On the Topology of the Web of Data

Markus Luczak-Rösch and Robert Tolksdorf
Freie Universität Berlin, Institute of Computer Science, Networked Information Systems
Königin-Luise-Straße 24/26, Berlin D-14195, Germany
luczak@inf.fu-berlin.de | tolk@ag-nbi.de

ABSTRACT

The Web of Data consists of the open accessible structured data on the Web. This includes the evolving number of Linked Open Data data sets but also the structured data which is embedded in Web pages. In this paper we address questions related to a unified definition of distinct data sets and factors that influence different network representations of structured Web data. The contributions are (1) an algorithm to generate a data set linking structure of the embedded structured data sourcing from (a) the Billion Triples Challenge corpus (b) the Web Data Commons corpus, and (c) the sindice crawl, (2) a discussion on the issue of identifying distinct data sets in a generic fashion, and (3) a high level visual abstraction of the current Web of Data topology.

Categories and Subject Descriptors

H.4 [**Information Systems**]: World Wide Web; H.4.7 [**World Wide Web**]: Web data description languages

General Terms

Experimentation, Algorithms, Measurement

1. INTRODUCTION

The Web of Data consists of the open accessible structured data on the Web. This includes the evolving number of Linked Open Data data sets but also the structured data which is embedded in Web pages. **Understanding and analyzing the topology of the Web of Data is crucial for use cases such as data set ranking in the context of Web of Data search engines or asessing data quality measures like reputation before consuming the data in applications**. But such analysis can also help to understand the overall adoption and evolution of the Web of Data towards a serious global dataspace with a significant size and robust structure as studied in [1].

Recently the Web Data Commons project made a representative subset of the Web of Data available for the research community, thus studies in this area are timely. A number of research questions are: Are the well-known "Linking Open Data cloud diagram" and recent adaptations based on it[1] a realistic view of the Web of Data? What are the standard entities to represent nodes and edges in a network representation of it? Are there consistent and commonly agreed definitions for what a data set is which also hold for Web pages fomenting something one could call "virtual data sets" in contrast to the "physical data sets" served by Linked Data endpoints? Which characteristics of structured data distinguish the Web of Data from former studies representing the Web of documents as a network? Is it possible to discover structures and properties of the classical Web also within the Web of Data? Is it better not to differentiate between "two Webs" and analyze the structure integratively instead?

In this paper we report on our work studying the network topology of the Web of Data and contribute (1) an algorithm to generate a data set linking structure of the embedded structured data sourcing from (a) the Billion Triples Challenge corpus (b) the Web Data Commons corpus, and (c) the sindice crawl, (2) a discussion on the issue of identifying distinct data sets in a generic fashion, and (3) a high level visual abstraction of the current Web of Data topology. To evaluate our approach we generate two link networks from the WDC 2012 and BTC 2011 corpora and compute network properties, which are commonly consulted to determine whether a network is scale-free which would characteristically distinguish the Web of Data from random networks and as it was discovered for the Web of documents.

The remainder of this paper is structured as follows: In Section 2 we refer to the related work on complex network analysis applied to the Web and structured Web data and survey on different notions of data sets. A novel link extraction algorithm is introduced in Section 3. Afterwards we present the evaluation of our approach (Section 4) before we finish the paper with the conclusions drawn and perspectives for future work (Section 5).

2. FOUNDATIONS AND RELATED WORK

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France

[1] Please refer to `http://commit.wim.uni-mannheim.de/uploads/media/commitWorkshop_Bizer.pdf` (seen on 2012-12-18) for Bizers adaptation of the original Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch available at `http://lod-cloud.net/` (seen on 2012-12-18)

Analyzing complex networks helps to understand the characteristics and sometimes predict the evolution of a variety of networked systems such as social networks, biological networks, citation networks but also the network stucture of the Web. One can distinct the visual analysis of networks which suits well for a macroscopic level of abstraction and a variety of underlying statistical measures and network properties which can be computed in order to interpret global as well as fine grained structures. One of the most noteworthy analysis of the conncetivity of the Web is the "Bow Tie" topology by Broder et al. [2]. On the mathematical level the finding of Barabasi that the Web is a scale-free network – thus a network where the degree distribution follows a power-law – is one of the most important ones [3, 4].

