Dataset Reuse: An Analysis of References in Community Discussions, Publications and Data

Kemele M. Endris[1], José M. Giménez-García[2], Harsh Thakkar[3], Elena Demidova[1],
Antoine Zimmermann[4], Christoph Lange[3,6], Elena Simperl[5]
[1]L3S Research Center, Leibniz University of Hannover, Germany {endris,demidova}@L3S.de
[2]UJM-Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
[4]Uni Lyon, MINES Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
[3]Universit,y of Bonn, Germany , [5]University of Southampton, UK, [6]Fraunhofer IAIS, Germany

ABSTRACT

Following the Linked Data principles means maximising the reusability of data over the Web. Reuse of datasets can become apparent when datasets are linked to from other datasets, and referred in scientific articles or community discussions. It can thus be measured, similarly to citations of papers. In this paper we propose dataset reuse metrics and use these metrics to analyse indications of dataset reuse in different communication channels within a scientific community. In particular we consider mailing lists and publications in the Semantic Web community and their correlation with data interlinking. Our results demonstrate that indications of dataset reuse across different communication channels and reuse in terms of data interlinking are positively correlated.

CCS CONCEPTS

• **Applied computing** → *Document management and text processing*;

1 INTRODUCTION

The number of datasets publicly available within the Linked Open Data (LOD) cloud more than tripled between 2011 and 2017, from 295 in September 2011 to 1146 in January 2017. Despite this growth, the interlinking, and, speaking more generally, *reuse* of the Web of Data remains limited, and is often focused on few well-known reference datasets, such as DBpedia [5] and YAGO [9]. In recent years a lot of research has focused on conformance of published datasets to the Linked Data best practices [4, 8], dataset profiling [2], scientific impact of published datasets [3] and different aspects of quality evaluation for Linked Data [10]. Less insights are available with respect to dataset reuse.

Dataset reuse leaves traces in different communication channels within scientific communities: by mentions of datasets in publications and mailing lists, and by means of citations of dataset papers. Further indications of reuse include references to data instances and vocabulary reuse that can be observed within other linked datasets. The aim of this paper is to quantify the correlation across the dataset references in different communication channels at the example of the Semantic Web community as well as dataset interlinking as an indication of dataset reuse. To better understand different indications of dataset reuse, we analyse dataset mentions in publications that appear in the proceedings of the key Semantic Web and Web conferences such as ISWC, ESWC and WWW in the time frame from 2007 to 2015, mailing list discussions about the datasets on the *public-lod@w3.org* and the *semantic-web@w3.org* mailing lists, citations of dataset papers published in the Linked Dataset Description track of the Semantic Web Journal [3] as well as mutual reuse of datasets in terms of interlinking. We propose metrics that estimate dataset reuse using these references.

Our contributions are: 1) We define metrics to estimate dataset reuse by a scientific community using reuse indications in different communication channels; 2) We analyse dataset reuse behaviour within the Semantic Web community over nine years applying these metrics to a large-scale collection of 1131 linked datasets.

2 PROBLEM STATEMENT & METHODOLOGY

The aim of this paper is to quantify dataset reuse indications across different sources in a scientific community and their correlations. Mentioning a dataset, either in a scientific publication or a mailing list, can potentially signal a usage or an interest by the authors. Hence, we rely on citations as an indicator of possible reuse. In addition, we measure the actual reuse of the terms of a dataset by others when creating their datasets. To facilitate this analysis we perform the following steps: (1) We define dataset reuse metrics relying on different information sources such as: i) publications and communication channels within a scientific community, and ii) data interlinking. (2) We compute reuse metrics on LOD datasets and analyse the correlations between these metrics.

2.1 Mentions in publications and mailing lists

Dataset mention. In the context of scientific publications and mailing list discussions, a dataset can be referenced by its metadata including its name, URI, etc. We call a reference to a dataset ds in a document d using (some of) its metadata a *mention*. Here, a document can be a scientific publication or a mailing list thread. We model a dataset mention in a document as a binary relation R_M where $(d, ds) \in R_M$ if and only if the dataset ds is mentioned in the document d. We consider mailing list posts at the granularity of threads, i.e. define $(d, ds) \in R_M$ if there exists an email within the thread mentioning the dataset. To determine dataset mentions, we build a dataset dictionary including dataset metadata (obtained

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3154461

from *datahub.io*), and match and disambiguate dataset metadata identified in documents against this dictionary, as discussed below[1].

