

# Ranking the Web of Data

José M. Giménez-García, Harsh Thakkar, Antoine Zimmermann

<sup>1</sup> Univ Lyon, UJM-Saint-Etienne, CNRS,  
Laboratoire Hubert Curien UMR 5516, F-42023 Saint Etienne, France  
`jose.gimenez.garcia@univ-st-etienne.fr`

<sup>2</sup> Enterprise Information Systems Lab,  
University of Bonn, Germany  
`hthakkar@uni-bonn.de`

<sup>3</sup> École Nationale Supérieure des Mines, FAYOL-ENSMSE,  
Laboratoire Hubert Curien, F-42023 Saint-Étienne, France  
`antoine.zimmermann@emse.fr`

**Abstract.** While a number of quality metrics have been successfully proposed for datasets in the Web of Data, there is a lack of trust metrics that can be computed for any given dataset. We argue that reuse of data can be seen as an act of trust. In the Semantic Web environment, datasets regularly include terms from other sources, and each of these connections express a degree of trust on that source. However, determining *what* is a dataset in this context is not straightforward. We use the concept of Pay-Level Domain to differentiate datasets, and consider usage of external terms as connections among them. Using these connections we compute the PageRank value for each dataset. This process has been performed for more than 300 datasets, extracted from the LOD Laundromat.

**Keywords:** linked data, trust, reuse, interlinking, PageRank, quality metric, quality assessment

## 1 Introduction

The WDAqua project<sup>4</sup> aims to advance the state of the art in data-driven question answering, with a special focus in the Web of Data. The Web of Data comprises thousands of datasets about varied topics, interrelated among them, which contain large quantities of relevant data to answer a question. Nonetheless, in an environment of information published independently by many different actors, data veracity is usually uncertain, and there is always the risk of consuming misleading data. While some quality metrics have been proposed that can help to identify good datasets [1], there is a lack of trust metrics to provide a confidence on the veracity of the data.

In this context, we argue that actual usage of data can be seen as an act of trust. In this paper we focus on reuse of resources by other datasets as an usage metric. We consider reuse of a resource of a dataset by any other given

---

<sup>4</sup> <http://wdaqua.informatik.uni-bonn.de>

dataset as an outlink from the later to the former. Under this purview, we can compute the PageRank [2] value of each dataset and rank them according to their reuse. In order to obtain a good measure of reuse, we perform the process on a large scale. We make use of the tools provided by the LOD Laundromat [3] to go beyond LOD Cloud, and process more than 38 billion triples, distributed in more than 600 thousand documents. The LOD Laundromat provides data from data dumps collected from the Internet, so it is not limited to dereferenceable linked data. However, what is regarded as a dataset is an important issue when dealing with data dumps. We make use of the concept of Pay-Level Domain (or PLD, also known as Top-Private Domain) to draw a distinction between datasets. We perform an aggrupation of the triples in 319 datasets and compute their PageRank values as a first measure of trust.

This document is organized as follows: in Section 2, we first discuss what should be considered a dataset in our context in order to clarify the problem we address; in Section 3 we present the tools we are using, namely the LOD Laundromat and the PageRank algorithm; Section 4 describes the experiments and results, which we further discuss in section Section 2; Section 5 presents relevant related work; finally, we provide some conclusions and directions for future work in Section 6.

## 2 Ranking the Web of Data

We would like to assess trust in datasets by measuring their popularity based on the reuse of resources from a dataset in another dataset. To do this, we rely on the PageRank algorithm (presented in more detailed in Section 3). To compute PageRank in a set of datasets, it is first necessary to define what is considered a dataset and what is a link between datasets. RDF graphs, although formally defined as a set of triples, can be seen as directed multigraphs in which predicates play the role of arcs. This view suggests that if a triple contains a resource of dataset  $A$  as subject, and a resource of dataset  $B$  as object, it can be seen as a link from dataset  $A$  to dataset  $B$ . However, the links formed by arcs in an RDF graph are irrelevant to the notion of dataset linking. In fact, only the presence of hyperlinks suffices to indicate a link between one source and destination, therefore any HTTP IRI in an RDF graph can be seen as a link. So the question is, what it means that a resource belongs to a dataset, and to what dataset a hyperlink “points to”. A naïve approach would be to consider that any IRI existing in a dataset belongs to the dataset and thus, that links connect two datasets having one same resource. However, this would imply, for instance, that any triple anywhere that uses a DBpedia IRI is considered to be linked to from the DBpedia dataset. As a result, any dataset that reuses a DBpedia IRI would increase their PageRank according to this definition.

