What is the Size of the Semantic Web?

Michael Hausenblas, Wolfgang Halb
(Institute of Information Systems & Information Management,
JOANNEUM RESEARCH, Austria
firstname.lastname@joanneum.at)

Yves Raimond
(Centre for Digital Music,
Queen Mary, University of London, United Kingdom
yves.raimond@elec.qmul.ac.uk)

Tom Heath
(Talis, United Kingdom
tom.heath@talis.com)

Abstract: When attempting to build a scaleable Semantic Web application, one has to know about the size of the Semantic Web. In order to be able to understand the characteristics of the Semantic Web, we examined an interlinked dataset acting as a representative proxy for the Semantic Web at large. Our main finding was that regarding the size of the Semantic Web, there is more than the sheer number of triples; the number and type of links is an equally crucial measure.

Key Words: linked data, Semantic Web, gauging
Category: H.m, D.2.8

1 Motivation

Developments in the last twelve months demonstrate that the Semantic Web has arrived. Initiatives such as the Linking Open Data community project[1] are populating the Web with vast amounts of distributed yet interlinked RDF data. Anyone seeking to implement applications based on this data needs basic information about the system with which they are working. We will argue that regarding the size of the Semantic Web, there is more to find than the sheer numbers of triples currently available; we aim at answering what seems to be a rather a simple question: *What is the size of the Semantic Web*?

We review existing and related work in section 2. Section 3 introduces the linked dataset we use for our experiments. Further, in section 4 we analyse the reference dataset syntactically and semantically attempting to answer the *size* question. Finally, we conclude our findings in section 5.

[1] `http://linkeddata.org/`

2 Existing Work

On the Web of Documents, typically the number of users, pages or links are used to gauge its size [Broder et al. 00, Gulli and Signorini 05]. However, Web links (`@href`) are untyped, hence leaving its interpretation to the end-user [Ayers 07]. On the Semantic Web we basically deal with a directed labelled graph where a fair amount of knowledge is captured by the links between its nodes.

From semantic search engines we learn that mainly the documents and triples as such are counted. No special attention is paid to the actual interlinking, i.e. the type of the links [Esmaili and Abolhassani 06]. In the development of the semantic search engine `swoogle` [Finin et al. 05] it has been reported that "... the size of the Semantic Web is measured by the number of discovered Semantic Web Documents". However, later, they also examined link characteristics [Ding and Finin 06]. Findings regarding the distribution of URIs over documents are well known in the literature [Tummarello et al. 07, Ding et al. 05]. Unlike other gauging approaches focusing on the schema level [Wang 06], we address the interlinking aspect of Semantic Web data represented in RDF, comparable to what Ding et. al. [Ding et al. 05] did in the FOAF-o-sphere.

3 Linked Datasets As A Proxy For The Semantic Web

The *reference test data set* (RTDS) we aim to use should be able to serve as a good proxy for the Semantic Web, hence it (i) must cover a range of different topics (such as people-related data, geo-spatial information, etc.), (ii) must be strongly interlinked, and (iii) must contain a sufficient number of RDF triples (we assume some millions of triples sufficient). As none of the available alternatives—such as the Lehigh University Benchmark dataset[2], Semantic Wikis (such as [Völkel et al. 06]) or embedded metadata—exhibit the desired characteristics, the Linking Open Data datasets were chosen as the RTDS. We note that embedded metadata (in the form of microformats, RDFa, eRDF and GRDDL) are constituting a large part of the openly published metadata. However, the interlinking of this data is not determinable unambiguously.

The basic idea of linked data was outlined by Sir Tim Berners-Lee; in his note[3], a set of rules is being provided. The Linking Open Data (LOD) project is a collaborative effort; it aims at bootstrapping the Semantic Web by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [Bizer et al. 07]. As of time of writing roughly two billion triples and three million interlinks have been reported (cf. Fig. 1[4], ranging from rather centralised ones to those that are very distributed. A detailed description of the datasets contained in the LOD is available in Table 1.

