Exploring the Dynamics of Linked Data

Tobias Käfer[1], Ahmed Abdelrahman[2], Jürgen Umbrich[2], Patrick O'Byrne[2], and Aidan Hogan[2]

[1] Institute AIFB, Karlsruhe Institute of Technology, Germany
[2] Digital Enterprise Research Institute, National University of Ireland, Galway

Abstract. Little is known about the dynamics of Linked Data, primarily because there have been few, if any, suitable collections of data made available for analysis of how Linked Data documents evolve over time. We aim to address this issue. We propose the Dynamic Linked Data Observatory, which provides the community with such a collection, monitoring a fixed set of Linked Data documents at weekly intervals. We have now collected eight months of raw data comprising weekly snapshots of eighty thousand Linked Data documents. Having published results characterising the high-level dynamics of Linked Data, we now wish to disseminate results: we wish to investigate how results from our experiment might benefit the community and what online services and statistics (relating to Linked Data dynamics) would be most useful for us to provide.

Summary

The Web of (Linked) Data is dynamic. Knowledge about Linked Data dynamics is important for a wide range of applications and can help to answer a wide variety of practical questions, including (but far from limited to):

For warehouses: Which remote datasets need to be updated most frequently? Which are static and do not need to be refreshed? Are the updates mostly additions or mostly deletions? How often do new documents appear?

For "live query" engines: Which documents are static and can be cached? Which query patterns involve dynamic patterns (e.g., are `foaf:knows` triples more dynamic than `foaf:name`)? For link traversal methods, how often do links change between documents?

For reasoning engines: Do schema data change more or less often than instance data? Which ontologies are the most dynamic? What kinds of semantics change? Are changes "monotonic" with respect to entailments? Which data are static and subject to materialisation? Which are dynamic and subject to query-rewriting?

For publishers: How often do target documents go offline? How often should deadlinks be checked and pruned? How often should certain link-types (e.g., `owl:sameAs`) be revised for updated remote content? How often do the semantics of vocabularies change?

Few works have looked at the nature of Linked Data dynamics or have tried to answer the types of questions posed above; this is probably due to to a lack

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7955, pp. 302–303, 2013.

of suitable collections tracking changes in Linked Data documents over time. Thus, we proposed the Dynamic Linked Data Observatory (DyLDO) [2,1] for monitoring weekly changes in 86,696 Linked Data documents.

We recently published results (in the main track of ESWC) based on initial analyses of the first six months of data that we have collected. These results begin to answer *some* of the questions listed above [1]. For example, therein:

1. We showed that 5% of documents went offline in six months, establishing an initial estimated death-rate for Linked Data documents.
2. We showed that the content of 62.2% of the monitored documents did not change in six months, and that most of the remaining documents either changed very infrequently (23.2%) or very frequently (8.4%).
3. We identified four types of data-sites: Static involving few infrequently changing documents, including linkedmdb.org; Dual involving a few frequently updated documents, including, loc.gov; Bulk involving many infrequently updated documents, including dbpedia.org; and Active involving many frequently updated documents, including dbtropes.org.
4. Of those documents that changed at least once, we showed that one quarter only ever updated individual values (often an object literal), one quarter only ever added new triples, and that the other half contained a mix of change types. We also showed, e.g., that the schema signature of a document (set of class and property terms used) rarely changes.
5. Links in the monitored documents are quite static over time, with the exception of a few domains such as sec.gov, identi.ca, zitgist.com, dbtropes.org and freebase.com that regularly contribute a small volume of fresh links.

We would now like to disseminate our collection and our results for the benefit of the community. We wish to collect use-cases for statistics on Linked Data dynamics that we can extract from our collection and provide online. We are currently working on a live site that visualises changes in Linked Data sites and allows interested users to see changes happen on a weekly basis, where they occur, and what types of changes they are (`http://swse.deri.org/dyldo`). We are also in the process of creating a SPARQL service that indexes details of weekly changes represented as RDF. This would enable external systems to find out which Linked Data sites were updated in the previous week(s) and what types of changes they exhibited. We hope that warehouses, live query engines, reasoners and publishers could then use our API to directly answer many of the questions about Linked Data dynamics highlighted at the outset.

References

1. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing Linked Data Dynamics. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 213–227. Springer, Heidelberg (2013)
2. Käfer, T., Umbrich, J., Hogan, A., Polleres, A.: DyLDO: Towards a Dynamic Linked Data Observatory. In: LDOW at WWW. CEUR-WS, vol. 937 (2012)