Alternatively, we could take advantage of the linked data principles which stipulate that IRIs should be addresses pointing to a location on the Web. Again, one could naïvely assume that the location that the address points to is what defines the dataset, that is, the document retrieved when one gets the resource

using the HTTP protocol. However, this would lead us, for instance, to define each DBpedia article as an individual dataset.

A second possibility would be to use the domain part of the URL. This way, it would be easy to determine what dataset is being linked to. Such approach would work well if all datasets were accessible from dereferenceable IRIs. However, there are large portions of the Web of Data that provide access to data dumps only [4, 5]. In this case, the domain of the dump does not necessarily match the domain of the individual IRIs found in the dataset. As an example, the DBpedia dumps are found at <http://downloads.dbpedia.org/> while all DBpedia IRIs start with <http://dbpedia.org/>.

Finally, we decided to use the the concept of *PLD*, *i.e.*, the subdomain component of a URL followed by a public suffix, to identify a dataset. This approach has already been used by other works [6, 7]. As an example, if a file found at <http://download.dbpedia.org/> contains the following triple:

```
<http://dbpedia.org/wiki/Europe>  
  <http://www.w3.org/2002/07/owl#sameAs>  
    <http://sws.geonames.org/6255148/>
```

we will consider that the dataset having the PLD `dbpedia.org` is linking to the dataset with PLD `geonames.org`. It is important to notice that the source of the link (`dbpedia.org`) is obtained from the URL of the document that contains the triple (<http://download.dbpedia.org/>), not from the subject of the example.

**Definition 1 (Dataset).** *A dataset is a non empty collection of triples that can be retrieved from a source accessible at a URL having a common Pay-Level Domain. The PLD identifies the dataset.*

In the previous example, we see that the predicate IRI is linking to the standard OWL vocabulary. It is very likely that predicates in general will be linking to vocabularies that are extensively reused. However, our intent is to evaluate trust on factual data rather than terminological definitions. For this reason, we do not take into account the predicates in our computation.

**Definition 2 (Dataset link).** *There exists a link from a dataset A to a dataset B if and only if there exists a triple in a file at a location having the PLD that identifies A in which the PLD of its subject, its object, or both matches the PLD that identifies B.*

This definition is in line with the PageRank algorithm [2] where the number of links between the same two nodes is irrelevant. Note that since datasets must be non empty, links to PLDs that do not host RDF have to be ignored. In the next section, we describe the tools that we used in our experiments.

### 3 Preliminaries

In order to provide a realistic assessment of reuse in Linked Open Data, we exploited a large number of datasets by way of the LOD Laundromat (described

in Section 3.1) from which we compute the dataset links that form the input of the PageRank algorithm (described in Section 3.2).

### 3.1 The LOD Laundromat and Frank

The LOD cloud<sup>5</sup>, and in general Linked Open Data, contains a wide variety of formats, publishing schemes, errors, that make it difficult to perform a large-scale evaluation. Yet, to be accurate, our study requires to be comprehensive. Fortunately, the LOD Laundromat [8, 9] makes this data available by gathering dataset dumps from the Web, including archived data. The LOD Laundromat “cleans” the data by fixing syntactic errors and removing duplicates, and then makes it available through download (either as gzipped N-Triples or N-Quads, or HDT [10] files), a SPARQL endpoint, and Triple Pattern Fragments [11].

Frank [3] is a command-line tool which serves as an interface of the LOD Laundromat, and makes it easy to run evaluations against very large numbers of datasets.

### 3.2 PageRank

PageRank [2] is the original algorithm developed by Page et al. that Google uses to rank their search results. It takes advantage of the graph structure of the web, considering each link from one page as a “vote” from the source to the destination. Using the links, the importance of a page is propagated across the graph, dividing the value of a page among its outlinks. This process is repeated until convergence is reached. The final result of PageRank corresponds to a stationary distribution, where each page value amounts to the probability for a random surfer to be at any moment in the page.

## 4 Experiments and Results

The process to compute PageRank involves the following steps, detailed further below.

