linked Closed Data as the logical counterpart to Linked Open Data, license restrictions are already common among 'Open' datasets (although they usually only require attribution as the source of the data); the addition of access restrictions is of greater significance. We argued that Linked Closed Data is likely to be the form of Linked Data publishing ultimately adopted by commercial Linked Data publishers when offering premium, paid access products, as they are likely to require some form of access restriction.

Finally, we identified six research challenges which are of new, or increased, significance when considering the effect of access and license restrictions on the Web of Data. These are: building dataset reputation despite access restrictions; developing and standardising access, authentication and payment protocols; fostering the creation of links to access restricted datasets; managing confidential data; tracking data licenses and ensuring license adherence; and the validation of Linked Data business models.

References

1. Anderson, C.: Free: The Future of a Radical Price: The Economics of Abundance and Why Zero Pricing Is Changing the Face of Business. Random House Books (Aug 2009)
2. Bizer, C., Jentzsch, A., Cyganiak, R.: State of the LOD Cloud (Aug 2011), http://www4.wiwiss.fu-berlin.de/lodcloud/state/
3. Brinker, S.: Business models for linked data and web 3.0 (Mar 2010), http://www.chiefmartec.com/2010/03/business-models-for-linked-data-and-web-30.html
4. Corlosquet, S., Sporny, M., Inkster, T., Story, H., Harbulot, B., Bachmann-Gmür, R.: WebID 1.0 - Web Identification and Discovery (Draft) (Feb 2011), http://www.w3.org/2005/Incubator/webid/spec/
5. Dodds, L.: Thoughts on Linked Data Business Models (Jan 2010), http://www.ldodds.com/blog/2010/01/thoughts-on-linked-data-business-models/
6. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext Transfer Protocol – HTTP/1.1. RFC 2616 (Draft Standard) (Jun 1999), updated by RFC 2817
7. Hanson, C., Kagal, L., Berners-Lee, T., Sussman, G., Weitzner, D.: Data-purpose algebra: Modeling data usage policies. In: POLICY '07. pp. 173 –177 (Jun 2007)
8. Hartig, O.: Provenance information in the web of data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009) (Jan 2009)
9. Miller, P., Styles, R., Heath, T.: Open data commons, a license for open data. In: Proceedings of the WWW2008 Workshop on Linked Data on the Web (LDOW) (2008)