Learning from Linked Open Data Usage: Patterns & Metrics

Knud Möller, Michael Hausenblas,
Richard Cyganiak, Siegfried Handschuh
Digital Enterprise Research Institute (DERI),
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway,
Ireland
knud.moeller@deri.org,
michael.hausenblas@deri.org,
richard.cyganiak@deri.org,
siegfried.handschuh@deri.org

Gunnar AAstrand Grimnes
German Research Center for Artificial
Intelligence (DFKI) GmbH, Knowledge
Management Group
D-67663 Kaiserslautern, Germany
grimnes@dfki.uni-kl.de

ABSTRACT

Although the cloud of Linked Open Data has been growing continuously for several years, little is known about the particular features of linked data usage. Motivating why it is important to understand the usage of Linked Data, we describe typical linked data usage scenarios and contrast the so derived requirement with conventional server access analysis. Then, we report on usage patterns found through an in-depth analysis of access logs of four popular LOD datasets. Eventually, based on the usage patterns we found in the analysis, we propose metrics for assessing Linked Data usage from the human and the machine perspective, taking into account different agent types and resource representations.

Keywords

linked data, Web of Data, access, usage patterns

1. INTRODUCTION

Linked Open Data (LOD) [6] is a recent community effort to alleviate the problem of missing, sufficiently interlinked datasets on the Web of Data. Through this effort, a significant number of large-scale datasets[1] have now been published in the LOD cloud[2], which is growing constantly.

At time of writing, 6.7 billion RDF triples and over 140 million links between datasets are available [5]. Though research is known to investigate search engine crawls and logs [8, 13, 17] the usage behaviour of agents — humans and machines alike — concerning linked data has so far not gained much attention. Hence, in this work we analyse access logs of LOD data sets in order to *learn how Linked Data is used*. Our contribution is twofold: (i) We report on usage patterns found in LOD datasets. (ii) Based on our observation we propose an initial set of dedicated Linked Data usage metrics. As a starting-point we briefly review related work (Sect. 1.1) and discuss the challenges of analysing linked data usage in Sect. 2. We then report on the results of our analysis in Sect. 3 and propose usage metrics for Linked Data in Sect. 4. Sect. 5 concludes our work and sketches next steps.

[1]For example, DBpedia (`http://dbpedia.org/`), BBC music (`http://www.bbc.co.uk/music/`), LinkedGeoData (`http://linkedgeodata.org/`), and only recently by the New York Times (`http://data.nytimes.com/`)

[2]`http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14.html`

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

1.1 Related Work

Analysing server logs is as old as the Web itself [22]. To this end, research has focused typically [27, 21, 26] on: (i) server-side: performance, optimisation (load balancing, etc.) (ii) client-side: customisation, etc. Our work can be seen as a case of Web use mining [17, 26] in the wider sense, with a focus on the analysis of semantically-enabled Web sites [20]. However, to the best of our knowledge none of the existing work has looked specifically at LOD or SPARQL (the query language of the Web of linked data) usage.

2. MOTIVATION

Much as the current Web (of documents) is mainly targeting human users, a particular strength of linked data is that applications can use it directly [12]. We hence differentiate two fundamental types of consumers concerning the usage of LOD: (i) **Human Users**, equipped with a generic Linked Data browser [14] such as Tabulator [4] or Sigma [7] on the one hand, and (ii) **Machine Agents**, that is, a piece of software not under the direct control of a human, on the other hand. One would assume that human and machine agents differ in terms of usage patterns. Whereas we suspect human users to browse the LOD datasets in a more traditional, rather unpredictable sense, we would imagine machine agents to be more "focused": machine agents are typically based on a fixed rule set encoded in their program. Additionally, if LOD is primarily targeting applications rather than humans, we would expect the majority of the usage caused by machine agents.

2.1 Motivating Challenges

Our motivation to better understand the usage of LOD data is tightly related to machine agents. There are a couple of challenges concerning LOD usage, especially from the machine agents point of view, which have so far been neglected by and large:

- Concerning **reliability**: with the recent additions from the commercial domain such as BBC and NY times, the LOD cloud developed into a commercial-strength global database. If one is about to use LOD from an application, the availability of the data is crucial.
- Concerning the **peak-load**: certain LOD datasets (or certain entities in LOD datasets) may be requested more frequently then others; this might be due to events (such as conferences, celebrations, launches, etc.) or due to their role as linking hubs.
- Concerning **performance**: knowing what queries are being executed against your store helps configuring caches and indexes.
- Concerning **usefulness**: what parts of your data is being accessed often, what links are people finding and following, or alternatively, searching for and NOT finding.
- Concerning **attacks**: with the growing commercial usage, targeted attacks, such as or spam[3] have to be taken seriously.