Recent research concentrated on Linked Open Data (LOD) when analyzing the size and topology of the Web of Data [1, 5, 6]. LOD refers to open available RDF data conforming to the Linked Data principles[2]. One of the most representative subsets of the overall amount of Linked Open Data is the Billion Triples Challenge (BTC) corpus which is the reference data set for the yearly Semantic Web Challenge series since 2009[3]. It consists of a number of gzipped text files containing RDF-quads crawled by starting from a set of seed data sets and then continuously following RDF links found in the retrieved data. The BTC corpora foment the basis for a number of network analysis focusing the topology of the Web of Data.

The recently published Web Data Commons (WDC) corpora helped to steer the attention again to integrate the other part of the Web of Data – structured data which is embedded in Web pages as Microformats, RDFa or Microdata – by making it available as a set of gzipped text files containing RDF-quads extracted from the publicly available Commons Crawl corpus [7, 8]. Two versions of the WDC corpus are available originating from the Common Crawl corpora of 2009/2010 and 2012. Another representative corpus of that kind is the sindice crawl[4] which covers a timeperiod ranging from 2009 to 2011.

The most common and intuitive network representation of the Web of Data is the interlinkage structure of distinct data sets and the widely known visualization is the manually generated Linking Open Data cloud diagram[5] which also can be found in the center of the above mentioned adaptation by Bizer in Figure **??**. Generating such a visualization requires a clear definition of what a distinct data set is. The VoID vocabulary[6] provides a socio-technical definition of a data set as "a set of RDF triples that are published, maintained or aggregated by a single provider" [9].[7] It is possible to describe base URIs for a data set[8] with VoID in a machine processable fashion. However, we experienced that the best practices and guidelines around VoID are focused on "physical data sets", meaning data sets which are served by a Linked Data endpoint such as Pubby[9] or a triple store. There is a lack of guidelines which support a Web site provider to embed a VoID description which defines the borders of her "virtual data set".

The definition of a data set, as it was introduced in the VoID specification, was the basis for the data set ranking approach by Toupikov et. al [10] who generate a weighted network from VoID descriptions in order to apply a ranking algorithm in comparison to PageRank and HITS on such networks. Most recently the BTC and WDC projects analysed their corpora reducing resource URIs to their respective pay level domains (PLD) which one can interprete as another notion of a data set. It is obvious that the VoID definition of a data set only holds for data sets which conform the Linked Data principles and publish VoID descriptions but fails for most of the structured data in Web pages which do not provide this. However, also the generalization of the BTC and WDC projects to PLDs has its shortcomings since it is too restrictive assuming that at one PLD only one discrete data set is served. As a counter example one can look at the Linked Open Drug Data project and the DBLP mirror hosted at FU Berlin. All these data sets are served at subpaths of the `www4.wiwiss.fu-berlin.de` server (e.g. `http://www4.wiwiss.fu-berlin.de/sider/`, `http://www4.wiwiss.fu-berlin.de/dblp/`) which will all be reduced to the PLD `fu-berlin.de`. The situation gets even worse taking account each personal blog hosted at `[someuniquealias].blogspot.com` and service providers which differentiate between user spaces by URI paths instead of subdomains, e.g. `twitter.com/[uniqueusername]` for example. Each of these unique URI spaces is one distinct "virtual data set" if the maintainer of the account provides emebedded structured data. From a perspective of stability of URIs the PLD granularity may be an adequate level of abstraction but from a perspective of distinct data sets as reflected by the Linking Open Data Cloud diagram it fails to be representative. We constitute that it is hard to implement a generic algorithm which can identify distinct data sets automatically when no structured information about data set base URIs is given in Web pages.