1. Extracting the document list from LOD Laundromat.
2. Parsing the content of each document to extract the outlinks.
3. Consolidating the results
4. Computing PageRank

### 4.1 Extracting the document list from LOD Laundromat

We use the Frank command line tool [3] to obtain a snapshot of the contents of the LOD Laundromat. While the output of Frank can be directly pipelined to our process, the next step is performed in parallel in several machines. For

---

<sup>5</sup> <http://lod-cloud.net/>

this reason, we need that every machine reads the exact same input. An update in the contents of the LOD Laundromat during the next process could have impacted the results in that case. We retrieve the list of documents in the LOD Laundromat with the following command.

```
$ frank documents > documents.dat
```

This command retrieves a list of pairs (*downloadURL-resourceURL*), where the first is the URL to download the gzipped datasets, and the second the resource identifier in the LOD Laundromat ontology. At the moment of the experiments, it retrieved 649,855 documents.

## 4.2 Parsing the content of each document to extract the outlinks.

A prototype tool<sup>6</sup> has been developed to stream the contents of the documents and extract the outlinks. This tool reads the list of pairs (*downloadURL-resourceURL*) by standard input, and accepts two optional parameters for partial: *Step* and *Start*. The first one tells how many lines the process reads in every iteration, processing the last one, while the second denotes what line to use for the first input. For each line processed, it queries the SPARQL endpoint to retrieve the URL where that dataset was crawled. This information can be found in the LOD Laundromat ontology connected to the resource, in the case the document was crawled as a single file, or connected to the archive that contains the document, if it was crawled compressed in a compressed file, possibly along other documents. In the first case, we retrieve the URL with Query 1, in the second case we retrieve the URL using Query 2, where %s is substituted by the *resourceURL*. The Pay-Level Domain is then extracted and stored. This will be considered as the identifier of the dataset.

```
SELECT ?url
WHERE {<%s> <http://lodlaundromat.org/ontology/url> ?url}
```

**Query 1:** Query to retrieve crawled URL of a non-archived document

Then, the gzipped file is streamed from the *downloadURL* and parsed the triples. The subject and object (in case it is a URI) are extracted the Pay-Level Domain and compared against their dataset PLD. If they have a valid PLD and is different from their dataset's Pay-Level Domain, the pair (*datasetPLD-resourcePLD*) is stored as an outlink for the dataset. The output of each dataset is stored in a different file, which will be appended more pairs if a different document is identified as the same dataset.

---

<sup>6</sup> <https://github.com/jm-gimenez-garcia/LODRank>

```

SELECT ?url
WHERE {
    ?archive <http://lodlaundromat.org/ontology/containsEntry> <%s> .
    ?archive <http://lodlaundromat.org/ontology/url> ?url
}

```

**Query 2:** Query to retrieve crawled URL of an archived document

This process makes use of Apache Jena<sup>7</sup> v3.0.1 to query the SPARQL endpoint of the LOD Laundromat and Google Guava<sup>8</sup> v19.0 to extract the Pay-Level Domain of the datasets.

In the experiments the process was launched in parallel in 8 different virtual machines using Google Cloud Platform<sup>9</sup> free trial resources, each one processing a different subset of the list downloaded in the previous step. A statistical description of the results of each process is detailed in table Table 1. “Documents” correspond to the number of dump files in the LOD Laundromat, while “Datasets” are the number of PLDs that the process is dealing with. There can overlap in the datasets of several processes, so the total number of datasets is not equal to the sum. We can see that the number of triples processed by each process is not proportional to the number of documents processed.

Process	Documents	Triples	Datasets
1	81,220	3,994,446,393	121
2	81,226	3,742,870,561	118
3	83,422	4,146,249,367	127
4	81,225	3,376,784,600	120
5	81,225	3,623,413,245	120
6	88,198	3,377,773,585	116
7	81,226	4,132,960,522	115
8	89,781	3,911,917,919	123

**Table 1:** Data extracted from the LODLaundromat by each process

### 4.3 Consolidating the results

Once the outlinks have been extracted, the different files have to be appended and duplicated removed. In the experiments, the data from each virtual machine was downloaded in a separate folder of a unique machine. Then files with the same name in each folder were concatenated and removed the duplicates using

<sup>7</sup> <https://jena.apache.org>

<sup>8</sup> <https://github.com/google/guava>

<sup>9</sup> <https://cloud.google.com>

a bash shellscript. The total number of datasets after consolidating the results is 319. The result was again concatenated in a single file.

#### 4.4 Computing PageRank

For PageRank computation we make use of the `igraph` R package [12].