In order to address the above challenges, one needs to understand in-depth who is using the data and how it is used. To better understand the usage, we have performed an analysis of the server access logs of major LOD sites and report on the findings in the next section.

3. ANALYSIS OF LINKED DATA ACCESS LOGS

In this section, we are first going to give an overview of the four different evaluated datasets, and the source data available for each. Afterwards, we will look into a number of questions which aim to increase our understanding of linked data usage.

3.1 Source Data

In order to analyse the usage of linked data sites, answer questions relating to usage patterns and classify them according to the metrics proposed in this paper, we take the basic approach of examining the server log files of the sites in question. Such log files record each individual HTTP request that is made to a server, keeping information about such things as the requested URI, the HTTP method used, the origin of the request, the exact time of the request, the agent performing the request and details about the response of the server (for an example see Fig. 1). While different Web servers use different log formats, the scope of the data recorded in each format is similar. By far the most common format is the *common log format* (CLF)[4], or the slightly extended *combined log format*. The latter was used by the servers hosting all four datasets in our analysis.

3.2 Evaluated Datasets

For our analysis, we had access to server log files from four different LOD sites: DBpedia, DBTune, RKBExplorer and SWC (aka "Semantic Web Dog Food"). All four sites differ greatly with respect to several of their basic characteristics, such as size (in number of RDF triples), connectedness in the LOD cloud, functionality beyond serving of linked data, etc. All four datasets together provide us with good coverage of the different types of datasets which make up the Web of linked data.

Below, we will give a brief introduction to each site, as well as an overview of some of their basic statistics in Tab. 1, such as size, period of time observed and number of hits in different categories. Specifically, we distinguish between requests to the SPARQL endpoints of each site and three related kinds of URIs which all reflect the same resource, in the sense that the *plain* resource URI is the identifier of a non-information resource [15] such as "WWW2009", while the related *RDF* and *HTML* document URI are identifiers for information resources, or representations in different formats about WWW2009. This is discussed in more detail in Sect. 3.3.1. For orientation purposes, the total number of hits to a site is also given. Because we had access to different amounts of log file data for each site, Tab. 1 gives both the absolute numbers for each site, as well as the average per day. For reasons discussed below, some of the statistics could not be generated for the RKBExplorer data set.

3.2.1 Semantic Web Dog Food

The smallest dataset in our analysis in terms of RDF triples (∼80,000 RDF) is served through the Semantic Web Conference metadata site (SWC or "Dog Food") [19, 18]. SWC holds RDF data about a number of large, international conferences in the Web and Semantic Web area, such as WWW, ISWC and ESWC, as well as a growing number of workshops. For each such event, detailed data about papers, authors, events and other entities is provided, both as RDF and as HTML documents. For this dataset, we had the best coverage over time, comprising of log files spanning well over 1 1/2 years, from 21/07/2008–03/10/2010.

3.2.2 DBpedia

By far the largest dataset in our analysis is the DBpedia [2], which provides linked data based on an extraction of structured data from Wikipedia. Because of its wide coverage in background knowledge entities such as people, places, species, etc., DBpedia can be considered a hub within the Web of linked data, in that it is used as a point of reference by many other datasets. The DBpedia site serves both RDF and HTML documents about its resources. For DBpedia, we had access to server log files dating from 30/06/2009–25/10/2009 (i.e., 118 days).

3.2.3 DBTune

DBTune[5] [24] is a meta-site which hosts different (currently 10) sub-datasets of linked data for a number of music-related non-LOD datasets, such as MusicBrainz, MySpace or Jamendo. While all datasets are hosted within the DBTune namespace, each of them is served in a slightly different manner, which makes an integrated analysis complicated. E.g., the MusicBrainz dataset is hosted through a D2R server instance and provides both HTML and RDF representations for its resources, for MySpace only RDF descriptions are provided at document-type URIs, while for Jamendo only RDF descriptions via SPARQL DESCRIBE

[3] `http://iandavis.com/blog/2009/09/linked-data-spam-vectors`

[4] CLF is an informal standard with no official reference. As a general introduction, we point the reader to `http://en.wikipedia.org/wiki/Common_Log_Format`.

[5] `http://dbtune.org`

Figure 1: The combined log format

	# triples	# days	total # hits	# plain hits	# RDF hits	# HTML hits	SPARQL
Dog Food	79,175	597	8,427,967	1,923,945	259,031	1,647,205	879,932
			(14,117)	*(3,223)*	*(434)*	*(2,759)*	*(1,471)*
DBpedia	109,750,000	118	87,203,310	22,821,475	7,008,310	22,999,237	20,972,630
			(739,011)	*(193,402)*	*(59,392)*	*(194,909)*	*(177,734)*
DBTune	74,209,000	61	7,467,125	1,952,185	1,135,509	677,904	3,055,493
			(122,412)	*(32,003)*	*(18,615)*	*(11,113)*	*(50,090)*
RKBExplorer	91,501,684	29	529,938	—	—	—	9,327
			(18,274)	*(—)*	*(—)*	*(—)*	*(322)*

Table 1: Overview of four LOD datasets

queries are served. For our evaluation, we had access to log files in two periods: from 24/05/2009–21/06/2009 and from 27/09/2009–29/10/2009, i.e., roughly two months.