3. GENERATING A LINKAGE STRUCTURE OF THE WEB OF DATA

As it was mentioned before in the context of the Web of Data the most commonly regarded entities are *data sets* which serve as nodes and *data links* connecting a resource from one data set to a resource within another serving as edges. We designed an algorithm that extracts the base URIs of all publicly known LOD data sets from the Data Hub repository and a set of links from one of the three corpora by WDC, BTC, and sindice. For all other URIs the simple subdomain granularity is applied. Consequently, the algorithm allows to compute a representation of all "physical" LOD data sets, all data sets which are rather "virtually" created from the embedded structured data, and all links which source from

[2] `http://www.w3.org/DesignIssues/LinkedData.html`

[3] `http://km.aifb.kit.edu/projects/[btc-2009, btc-2010,btc-2011,btc-2012]`

[4] `http://data.sindice.com/trec2011/download.html`

[5] `LinkingOpenDataclouddiagram, byRichardCyganiakandAnjaJentzsch.http://lod-cloud.net/`

[6] `http://www.w3.org/TR/void`

[7] Please note that there is no difference in meaning between the different spellings of "data set" or "dataset" and that we adopted the former one consistently while the VoID definition adopted the latter one.

[8] http://www.w3.org/TR/void/dataset-uris

[9] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`

the latter. The algorithm works as follows:

1. extract all data set URIs of Linked Open Data Cloud data sets as listed at The Data Hub in the group "lod-cloud" and store them in a database table "datasets"
2. access the extraction file list of BTC, WDC or sindice and initialize the URIs of all extraction files containing quads (WDC and BTC) or triples (sindice)
3. process all lines of the extraction files and create triples or quads using the Jena RIOT[10] parser
4. generalize the subject and the object to
   (a) a LOD data set base URI extracted in step 1 if there is a subpartial string overlap with one of the listed LOD data sets or otherwise
   (b) to subdomain granularity (e.g. xyz.example.org)
5. if the generated subject and object data set URIs are not equal, add subject, predicate, object, and graph (graph = "sindice" in case of processing the sindice extraction files) to the database table "links" complying with the unique constraint consisting of subject, predicate and object

4. EVALUATION

Our evaluation is intended to (1) demonstrate the functionality of our link extraction algorithm, (2) describe the characteristics of the resulting network representations, and (3) disclose general properties of the current state of the Web of Data and the RDFa subset as a specific part of it. We created two different link databases – one for the 2012 WDC corpus resulting in a total number of 2680692 distinct links and one for the 2011 BTC corpus resulting in a total number of 773487 links.

In the first network representation we generated from this data each node represents one distinct data set[11]. We only regard links between data sets which are published either in the source or the target data set of the link and not those which are published by third parties connecting two remote data sets. The simple reason for this is that we regard data set publishers as an authority for publishing links from or towards the own data. We do not argue about messy links at this point and think that our simple approach at least reduces the probability of those. Table 1 lists the number of overall nodes and edges for the two resulting networks as well as the number of nodes when only interlinked nodes are regarded which means that these nodes have a degree ≥ 1 (kcore= 1).

Computing the modularity of a network helps to detect communities or clusters of nodes. A total number of 122435 communities and an overall modularity of 0.896 allow the interpretation that nodes in different communities are rather sparsely connected. A huge number of nodes with a very low degree are opposed by a small number of nodes with a high degree, which is an indicator that the Web of Data is a scale-free network and further stressed when plotting the statistic on a double logarithmic scale (Figure 1) which yields a staight line.

Table 1: Network sizes of the link structure extracted from the WDC 2012 and BTC 2011 corpora in Gephi.

	WDC 2012	*BTC 2011*
# of nodes	1835909	620806
# of nodes with degree ≥ 1	1835698 (99.99%)	620578 (99.96%)
# of edges	2254269	668008

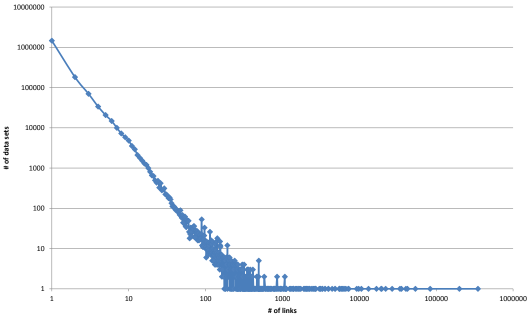

Figure 1: Degree distributon for our integrated network representation of the WDC corpus on a double logarithmic scale.