[2] `http://swat.cse.lehigh.edu/projects/lubm/`
[3] `http://www.w3.org/DesignIssues/LinkedData.html`
[4] by courtesy of Richard Cyganiak, `http://richard.cyganiak.de/2007/10/lod/`

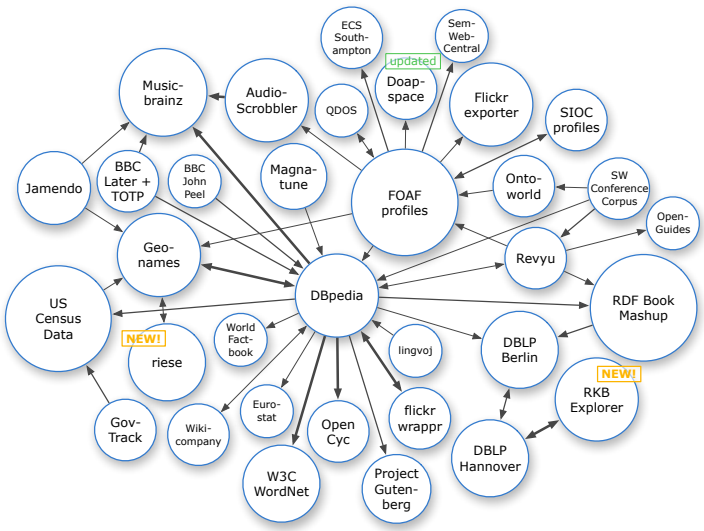

Figure 1: The Linking Open Data dataset at time of writing.

4 Gauging the Semantic Web

In order to find metrics for the Semantic Web we examine its properties by inducing from the LOD dataset analysis. One possible dimension to asses the size of a system like the Semantic Web is the data dimension. Regarding data on the Semantic Web, we roughly differentiate into: (i) the **schema level** (cf. ontology directories, such as OntoSelect[5]), (ii) the **instance level**, i.e. a concrete occurrence of an item regarding a certain schema (see also [Hausenblas et al. 07]), and the actual **interlinking**: the connection between items; represented in URIrefs and interpretable via HTTP. This aspect of the data dimension will be the main topic of our investigations, below.

As stated above, the pure number of triples does not really tell much about the size of the Semantic Web. Analysing the links between resources exhibits further characteristics. The LOD dataset can roughly be partitioned into two distinct types of datasets, namely (i) **single-point-of-access datasets**, such as DBpedia or Geonames, and (ii) **distributed datasets** (e.g. the FOAF-o-sphere or SIOC-land). This distinction is significant regarding the access of the data in terms of performance and scalability.

Our initial approach aimed at loading the whole LOD dataset into a relational database (Oracle 11g Spatial). Due to technical limitations this turned

[5] `http://olp.dfki.de/ontoselect/`

Name	**Triples** (millions)	**Interlinks** (thousands)	**Dump** download	**SPARQL** endpoint
BBC John Peel	0.27	2.1		
DBLP	28	0		yes
DBpedia	109.75	2,635	yes	yes
Eurostat	0.01	0.1		yes
flickr wrappr	2.1	2,109		
Geonames	93.9	86.5	yes	
GovTrack	1,012	19.4	yes	yes
Jamendo	0.61	4.9	yes	yes
lingvoj	0.01	1.0	yes	
Magnatune	0.27	0.2	yes	yes
Musicbrainz	50	0		
Ontoworld	0.06	0.1	yes	yes
OpenCyc	0.25	0	yes	
Open-Guides	0.01	0		
Project Gutenberg	0.01	0		yes
Revyu	0.02	0.6	yes	yes
riese	5	0.2	yes	yes
SemWebCentral	0.01	0		
SIOC	N/A	N/A		
SW Conference Corpus	0.01	0.5	yes	yes
W3C Wordnet	0.71	0	yes	
Wikicompany	?	8.4		
World Factbook	0.04	0		yes

Table 1: Linking Open Data dataset at a glance.

out not to be feasible—the overall time to process the data exceeded any sensible time constraints. As not all LOD datasets are available as dumps, it became obvious that additional crawling processes were necessary for the analysis. We finally arrived at a hybrid approach. The available and the self-crawled dumps together were loaded into the relational database, were the analysis took place using SQL. Additionally, we inspected the descriptions provided by the LOD dataset providers in order to identify parts of the dataset which are relevant for interlinking to other datasets. Where feasible, we also used the available SPARQL-endpoints.