3.2.4 RKBExplorer

RKBExplorer[6] [11] is another meta-dataset currently comprising 44 sub-datasets covering various topics and sources within the domain of academic research, as well as a Web application that allows users to access and browse its content in an integrated fashion. Both RDF and HTML documents about the resources in all datasets are available. Apart from serving linked data, the site also features a module that provides co-reference resolution functionality [10]. For our evaluation, we had access to log files in the period from 24/05/2009–21/06/2009, i.e., roughly one month. However, since the log files were partially broken (no referrer IPs were recorded), and because their structure was slightly modified in comparison to the conventional log file format, we were only able to make use of the dataset in some of our experiments.

3.3 A New Breed of Agents

Since we expect usage of linked data to be different from conventional Web usage, we can also expect to find new kinds of agents. In this section we define what we consider to be "semantically aware" agents, which are explicitly targeted at the Web of linked data.

3.3.1 Detecting Semanticity

By classifying an agent as "semantic", we imply that it is capable of processing structured, semantic data, i.e., RDF. Whether or not an agent has this capability can only be determined indirectly from the log files, based on some heuristics. Making the assumption that any agent which explicitly requests semantic data from a server also knows how to process it, we will classify such agents as "semantic". In detail, we use the following two heuristics:

- **SPARQL requests:** if an agent sends a request containing a SPARQL query, we assume that it is capable of handling the query result, i.e., either a set of bindings (in the case of a SELECT query), potentially containing URIs of RDF resources, or an RDF graph (in the case of a CONSTRUCT or DESCRIBE query).
- **RDF requests:** if an agent *directly* requests RDF from a server, we assume that it knows how to process data in this format. Directly here means that the agent specified an RDF syntax such as rdf/xml as an acceptable response in the header of its request. Merely requesting the URI of an RDF representation does not suffice to indicate semanticity, as this could simply mean that the agent followed a link to this representation.

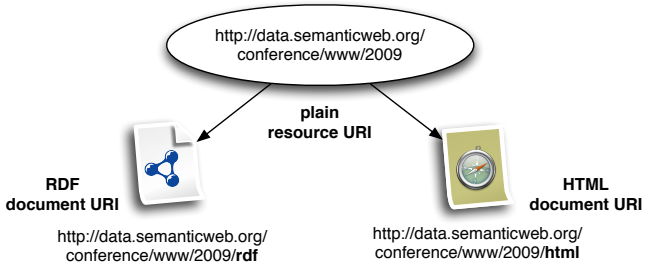


Figure 2: Plain resource, RDF and HTML representations

Detecting SPARQL requests is straightforward, since the requested URI will contain the actual SPARQL query. However, log files of Web servers do not normally record the header for each request[7], which makes it less straightforward to apply the second heuristic. Nevertheless, there is an indirect way to apply it in some cases, based on the

[6] http://www.rkbexplorer.com

[7] Web servers can be configured to also log information such as request headers. In fact, this has been done by the administrators of RKBExplorer, which makes it easy to detect semantic agents in this site's log files.

way data is served from many linked data sites. As mentioned in Sect. 3.2, the sites we analyse serve each resource in a data set according to the best-practice [25] of assigning different identifiers to the resource itself and its representations in RDF and HTML (and possibly other formats). This principle is illustrated in Fig. 2, showing the three different URIs for the plain resource (the *non-information resource* [16]) and its representations (the *information resources*) from the Dog Food dataset. Agents can either request the various representations directly, or they can request the plain resource and indicate the desired format to the server in the so-called *request header*, e.g. like this: `Accept: application/rdf+xml`. The server will then redirect to the corresponding representation, using a a method called *content negotiation*. This process is reflected in the log files by two entries: the first one will be a request for the plain resource, and is answered by the server with the HTTP 303 code, indicating a redirection. The second entry will be for the corresponding resource representation, and is answered by the server with the HTTP 200 code, indicating a successful request. An example of how this looks like in the logs of the SWC server is shown in List. 1: the first request for the URI of the resource representing VU Amsterdam is redirected to an RDF document about this resource, which indicates that the "rdflib-2.4.0" user agent had requested rdf/xml.

Following this approach, we can determine if a request for an RDF representation of a resource was in fact a "semantic request", and therefore whether or not the requesting agent can be classified as semantically aware.