We created a second network out of the RDFa subset of the WDC corpus, which means all structured data which are embedded as RDFa within Web pages, in order to compare these two network representations. Table 2 lists the extraction statistics about distinct links and data sets, the number of overall nodes and edges as well as the number of nodes which have a degree ≥ 1.

Table 2: Statistics for the WDC 2012 RDFa subset[12].

# of LOD data sets	328
# of distinct data sets	157339
# of distinct links incl. links served by third party	225081
# of distinct links served by source or target data set	216313
# of nodes	157638
# of nodes with degree ≥ 1	157422 (99.86%)
# of edges	189653

In case of the WDC RDFa subset a total number of 1777 communities has been detected by computing the modularity of the network which is with a value of 0.662 rather high, even though it is significantly lower than the modularity of the integrated WDC network (0.896, as mentioned above). We conclude that the RDF data embedded in Web pages as RDFa is better integrated with other RDF data sets on the Web than the structured data shared as microdata or microformats. However, it is again statistically shown that the embedded structured data on the Web foments a network involving a number of sparsely connected sub-networks. Plotting the degree distribution on a double logarithmic scale does not yield the characteristic straight line (see Figure 2)

[10] `http://jena.apache.org/documentation/io/riot.html`

[11] To generate network representations and to compute common network metrics we apply a Web-based port of the Gephi (`http://gephi.org/`) network analysis and visualization tool.

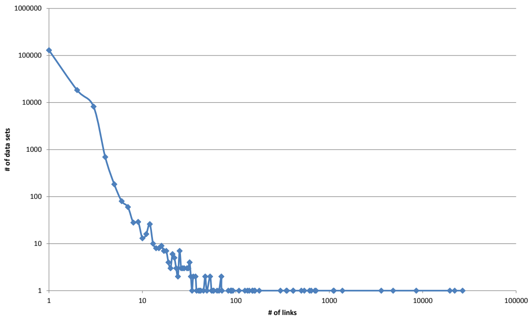

Figure 2: Degree distributon for the RDFa subset of the WDC corpus on a double logarithmmic scale.

as this was the case for the entire WDC network representation. That means that the RDFa subset of the WDC corpus cannot be clearly defined as a scale-free network.

We furthermore exported the degree and the reference whether the data set is a part of the LOD cloud of the data sets with a degree higher than 1000 as shown in Table 3[13] in order to determine whether some data sets have a special position within the network. The data set with the highest degree is DBpedia with an overall degree of 83431. Interestingly the in-degree of DBpedia massively outperforms the out-degree. It can be generally deduced that the difference between the in- or out-degree of all data sets is significantly high. The listed LOD data sets have a high out-degree, data sets which can be manually identified as vocabulary namespaces (such as `http://xmlns.com/` or `http://purl.org/`) have a high in-degree, but for all other data sets no regularity can be detected.

Table 3: Degree statistics of data sets in the WDC RDFa subset with a degree higher than 1000 and the respective modularity class id a data set has been assigned to.

Node id (data set base URI, "http://" stripped to save space)	**in-degree**	**out-degree**	**degree**
dbpedia.org/ *	34	83 397	83 431
xmlns.com/	26 712	1	26 713
purl.org/	21 866	123	21 989
rdfs.org/	19 516	0	19 516
www.w3.org/	8 393	137	8530
rdf.data-vocabulary.org/	4 823	0	4 823
www.n49.ca/	0	3 600	3 600
d1.scribdassets.com/	1 381	1	1 382
www.biologeek.com/	1	1 120	1 121
www.bbc.co.uk/ programmes/ *	2	1 105	1 107

With reference to our extraction algorithm this listing is an ambivalent result. The trivial PLD approach would never detect the BBC programmes data set (which is distinct from several other BBC data sets served at `http://www.bbc.co.uk/`) as an important data set which was possible thanks to involving the Data Hub listing. On the contrary the list of obvious vocabulary namespaces (`http://xmlns.com/`, `http://purl.org/`, `http://rdfs.org/`, `http://www.w3.org/`, or `http://rdf.data-vocabulary.org/` proves that still a smarter lookup of URIs that distinguish base URIs by URI paths instead of subdomains is necessary.