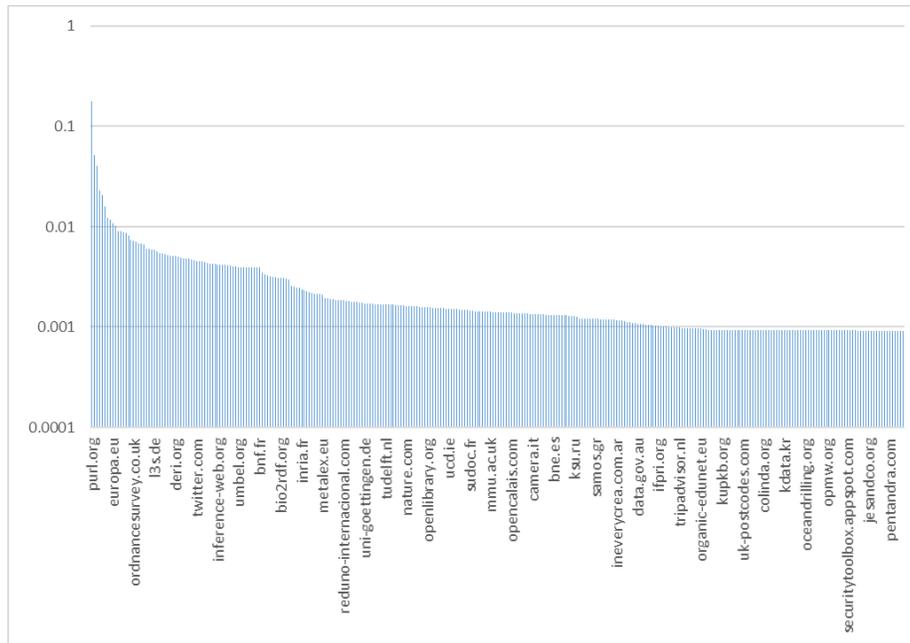
Rank	Dataset	PageRank value
1	<code>purl.org</code>	0.177783059
2	<code>creativecommons.org</code>	0.051508485
3	<code>dbpedia.org</code>	0.040423981
4	<code>rdfs.org</code>	0.022861256
5	<code>geonames.org</code>	0.020872769
6	<code>loc.gov</code>	0.01596783
7	<code>fao.org</code>	0.012239963
8	<code>holygoat.co.uk</code>	0.01182691
9	<code>europa.eu</code>	0.010810562
10	<code>data.gov.uk</code>	0.010097518

**Table 2:** PageRank results for the extracted datasets

Table 2 shows the 10 highest ranked datasets together with their PageRank. Not so surprisingly, `purl.org` is first because it is a popular way of providing persistent URLs that redirect to various data owners. We observe that the PageRank value is quickly decreasing. `creativecommons.org` is often referenced when a dataset has an explicit license attached to it. `dbpedia.org` is, as expected, high in the ranking. `rdfs.org` is most well known for hosting the SIOC ontology<sup>10</sup>, often used to describe online communities, forums, blogs, and the likes. `geonames.org` provides reference geospatial data; geolocated data are quite common in Linked Open Data. `loc.gov` hosts bibliographical data from the United States Library of Congress. `fao.org` is the domain of the Food and Agriculture Organization of the United Nations that not only hosts data about food and agriculture in many countries but also geopolitical data. `holygoat.co.uk` is a real surprise: it is the only domain in the top 10 owned by an individual person, Richard Newman, a computer scientist who wrote several ontologies in the early days of the Semantic Web. `europa.org` is the official portal of the European Union. `data.gov.uk` is the main British open data platform.

*Discussion* Here we provide additional information about the datasets, especially the top-ranked ones, in order to understand how ranking correlates with other statistical values, such as number of triples, number of documents. We also discuss how our own choices influenced the results.

We used two queries, (Query 3 and Query ??), to obtain the number of documents and triples for each PLD, from the LOD Laundromat.