3.3.2 *Kinds of Semantic Agents*

By using the method defined above, we were able to detect semantic agents in all four datasets. As with conventional agents, also semantic agents can be divided into different sub-classes such as browsers (human usage) and bots (machine usage), as well as tools (curl, wget, ...) and data-services. The latter is a term introduced for agents which provide a service for other agents by processing some data on the Web. In contrast to crawlers, the purpose here is not archiving or indexing. Examples of data-services are format converters, snapshot generators, etc. Fig. 3 shows the distribution of agents according to those classifications, using the Dog Food dataset as an example (the other datasets show similar distributions). With the exception of data-services, in all agent types the distribution is still very clearly in favour of conventional agents.

3.4 Demand for Semantic Data

Many linked data sets are published in a way that provides both RDF data to semantically enabled agents, as well as simple HTML representations for human browsing. In other words, such data sets are exposed both to the conventional eye-ball Web and to the semantic Web of linked data. The traditional metric of the development of traffic over time (in terms of hits or visits) for a particular site or data set can therefore not be applied straight away to the domain of linked data: by indiscriminately measuring traffic, the distinction between both kinds of access is lost.

When asking how the demand for a particular data set has developed over time from the point of view of the Web of linked data, we therefore distinguish between different kinds of traffic. In particular, we measure traffic for plain

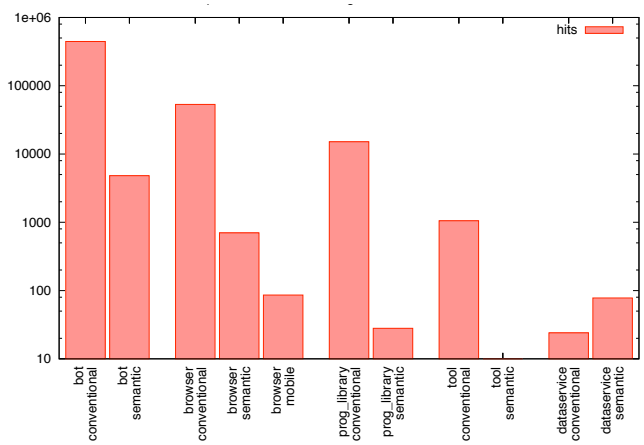


Figure 3: Distribution of agents grouped by type and semanticity

resource URIs, HTML representations, RDF representations and semantic requests a a subset of RDF requests (following the approach defined in Sect. 3.3.1). Figure 4 gives a general overview in these terms for each dataset in our evaluation (with the exception of RKBExplorer, due to the restrictions discussed above). The graphs show that demand for RDF representations is low across the board, with the amount of actual semantic requests being almost negligible. Also, the ratio of plain requests vs. HTML + RDF requests shows that agents do not always use the plain resource URIs as the entry point, but also request representations directly.

In order to get a more detailed analysis over time, Fig. 5 plots[8] the traffic for all four request types on the Dog Food server. As discussed above, representations can also be requested directly, which explains why the sum of HTML and RDF requests on any given day will usually exceed the number of plain requests[9].

The graph clearly shows that conventional traffic on the Dog Food site has been increasing steadily since its inception in July 2008, while demand for RDF data has been more or less static on a comparably low level. In fact, true semantic requests are even lower than requests for RDF representations, which indicates that most agents requesting semantic data blindly follow links, without actually "knowing" what to do with the received RDF data.

The same evaluation on the DBpedia logs supports the observation that no increasing demand for semantic data can be observed, at least not within the time frame accessible to our experiments. This is illustrated in Fig. 6, which shows that also here the amount requests for conventional data is much higher than for semantic data. We did not evaluate the other two datasets in this way, since the time frames of one and two months available to was too small to give conclusive results.

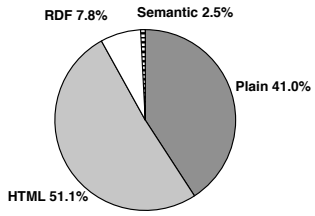
3.5 Real-world-influenced Interest in Datasets

[8]The raw hit count data has been smoothed using an exponential moving average algorithm to remove noise and highlight possible trends.

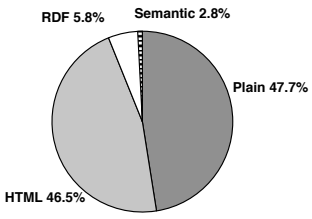
[9]Exceptions to this rule are likely to be explained with high server load, which can lead to some unsuccessful redirections for plain resource requests.

```
90.21.243.141 - - [06/Oct/2008:16:07:58 +0100] "GET /organization/vrije-universiteit-amsterdam-the-
    netherlands HTTP/1.1" 303 7592 "-" "rdflib-2.4.0 (http://rdflib.net/; eikeon@eikeon.com)"
90.21.243.141 - - [06/Oct/2008:16:08:02 +0100] "GET /organization/vrije-universiteit-amsterdam-the-
    netherlands/rdf HTTP/1.1" 200 45358 "-" "rdflib-2.4.0 (http://rdflib.net/; eikeon@eikeon.com)"
```