4.1 Single-point-of-access Datasets

It has to be noted that only a certain subset of the links actually yields desirable results in the strict sense, i.e. return RDF-based information when performing an `HTTP GET` operation. Taking the DBpedia dataset as an example yields that

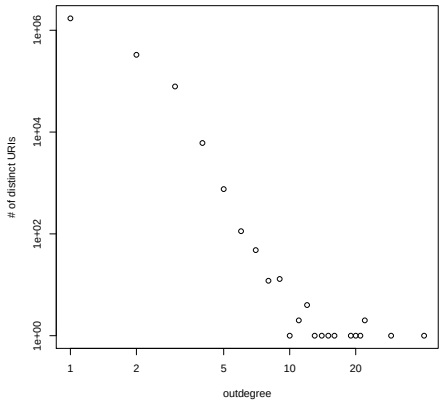

Figure 2: Outgoing Links From the DBpedia dataset.

only half of the properties in this dataset are dereferenceable. Fig. 2 depicts the distribution of the dereferenceable outgoing links from the DBpedia dataset. We would expect this distribution to be modelled by a power-law distribution considering the degree of DBpedia resources (the number of resources having a given number of links to external datasets). However, Fig. 2 does not clearly suggest this, which may be due to too little data or due to the fact that links from DBpedia to other datasets are created in a supervised way, whereas scale-free networks tend to represent organic and decentralised structures. We found

Property (Link Type)	**Occurrence**
`http://dbpedia.org/property/hasPhotoCollection`	2.108.962
`http://xmlns.com/foaf/0.1/primaryTopic`	2.108.962
`http://dbpedia.org/property/wordnet_type`	338.061
`http://www.w3.org/2002/07/owl#sameAs`	307.645
`http://xmlns.com/foaf/0.1/based_near`	3.479
`http://www.w3.org/2000/01/rdf-schema#seeAlso`	679

Table 2: Overall Occurrence of Link Types in the LOD dataset.

only a limited number of dereferenceable links in the LOD dataset (Table 2); this distribution is biased towards the DBpedia dataset and the flickr wrapper, however. In case of the single-point-of-access datasets, we found that mainly one

or two interlinking properties are in use (Fig 3). The reason can be seen in the way these links are usually created. Based on a certain template, the interlinks (such as `owl:sameAs`) are generated automatically. As the data model of the

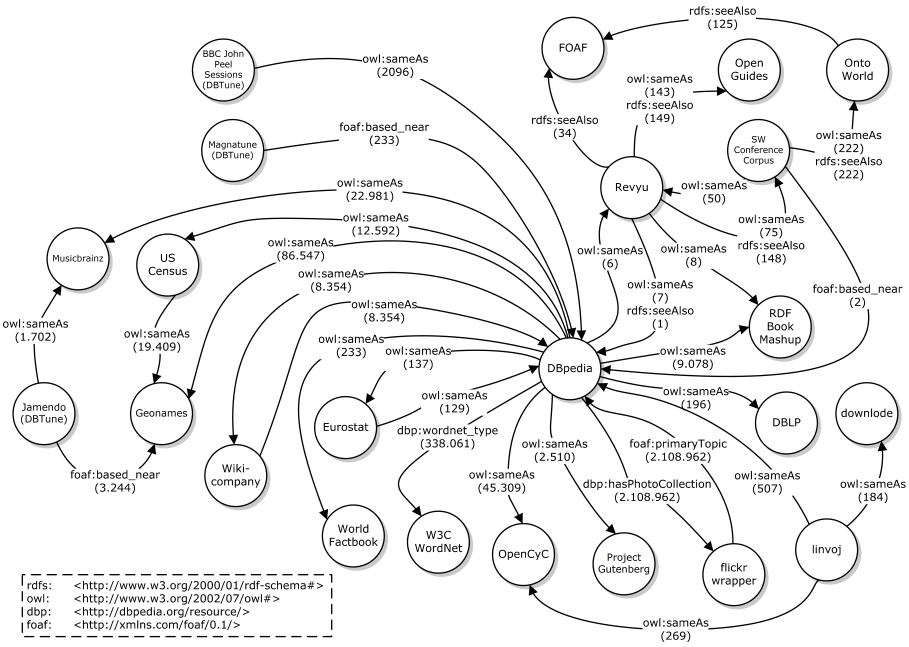

Figure 3: Single-point-of-access Partition Interlinking.

Semantic Web is a graph the question arises if the density of the overall graph can be used to make a statement regarding the system's size. The LOD dataset is a sparse directed acyclic graph; only a few number of links (compared to the overall number of nodes) exist. Introducing links is costly. While manual added, high-quality links mainly stem from user generated metadata, the template-based generated links (cheap but semantically low-level) can be added to a greater extent.

4.2 Distributed Datasets

In order to analyse the partition of the LOD covering the distributed dataset, such as the FOAF-o-sphere, we need to sample it. Therefore, from a single seed URI[6], approximately six million RDF triples were crawled. On its way, 97410 HTTP identifiers for persons were gathered. We analysed the resulting sampled FOAF dataset, yielding the results highlighted in Table 3.