**Fig. 1:** PageRank values in decreasing order

```

PREFIX llo: <http://lodlaundromat.org/ontology/>
PREFIX ll: <http://lodlaundromat.org/resource/>
SELECT (COUNT(DISTINCT ?resource) AS ?count)
WHERE {
  {
    ?resource llo:url ?url
    FILTER regex(?url, "[^/\\.]*\\.?.?s/", "")
  }
  UNION
  {
    ?archive llo:containsEntry ?resource ;
      llo:url ?url
    FILTER regex(?url, "[^/\\.]*\\.?.?s/", "")
  }
}

```

**Query 3:** Query to retrieve the number of documents per dataset

```

PREFIX llo: <http://lodlaundromat.org/ontology/>
PREFIX ll: <http://lodlaundromat.org/resource/>
SELECT (COUNT(DISTINCT ?resource) AS ?count) (SUM(?triples) as ?sum)
WHERE {
  {
    ?resource llo:url ?url ;
      llo:triples ?triples
    FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
  }
  UNION
  {
    ?archive llo:containsEntry ?resource ;
      llo:url ?url .
    ?resource llo:triples ?triples
    FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
  }
}

```

**Query 4:** Query to retrieve the number of documents with triples and number of triples

The result of the queries are given in Table 3 for the 10 top-ranked PLDs.

Rank	Dataset	Documents	Documents with triples	Triples
1	purl.org	9,166	9,135	254,548,441
2	creativecommons.org	1	1	117
3	dbpedia.org	1,888	1,823	1,257,930,891
4	rdfs.org	6	6	1,808
5	geonames.org	6	4	9,762
6	loc.gov	16	12	263,653,979
7	fao.org	17	16	48366
8	holyoat.co.uk	1	1	95
9	europa.eu	7734	7733	3,414,066,228
10	data.gov.uk	157	103	51,401,490

**Table 3:** Documents and trilpes per dataset in LOD Laundromat

As we can see, popularity is not at all correlated with the size of the datasets. Indeed, one of the top ten dataset only has 95 triples and the second overall most popular dataset of the LOD has 117 triples, while `dbpedia.org` and `europa.eu` both have billions of triples.

The enormously high page rank of `purl.org` should be mitigated by the fact that `purl.org` does not actually host any data. It is a redirecting service

<sup>10</sup> <http://rdfs.org/sioc/spec/>

that many data publishers are using. This result highlights a drawback in our heuristic for identifying datasets: the PLD is not always referring to a single dataset. To overcome this particular case, we could consider the PLD of the URL of the document obtained after dereferencing the IRI.

Another possible drawback of the approach is that triples with `rdf:type` in predicate position have their object pointing to a class in an ontology. This is in contradiction with our remark in Section 2 where we say that we want to rank instance data rather than terminological knowledge. This can have a major impact the results since `purl.org` is most often used to redirect to vocabularies more than datasets, and `rdfs.org` only hosts ontologies.

## 5 Related work

The authors of Semantic Web Search Engine (SWSE [6]) strongly advocate that the use of a ranking mechanism is very crucial for prioritizing data elements in the search process. Their work is inspired by the Google PageRank algorithm, which treats hyperlinks to other pages as a positive score. The PageRank algorithm is targeted for hyperlink documents and its adaptation to the LOD is however non-trivial, as we have seen. They point out that the primary reason for this is that LOD datasets may not have direct hyperlinks to other datasets but rather in most cases make use of implicit links to other web pages via the re-use of dereferenceable URIs. Here the unit of search becomes the entity and not the document itself. The authors briefly re-introduce the concept of naming authority, from their previous work [13] in order to rank structured data from an open distributed environment. They assume that the naming authority should match the Pay-level domain such that computing PageRank is performed on a naming authority graph where the nodes are PLDs. Their intuition therefore is in accordance with our reasoning from Section 2. They have discussed and contrasted the interpretation of naming authorities on a document level (e.g. <http://www.danbri.org/foaf.rdf>) and a PLD level (`danbri.org`). Also, they make use of a generalization for the method discussed in the paper [14] for ranking entities and carry out links analysis on the PLD abstraction layer.

The authors of Swoogle [15] develop OntoRank algorithm in order to rank documents. OntoRank, a variation of Google PageRank, is an iterative algorithm for calculating the ranks for documents built on references to terms (i.e., classes and properties) which are defined in other documents.

In the paper [16], the authors calculate the rank of entities (or as they call them objects) based on the logarithm of the number of documents where that particular object is mentioned.

In their work [17] present LinkQA, an extensible data quality assessment framework for assessing the quality of linked data mappings using the network measures. For this, they assess the degree of interlinking of datasets using five network measures, out of which two network measures are specifically designed for Linked Data (namely, Open Same-As chains and Description Richness) and the other three standard network measures (namely, degree, centrality, and the

clustering coefficient) in order to assess variation in the quality of the overall linked data with respect to a certain set of links.