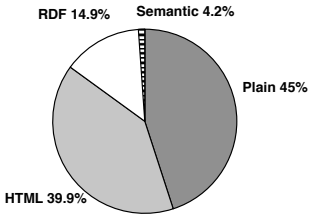
Listing 1: Redirection evidence to RDF representation in CLF



(a) Semantic Web Dog Food



(b) DBpedia



(c) DBTune

Figure 4: Distribution of traffic by request type

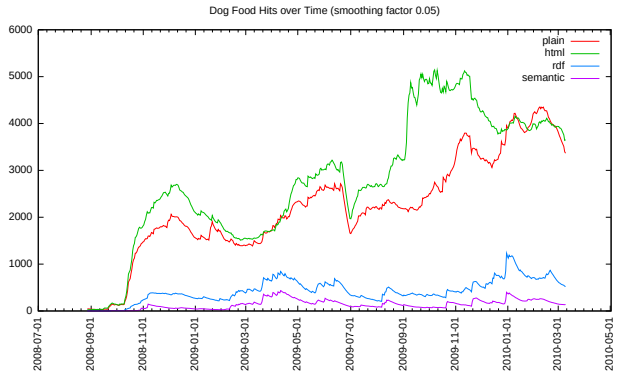


Figure 5: Hits over time for different request types on the Semantic Web Dog Food site

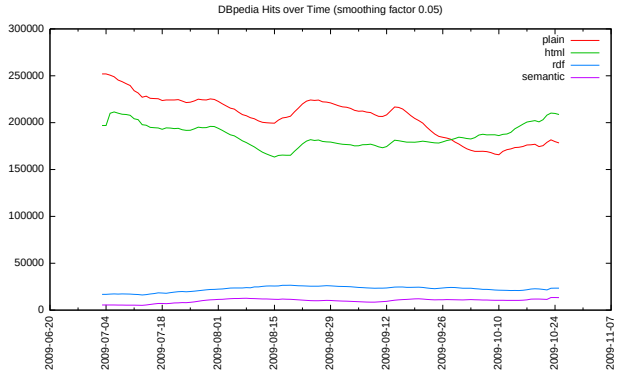


Figure 6: Hits over time for different request types on DBpedia

A possible metric to indicate the relevance of a dataset in the linked data cloud is the level of influence that real-world events have on its access statistics. We have measured this by filtering the hit curve by specific resources which are central to those events. The hit curve shows the sum of requests for HTML and RDF representations per day. Since the curve for semantic requests was not distinguished by any unique features, we did not include it in the presentation here. We restricted this analysis to the Dog Food dataset and DBpedia, since RKBExplorer and DBTune are either too static in nature, or our coverage over time was too limited.

In the case of the Dog Food dataset, the hypothesis is that requests for data from specific conferences would be noticeably higher around the time when the event took place. Since all resources unique to an event fall into the same namespace, we were able to extract requests to those resource by using the corresponding namespace as a filter. We did a separate measurement for every major conference that falls into the time period covered by the log files available to us: ESWC2009, ISWC2008, ISWC2009 and WWW2009. Fig. 7 shows the results of our analysis. Access for each conference is plotted in a different colour and line style, while the time frame of the conference is marked by a bar in the same colour. Contrary to our expectations, there are no significantly higher access rates around the time of the event. Instead, we see an initial increase in traffic around the time of the event (when the dataset was published), and afterwards a more or less stable rate of access. One exception to this observation could be WWW2009, where there is significantly higher traffic in the weeks after the conference. The spikes in the second half of 2009 seem to be anomalies and not part of a general trend. However, this finding is not enough to conclude a general trend of higher targeted traffic around time of a conference.

For DBpedia, we picked two events within the time period covered by the log files that have generated significant public interest internationally, expecting to see this increased interest reflected in the usage of the dataset as well: (i) Michael Jackson's memorial service 7th July 2009 (shortly after the first day of log coverage) and (ii) the second Irish referendum on the Treaty of Lisbon on 2nd October 2009 (shortly before the end of the log coverage). For each event, we

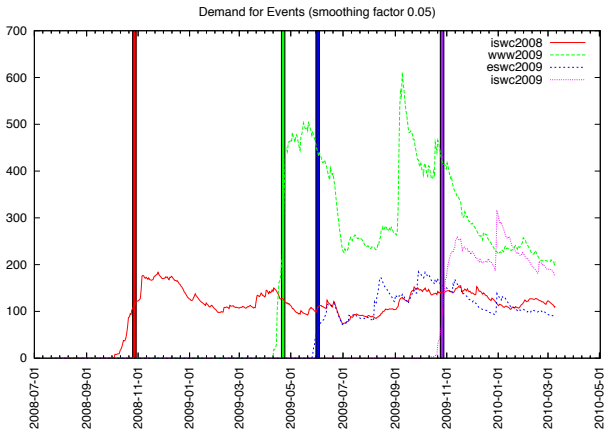


Figure 7: Requests for different conferences

chose three relevant resources and measured their access statistics over time. In particular, those resources were `dbpedia:Michael_Jackson`, `dbpedia:Michael_Jackson_me\-morial_service` and `dbpedia:Staples_Center` (the location of the memorial service in Los Angeles) for (i), and `dbpedia:\-Republic_of_Ireland`, `dbpedia:European_Union`, and `dbpedia:Treaty_of_Lisbon` for (ii).