[6] `http://kmi.open.ac.uk/people/tom/`

To	Interlinking Property	Occurrence
FOAF	`foaf:knows` (direct)	132.861
FOAF	`foaf:knows`+`rdfs:seeAlso`	539.759
Geonames	`foaf:based_near`	7
DBLP	`owl:sameAs`	14
ECS Southampton	`rdfs:seeAlso`	21
ECS Southampton	`foaf:knows`	21
DBPedia	`foaf:based_near`	4
DBPedia	`owl:sameAs`	1
RDF Book Mashup	`dc:creator`	12
RDF Book Mashup	`owl:sameAs`	4
OntoWorld	`pim:participant`	3
Revyu	`foaf:made`	142
Other LOD datasets	-	0
Total inter-FOAF links	-	672.620
Total of other links	-	229

Table 3: Interlinking from a sampled FOAF dataset to other datasets.

Although the intra-FOAF interlinking is high (in average, a single person is linked to 7 other persons), the interlinking between FOAF and other datasets is comparably low; some $2 * 10^{-}3$ interlinks per described person have been found. Also, the proportion of *indirect* links from a person to another (using `foaf:knows` and `rdfs:seeAlso`) is higher than *direct* links (through a single `foaf:knows`).

5 Conclusion

We have attempted to make a step towards answering the question: *What is the size of the Semantic Web*? in this paper. Based on a syntactic and semantic analysis of the LOD dataset we believe that answers can be derived for the entire Semantic Web. We have identified two different types of datasets, namely single-point-of-access datasets (such as DBpedia), and distributed datasets (e.g. the FOAF-o-sphere). At least for the single-point-of-access datasets it seems that automatic interlinking yields a high number of semantic links, however of rather shallow quality. Our finding was that not only the number of triples is relevant, but also how the datasets both internally and externally are interlinked. Based on this observation we will further research into other types of Semantic Web data and propose a metric for gauging it, based on the quality and quantity of the semantic links. We expect similar mechanisms (for example regarding automatic

interlinking) to take place on the Semantic Web. Hence, it seems likely that the Semantic Web as a whole has similar characteristics compared to our findings in the LOD datasets. Finally we return to the initial question: *What is the size of the Semantic Web?* In a nutshell, the answer is: just as the surface of a sphere is bounded but unlimited, the Semantic Web is.

Acknowledgement

The research leading to this paper was carried out in the "Understanding Advertising" (UAd) project[7], funded by the Austrian FIT-IT Programme, and was partially supported by the European Commission under contracts FP6-027122-SALERO and FP6-027026-K-SPACE.

References

[Broder et al. 00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 33(1-6):309–320, 2000.

[Gulli and Signorini 05] A. Gulli and A. Signorini. The Indexable Web is More than 11.5 Billion Pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, 2005.

[Ayers 07] D. Ayers. Evolving the Link. *IEEE Internet Computing*, 11(3):94–96, 2007.

[Esmaili and Abolhassani 06] K. S. Esmaili and H. Abolhassani. A Categorization Scheme for Semantic Web Search Engines. In *4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06)*, Sharjah, UAE, 2006.

[Finin et al. 05] T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In *AAAI 05 (intelligent systems demo)*, 2005.

[Ding and Finin 06] L. Ding and T. Finin. Characterizing the Semantic Web on the Web. In *5th International Semantic Web Conference, ISWC 2006*, pages 242–257, 2006.

[Tummarello et al. 07] G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the Open Linked Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 552–565, 2007.

[Ding et al. 05] L. Ding, L. Zhou, T. Finin, and A. Joshi. How the Semantic Web is Being Used:An Analysis of FOAF Documents. In *38th International Conference on System Sciences*, 2005.

[Wang 06] T. D. Wang. Gauging Ontologies and Schemas by Numbers. In *4th International Workshop on Evaluation of Ontologies for the Web (EON2006)*, 2006.

[Hausenblas et al. 07] M. Hausenblas, W. Slany, and D. Ayers. A Performance and Scalability Metric for Virtual RDF Graphs. In *3rd Workshop on Scripting for the Semantic Web (SFSW07)*, Innsbruck, Austria, 2007.

[Völkel et al. 06] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic Wikipedia. In *15th International Conference on World Wide Web, WWW 2006*, pages 585–594, 2006.

[Bizer et al. 07] C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In *4th European Semantic Web Conference (ESWC2007)*, pages 802–815, 2007.

[7] `http://www.sembase.at/index.php/UAd`