In [18], PageRank is used to compute a measure that is in turn associated to individual statements in datasets for the purpose of incorporating trust in reasoning. Therefore, as in our own approach, they consider that PageRank is an indication of trustworthiness. However, they only compute PageRank on a per document basis, and report on the PageRank values of the top 10 documents obtained from their web crawl.

## 6 Conclusion & Future work

Data-driven question answering, the aim of project WDAqua mentioned in the introduction to this paper, requires quality data in which one can trust. Our aim has been to provide a trust measure based on dataset interlinking, assuming Pay-Level Domains were identifying a unique dataset in order to compute PageRank on them. Our results show that the choice of such definition of dataset is questionable, as some PLDs group many data dumps. In order to improve this, we could associate well known datasets to IRI patterns, such as `it.dbpedia.org` for the Italian version of DBpedia.

In addition, we also intend to explore further applications of PageRank that may be useful for question answering. User interaction that provides trust values in a number of dataset could be used to compute PageRank values with those datasets as a teleport set, as suggested by Gyöngyi et al. [19]. Also, Topic-Sensitive PageRank [20] could help a question-answering system to select different datasets when a question is identified to belong to a specific topic.

Finally, this work is part of a broader objective that we want to pursue: to ascertain the relationship between the perceived trust on a dataset and its objective quality. We will explore this area in a future work where other data reuse metrics will be considered and compared against different quality metrics.

## 7 Acknowledgement

This project is supported by funding received from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 642795.

## Bibliography

- [1] Harsh Thakkar, Kemele M. Endris, Jose M. Giménez-García, Jeremy De-battista, Christoph Lange, and Soren Auer. Are linked datasets fit for open-domain question answering? a quality assessment. 2016.
- [2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.
- [3] Laurens Rietveld, Wouter Beek, and Stefan Schlobach. LOD lab: Experiments at LOD scale. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, pages 339–355, 2015.
- [4] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *J. Web Sem.*, 14:14–44, 2012.
- [5] Ivan Ermilov, Michael Martin, Jens Lehmann, and Sören Auer. Linked open data statistics: Collection and exploitation. In *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*, pages 242–249, 2013.
- [6] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stephan Decker. Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. *Journal of Web Semantics*, 9(4), 2011.
- [7] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 245–260, 2014.
- [8] Wouter Beek, Laurens Rietveld, Hamid R. Bazoobandi, Jan Wielemaker, and Stefan Schlobach. LOD laundromat: A uniform way of publishing other people’s dirty data. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 213–228, 2014.
- [9] Laurens Rietveld, Ruben Verborgh, Wouter Beek, Miel Vander Sande, and Stefan Schlobach. Linked data-as-a-service: The semantic web redeployed. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, pages 471–487, 2015.
- [10] Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF Representation for Publication and Exchange (HDT). *Journal of Web Semantics*, 2013. ISSN 1570-8268. URL <http://dataweb.infor.uva.es/wp-content/uploads/2013/01/jws2013.pdf>.
- [11] Ruben Verborgh, Miel Vander Sande, Pieter Colpaert, Sam Coppens, Erik Mannens, and Rik Van de Walle. Web-scale querying through linked data

- fragments. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, 2014. URL [http://ceur-ws.org/Vol-1184/ldow2014\\_paper\\_04.pdf](http://ceur-ws.org/Vol-1184/ldow2014_paper_04.pdf).
- [12] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.
  - [13] Andreas Harth, Sheila Kinsella, and Stefan Decker. Using naming authority to rank data and ontologies for web search. In *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pages 277–292, 2009.
  - [14] Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari. Finding and ranking knowledge on the semantic web. In *The Semantic Web-ISWC 2005*, pages 156–170. Springer, 2005.
  - [15] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 652–659. ACM, 2004.
  - [16] Gong Cheng and Yuzhong Qu. Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. *International Journal of Semantic Web and Information Systems*, 5(3):49–70, 2009.
  - [17] Christophe Guéret, Paul T. Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pages 87–102, 2012.
  - [18] Piero A. Bonatti, Aidan Hogan, Axel Polleres, and Luigi Sauro. Robust and scalable Linked Data reasoning incorporating provenance and trust annotations. *Journal of Web Semantics*, 9(2):165–201, 2011.
  - [19] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan O. Pedersen. Combating web spam with trustrank. In *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3 2004*, pages 576–587, 2004.
  - [20] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference, WWW 2002, May 7-11, 2002, Honolulu, Hawaii*, pages 517–526, 2002.