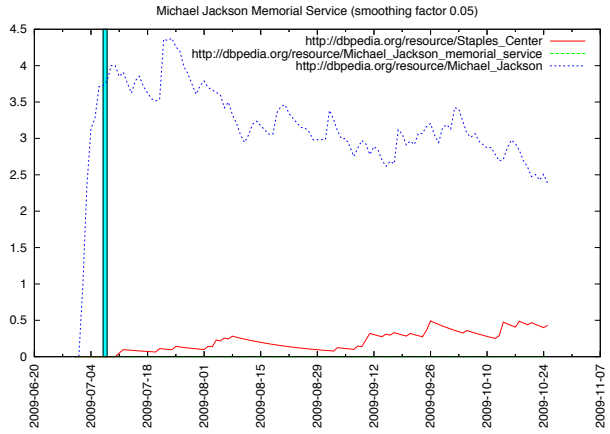


Figure 8: Interest in the Michael Jackson memorial service

In the case of Michael Jackson's memorial service, Fig. 8 shows dropping interest in the `dbpedia:Michael_Jackson` resource after the event (we do not have enough log file coverage to report on interest before the event), pointing towards a correlation between the real-world event and usage of the DBpedia dataset. The other two resources show no such correlation. In fact, there have been no requests for `dbpedia:Michael_Jackson_memorial_service` at all, possibly due to the delay in the propagation of changes from Wikipedia to DBpedia.

Figure 9 shows a similar situation for the Lisbon Treaty referendum. While two of the chosen resources (`dbpedia:\-European_Union` and `dbpedia:Treaty_of_Lisbon`) show no correlation between the real-world event and dataset usage,

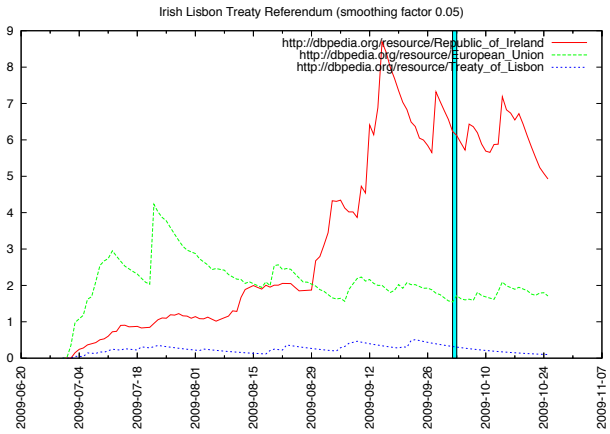


Figure 9: Interest in the Irish referendum on the Lisbon Treaty

the `dbpedia:Republic_of_Ireland` resource shows a definite increase in traffic in the weeks preceding the referendum, as well as a possible drop afterwards. This, again, could be interpreted as evidence for a correlation between between real-world events and dataset usage.

3.6 SPARQL Evaluation

Apart from direct access to resources and their representations, structured querying is the other significant aspect of linked data usage. To get a better understanding of how SPARQL is used "in the wild", we performed a basic analysis of the queries encountered in all four data sets, on which we report in this section. In our analysis, we looked at the type of queries issued, the complexity of each query and what resources/predicates were queried for. Table 2 shows the total number of queries and break-down for each SPARQL query-type in percent. Note that only the RKB data contained *no* CONSTRUCT, DESCRIBE or ASK queries, but for the other data-sets these query types still only made up a fraction of a percent of the total. We also show the percentage of queries that did not parse, these errors are mainly due to missing PREFIX declarations, illegal URIs or literals. For all data-sets over 90% of the queries are of the SELECT type.

Type	DBTune	DogFood	RKB	DBpedia
% ASK	0	0	0	0.05
% CONSTRUCT	0	0	0	0
% DESCRIBE	0.02	0.01	0	0
% SELECT	0.94	0.99	0.95	0.91
% Error	0.03	0	0.05	0.04
Total	3,055,084	253,752	9,319	12,217,986

Table 2: SPARQL Query-type breakdown.