The interesting observation is, that DBpedia has the same degree distribution in the complete WDC network as it has in the RDFa subset. That means, DBpedia is not linked from data represented in any other format than RDFa. BBC programmes gains at least attention of 8 additional incoming links. Altogether this lets us conclude that the RDF data in the LOD cloud is not directly interlinked with the data shared as microformats or microdata on the Web.

5. CONCLUSIONS AND FUTURE WORK

In this paper we studied and discussed the analysis of network representations of the Web of Data. We introduced an algorithm which accesses the Data Hub repository in order to integrate the publicly available LOD data sets with the structured data crawled by the Web Data Commons project, the Billion Triples Challenge corpus, and the sindice engine. The algorithm also exploits the list of LOD data set URIs for a more accurate identification of distinct data sets. We mentioned that the notion of a data set is a fundamental factor for the network representation and analysis of the entire Web of Data. In this context one of the most obvious experience of our experiments is that it is necessary to embed additional structured information into Web pages which help to identify the borders of distinct data sets more fine grained than pay level domain or subdomain granularity allows it. So we claim for an extension of the documentation and guidelines around the VoID vocabulary to make VoID being easier applicable in the context of embedded structured data published by Web site providers.

5.1 An ambivalent scale-free characteristic of the Web of Data

We ran experiments on the WDC and the BTC corpora individually and provided a detailed analysis of the integrated WDC corpus as well as the RDFa subset of it. Combining the links extracted from more than one up to all three sources (WDC, BTC, and sindice) as well as the generation of network representations which leverage a different set of the dimensions which we introduced in this paper is a next step of our work. By today we can approve that our representation of the Web of Data is a scale-free network. That means that the Web of Data will further evolve following the principle of preferential attachment [3]. However, our experiments resulted that this characteristic does not hold for the RDFa subset of the Web of Data.

5.2 Towards an integrated view on the Web of documents and the Web of Data

A well-prepared visualization of large-scale networks can help for didactic purposes on a macroscopic level as it was shown by the "Bow Tie" representation of the Web [2]. Due to the huge amount of data, the lack of a proper notion of distinct data sets, and a missing insight into the adequate level of abstraction we were not able to generate a visualization of the integrated Web of Data which is as representative as the "Bow Tie". However, we created a meaningful force directed layout of the WDC RDFa subset network which allows us to deduce the following interpretations[14]: (1) The RDFa subset of the Web of Data consists of a core network of

[13] Data sets marked with a * in the table are LOD cloud data sets listed in the Data Hub repository.

[14] We applied the Yifan Hu layout algoithm followed by the Force Atlas 2 layout algorithm (both in their Gephi implementation). Due to the fine granularity of this large scale network it does not make sense to publish it in this paper. You can ac-

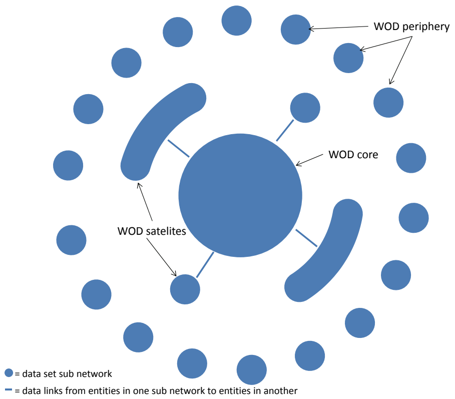
WOD periphery
WOD core
WOD satelites
● = data set sub network
– = data links from entities in one sub network to entities in another

Figure 3: High level abstraction of the RDFa subset of the WDC corpus into WOD core, WOD satelites, and WOD periphery.