Dataset popularity. We estimate dataset reuse using the number of dataset mentions in publications and mailing lists. This is done by computing the proportion of documents in a given corpus that mention the dataset. We call this measure dataset popularity *ds*. Formally, given a collection of documents *D*, e.g., a proceedings volume or a mailing list archive, we compute *ds* as follows:

$$popularity(ds, D) = \frac{\#\{(d, ds) \in R_M | d \in D\}}{\#D}. \qquad (1)$$

Extraction and disambiguation of dataset mentions. Datasets are not referenced in publications in a standardised way [1]. Dataset properties as they appear in the full text of publications or mailing lists can be ambiguous. In our work the dataset mention must be uniquely identified in the document using either: a) One of the dataset unique attributes, such as the name (if non-ambiguous) or the URL; or b) A citation of the dataset description paper in the reference section of the document. A manual evaluation on a random sample of 25 documents indicates high precision of 0.93.

2.2 Dataset references in linked datasets

It is a best practice to reuse resources from external datasets where possible. We consider a dataset *ds* to be reusing a resource from an external dataset ds_e if in the original dataset *ds* there is a triple that contains an IRI from the namespace of ds_e. In short, we speak of the original dataset *ds referencing* the external dataset ds_e. To estimate dataset reuse based on dataset references, we model the dataset collection as a directed graph whose nodes are datasets and whose edges are dataset references. On such a graph one can compute a PageRank [6] value for each dataset. PageRank value represents the steady-state probability of the random walk in a node; it is an interlinking metric that measures the popularity of the dataset based on the link structure in the graph.

2.3 Dataset paper citations

The importance of publishing datasets is increasingly recognised in the scientific community [3, 7]. In case a dedicated dataset paper is available, reuse can also be measured in terms of the number of citations of this paper, as specified in [3].

2.4 Correlation analysis

To better understand the similarities in the dataset reuse indications across the different communication channels, we analyse the correlation across the reuse metrics using the Pearson Correlation Coefficient (PCC). This computation requires our dataset collection to be represented as a vector space, where each dataset represents a dimension, and each metric is represented as a vector of values of this metric for all datasets. $PCC \in [-1, 1]$ is a measure of the linear correlation between two variables, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. We show the results in Sec. 4.

3 DATA COLLECTION

Publications: We consider publications from the ISWC, ESWC and WWW conferences from 2007 to 2015. All papers except Part II of the ESWC 2010 and 2011 proceedings are included. In total our collection contains 2,162 papers. We extracted dataset mentions from the main content, the evaluation section and the reference part of each paper. To this end we used pdfminer[2] and regular expressions and performed the extraction and disambiguation procedure described in Sec. 2.1. As we observed in our test collection, there is a strong correlation (PCC of 0.955) between the mentions of datasets in the evaluation section and the rest of the full text in a publication. Therefore, we do not further differentiate between the mentions in different sections of the publications.

Mailing Lists: We downloaded a crawl containing JSON encodings of the discussions on the *public-lod@w3.org* (2008 to 2015) and *semantic-web@w3.org* (2007 to 2015). From the mailing lists, we extracted the title, date, body and answer(s) body of each thread. For extracting and disambiguating dataset mentions from publications and mailing lists, we used the approach discussed in Sec. 2.1. In total, we extracted 9,046 discussion threads from *semantic-web@w3.org* and 4,661 discussion threads from *public-lod@w3.org*.

Dataset Dictionary: We extracted metadata of 1,131 datasets tagged as Linked Open Data (“lod”) from *datahub.io*. 818 of these are available under an open licence according to the open definition[3] (“isopen” tag) and 313 other datasets (194 of them do not specify any licence information and 119 explicitly specify other licences, e.g., a commercial licence). We refer to this group of 313 datasets as “non-open” in the following. From the metadata we used the name (unique in the catalogue) of a dataset, the title (long name), the homepage URL and the resources URL (downloadable resources and/or service APIs such as endpoints) to construct our dictionary. We also recorded the year the dataset entry was added to *datahub.io*. Furthermore, we identified publications describing 142 of the datasets in the dictionary by manually inspecting homepages of datasets and searching Google Scholar for the dataset title. To provide a comprehensive metadata collection, during our manual inspection we also added alternative dataset names used by the dataset authors on their homepages to the dictionary.