Looking at these SELECT queries only, Table 3 shows the number of triple-patterns per query, again as percentage of the total number of select queries for easy comparison. By triple-pattern we mean a single pattern matching an RDF subject, predicate, and object triple, FILTER and similar constructs are not included. For both DBTune and the Dog Food dataset over 99% of the SELECT queries contain a

single triple-pattern only. Another potentially interesting observation is the number of variables and the number of *joins*, i.e., the number of re-occurring variables per query. However, these values strongly correlated with the length of the query, and are for space-reasons not shown here. For all data-sets larger queries with more triple patterns constitute a too small fraction to show in the table, but our data-sets do contain some queries with up to 16 triple-patterns and 12 different variables.

# Patterns	DBTune	DogFood	RKB	DBpedia
1	0.99	0.99	0.28	0.58
2	0.01	0.01	0.72	0.06
3	0	0	0	0.35
4	0	0	0	0

Table 3: % of SELECT queries with N triple patterns.

For the SELECT queries containing only a single pattern we have looked at the generic types of patterns being used, ignoring the specific URIs and looking at the types only. Table 4 shows the three most common structures of triple-patterns, here r signifies a resource, i.e. a URI, l a literal and ? a variable. The numbers are again in percent of all triple patterns, and one can see that these three patterns alone cover 85–99% of the queries. This is signifiant when chosing which indexes to pre-compute and store for serving LOD data. Another statistic relevant for optimsing storage is which predicates are used, for space reasons the details are not included, but again one can cover the majority of queries with a few predicates, such as *rdf:type*, *rdfs:label*, *dbpedia:abstract*, *foaf:name*.

Pattern	DBTune	DogFood	RKB	DBpedia
$(?, r, l)$	0.06	0.29	0	0.03
$(?, r, r)$	0.25	0.26	0.18	0.46
$(r, r, ?)$	0.68	0.43	0.68	0.5

Table 4: Main query pattern-types.

4. DISCUSSION AND PROPOSED METRICS

Performing usage analysis based on server access logs has known limitations[10]. Due to the Web architecture [15], one finds many intermediary components [9] such as proxies, caches, etc. between a server and a client. Hence, when analysing server access logs, we must consider that we are only aware of the server-side and can, in general, not assess the entire communication.

Regarding the application server access log analysis to linked data, there are further issues that need to be considered: (i) For embedded RDF serialisations such as RDFa, one cannot use the requested content type to determine the type of agent. In this work we have limited our investigations to a setup based on content negotiation, which does not have this limitation. However, in order to address this, one could additionally evaluate the logs from RDF converter services (such as `http://www.w3.org/2007/08/pyRdfa/`) to determine if an RDF representation has been obtained, subsequently to the initial request. (ii) Independent of the RDF serialisation, tracking movement of an agent between LOD datasets is very challenging. Beside the issue that one needs both access logs, one must also be able to determine which requests stems from a certain agent. (iii) Another limitation is the fact that we are only considering traffic on the particular site where a dataset is hosted. If the data has been aggregated through a service such as Sindice[11] and is used there, our approach will not be aware of this usage, except for the fact that the site has been crawled by Sindice.

Based on our analyses and discussions, we propose the following generic LOD usage metrics to assess the usage of LOD from a machine agent's perspective:

- the "machine-agent-awareness" factor ($ma2$):

$$ma2 = \frac{\#agents_{RDF-aware}}{\#agents_{total}} \quad (1)$$

- the "data-request"–ratio (dr):

$$dr = \frac{\#requests_{RDF}}{\#requests_{total}} \quad (2)$$

- the "redirection"–ratio (r):

$$r = \frac{\#requests_{303}}{\#requests_{200}} \quad (3)$$

Concerning SPARQL-specific usage, we propose:

- the "query-lookup"–ratio (ql):

$$ql = \frac{\#GET_{SPARQL}}{\#GET_{total}} \quad (4)$$

- the "query-complexity" levels (qc_i):

$$qc_i = \frac{\#pattern_i}{\#GET_{SPARQL}} \quad (5)$$

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have reported on our ongoing work of the analysis of linked data usage as a new phenomenon that needs to be analysed differently that conventional Web usage. We have proposed the foundation for a methodology that is based on the analysis of commonly found server log files and typical linked data site configurations. In using our approach, we have analysed four different linked data sets and addressed questions such as which new kinds of agents can be encountered, how a change in demand for semantic, linked data can be measured and whether or not any influence of real-world events on linked data usage can be observed. All of these questions can be used as indicators for the general state of acceptance and uptake of the Web of linked data, aka the Semantic Web. Additionally, we have used the same datasets to present a first iteration of an analysis of SPARQL queries "in the wild", as a second kind of linked data usage, and as a possible source of real-world information for SPARQL implementers. Finally, we have discussed some general problems with our analysis methodology and proposed a set of metrics to classify linked data sites.