connected data sets. (2) A number of data set exist around the core which link to it but are not that deep interlinked as the core itself. (3) There is a large periphery of small data set networks which are not connected to the core. Figure 3 depicts our high level abstraction for the RDFa subset of the WDC corpus we generated as a result of this interpretation. We distinguish the Web of Data core (WOD core), the Web of Data satelites (WOD satelites), and the Web of Data periphery (WOD periphery). The WOD core is the heavily interlinked set of LOD data sets and the vocabularies applied for Linked Data publication as well as a small number of data embedding pages. The WOD satelites are those Web pages embedding structured data which exploit the resource URIs of the WOD core as reference URIs for things described in their respective contents. And the WOD periphery is the disconnected rest of Web pages which only apply data links to refer to other Web site in the WOD periphery.

The network representations generated for this early stage report on our work have several shortcomings. We extracted the linking structure in a very trivial but generic fashion which keeps the resulting network not free of links which better should not be treated as “data links”. Some examples are RDF links to images or other media documents on the Web (like Web resources referred to by properties such as `http://ogp.me/ns#image`). In a next step we will come up with a heuristic approach to be more precise which links can be treated as real “data links”. Furthermore we will implement a much more sophisticated data set identification approach which exploits URI similarity calculation as described in [11] and looks up further information about data set URIs within the data itself but also about ontology namespaces within respective sources on the Web, such as the vocab.cc or the LOV services. A study on the performance of the different approaches to find the most precise one will be a valuable contribution.

The aforementioned observation – data links pointing to classical information resources such as images – is an issue at a first sight. But thinking further this could also stimulate discourse about the question if it is necessary to distinguish the Web into a classical Web of ducuments and a Web of Data and analyze them separatly. We think it is interesting to see how it is possible to browse not only through information resources but also to switch over to non-information resources when both charateristic parts of the Web are analyzed integratively. That would open further questions if it makes sense that future Web browsers also visualize the embedded links to structured data to allow the user to navigate along them and retrieve information she would not have retrieved when browsing through the classical Web of documents only. The following figure drafts a visual model how classical hyperlinks connect the WOD periphery introduced in the last section to the WOD core, mediated by the WOD satelites, resolving the disconnection of these components on the level of data links.

6. REFERENCES

[1] Guéret, C., Groth, P.T., van Harmelen, F., Schlobach, S.: Finding the achilles heel of the web of data: Using network analysis for link-recommendation. In Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., 0007, L.Z., Pan, J.Z., Horrocks, I., Glimm, B., eds.: International Semantic Web Conference (1). Volume 6496 of Lecture Notes in Computer Science., Springer (2010) 289–304

[2] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Comput. Netw. **33** (June 2000) 309–320

[3] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439) (1999) 509–512

[4] Barabási, A.L., Bonabeau, E.: Scale-free networks. Scientific American **288**(60-69) (2003)

[5] Guéret, C., Wang, S., Schlobach, S. In: The Web of Data is a Complex System - first insight into its multi-scale network properties. Volume D. (2010) 1–12

[6] Guéret, C., Wang, S., Groth, P., Scholbach, S.: Multi-scale analysis of the web of data: A challenge to the complex system's community. Advances in Complex Systems **14**(04) (2011)

[7] Mühleisen, H., Bizer, C.: Web Data Commons - Extracting Structured Data from Two Large Web Corpora. In: WWW 2012 Workshop: Linked Data on the Web (LDOW2012), Lyon, France (2012)

[8] Mika, P., Potter, T.: Metadata statistics for a large web corpus. In Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M., eds.: LDOW. Volume 937 of CEUR Workshop Proceedings., CEUR-WS.org (2012)

[9] Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In: WWW 2009 Workshop: Linked Data on the Web (LDOW2009), Madrid, Spain (2009)

[10] Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., Tummarello, G.: DING! Dataset Ranking using Formal Descriptions. (2009)

[11] Tim Bray: How to compare uniform resource identifiers. http://www.textuality.com/tag/uri-comp-2.html Visited on December 14th, 2012.

cess this visualization at `https://dl.dropbox.com/u/60766512/wod-rdfa-subset-yifan500-fa2-500.pdf`