The main focus of future work will be in extending our log file corpus, particularly in order to have better coverage over

[10] `http://www.boxesandarrows.com/view/the-limitations-of`

[11] `http://sindice.com/`

time for individual datasets. Coverage of one or two months has proven to be insufficient for many relevant queries. Also, we will extend, apply and evaluate the SPARQL analysis and our proposed metrics.

Acknowledgments

We would like to thank Chris Bizer and the DBpedia team, Yves Raimond and the DBTune team and Hugh Glaser and the RKBExplorer team for providing us with the logfiles of their respective sites and in general for their great help in compiling the data for this research. The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon-2) and (in part) by the European project FAST No. FP7-ICT-2007-1.2 216048.

6. REFERENCES

[1] C. Anderson. *The Long Tail*. Random House, 2006.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 722–735, 2007.

[3] T. Berners-Lee. Linked Data–Design Issues, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`.

[4] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing linked data on the Semantic Web. In *In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06)*, Athens, Georgia, USA, 2006.

[5] C. Bizer. The Emerging Web of Linked Data. *IEEE Intelligent Systems*, 24(5):87–92, 2009.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[7] R. Cyganiak, M. Catasta, and G. Tummarello. Towards ECSSE: live Web of Data search and integration. In *WWW09: Semantic Search Workshop*, Madrid, Spain, 2009.

[8] L. Ding and T. Finin. Characterizing the Semantic Web on the Web. In *5th International Semantic Web Conference (ISWC06)*, Athens, GA, USA, 2006.

[9] R. Fielding and R. Taylor. Principled design of the modern Web architecture. *ACM Trans. Internet Technol.*, 2(2):115–150, 2002.

[10] H. Glaser, A. Jaffri, and I. Millard. Managing co-reference on the Semantic Web. In *Linked Data on the Web (LDOW2009) at WWW2009 Madrid, Spain*, 2009.

[11] H. Glaser, I. Millard, and A. Jaffri. RKBExplorer.com: A knowledge driven infrastructure for linked data providers. In *5th European Semantic Web Conference (ESWC2008), Tenerife, Spain*, volume 5021 of *LNCS*, pages 797–801. Springer, 2008.

[12] M. Hausenblas. Exploiting Linked Data to Build Web Applications. *IEEE Internet Computing*, 13(4):68–73, 2009.

[13] M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. What is the Size of the Semantic Web? In *I-Semantics 2008: International Conference on Semantic Systems*, Graz, Austria, 2008.

[14] T. Heath. How Will We Interact with the Web of Data? *IEEE Internet Computing*, 12(5):88–91, 2008.

[15] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004, W3C Technical Architecture Group (TAG), 2004.

[16] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One. W3C Recommendation. Recommendation, W3C, December 2004. `http://www.w3.org/TR/2004/REC-webarch-20041215/`.

[17] P. Mika, E. Meij, and H. Zaragoza. Investigating the Semantic Gap through Query Log Analysis. In *8th International Semantic Web Conference (ISWC2009)*, Washington DC, USA, 2009.

[18] K. Möller. *Lifecycle Support for Data on the Semantic Web*. PhD thesis, National University of Ireland, Galway, 2009.

[19] K. Möller, T. Heath, S. Handschuh, and J. Domingue. Recipes for Semantic Web Dog Food—The ESWC2006 and ISWC2006 Metadata Projects. In *6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC+ASWC2007), Busan, South Korea*, pages 802–815. Springer, 2007.

[20] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez. Conceptual user tracking. In *Advances in Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003, Proceedings*, volume 2663 of *Lecture Notes in Artificial Intelligence*, pages 142–154. Springer, 2003.

[21] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. Spyropoulos. Web Usage Mining as a Tool for Personalization: A Survey. *User Modeling and User-Adapted Interaction*, 13(4):311–372, 2003.

[22] J. E. Pitkow and K. Bharat. WEBVIZ: A Tool for World Wide Web Access Log Visualization. In *Proceedings of the First International Conference on the World-Wide Web (WWW94)*, Geneva, Switzerland, 1994.

[23] E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. W3C recommendation 15 january 2008, W3C RDF Data Access Working Group, 2008.

[24] Y. Raimond, C. Sutton, and M. Sandler. Interlinking music-related data on the web. *IEEE MultiMedia*, 16(2):52–63, April-June 2009.

[25] L. Sauermann, R. Cyganiak, D. Ayers, and M. Völkel. Cool URIs for the Semantic Web. Interest group note, W3C, December 2008. `http://www.w3.org/TR/cooluris/` 05/05/2009.

[26] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, 2000.

[27] C. Stolz, M. Viermetz, M. Skubacz, and R. Neuneier. Guidance Performance Indicator " Web Metrics for Information Driven Web Sites. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 186–192, Washington, DC, USA, 2005. IEEE Computer Society.