3.1 Resulting collection: An overview

Fig. 1 shows an overview of data collected for our evaluation (the *y* axis is logarithmic). Fig. 1.a) illustrates the number of datasets published on *datahub.io* in each year and the proportion between open and non-open license datasets. The number of new entries on *datahub.io* was particularly high in 2009–2012 and 2014, with a total of 818 datasets added in these years (i.e. 72% of the overall 1131 datasets in our collection). While the majority of the datasets in our collection are open, most non-open datasets were added in 2010–2012. A total of 245 out of 313 non-open datasets, i.e., 78.2%, was added in this period. Fig. 1.b) shows the aggregated number of papers from three scientific conferences—ISWC, ESWC and WWW—in each year and the number of dataset mentions in the full text. The number of publications remained stable over the whole time interval, whereas the number of dataset mentions in the publications has grown from 50 in 2007 to 110 in 2015.

[1] The code is publicly available at: https://github.com/keme686/dataset_reuse

[2] https://euske.github.io/pdfminer/

[3] http://opendefinition.org

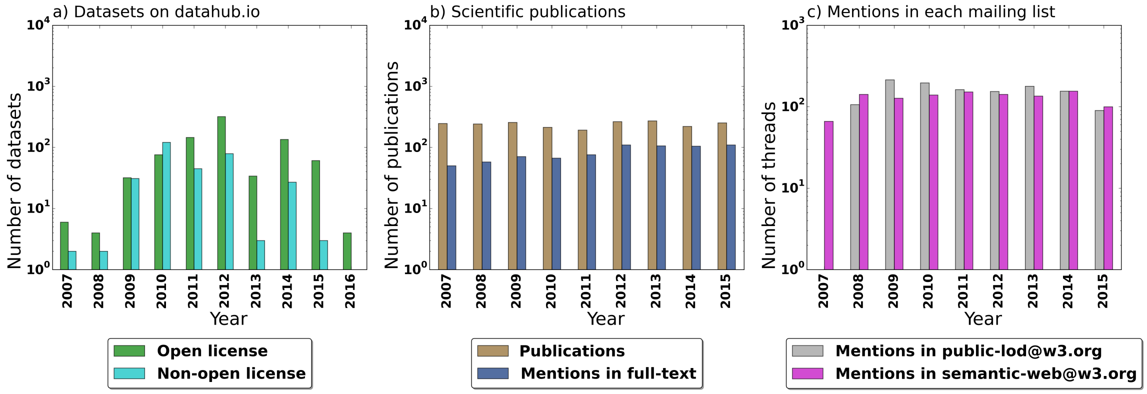

Figure 1: Data collection overview

Finally, Fig. 1.c) shows the number of dataset mentions extracted from the *semantic-web@w3.org* and *public-lod@w3.org* mailing lists. The overall number of threads in both mailing lists remained rather stable over time with an average of 1,005 per year for the *semantic-web@w3.org* and 583 per year for *public-lod@w3.org* (starting from 2008 in the case of *public-lod@w3.org*).

4 ANALYSIS RESULTS

The goal of our analysis is to provide insights in the dataset reuse indications in the communication channels we observed, and to analyse the correlation across the reuse metrics.

4.1 Dataset mention results

Fig. 2 shows the top datasets mentioned in publications and mailing lists. Part a) shows the top-10 datasets mentioned in the full text of scientific publications. The number of mentioned datasets has grown over the years with a particular increase starting from 2012—the same year for which we observed an increased number of *datahub.io* entries; cf. Fig. 1.a). Overall, DBpedia, Freebase, YAGO and GeoNames are the most widely mentioned datasets across all channels. Overall each of them were mentioned more than 100 times in publications. Fig. 2.b) shows the top-10 datasets mentioned in the *public-lod@w3.org* mailing list. Fig. 2.c) shows the top-10 datasets mentioned in the *semantic-web@w3.org* mailing list. There was a particularly high number of dataset discussions in the *public-lod@w3.org* mailing list between 2009 and 2010. We can observe, that *public-lod@w3.org* mailing list was overall more popular for the dataset discussions than the *semantic-web@w3.org* mailing list.

4.1.1 Correlation between reuse metrics in publications and mailing lists. Table 1 presents an overview of the correlation between the reuse metrics computed for different communication channels using PCC. This table shows the correlation between the different indicators of reuse, showing that in fact if interest in a dataset is expressed in mailing lists, it is commonly also mentioned in scientific publications. There is a strong correlation between the dataset mentions in the overall full text of the publications, the evaluation section and other sections (Non-eval section). Other communication channels are also strongly correlated, with $PCC = 0.86$ between the two mailing lists, and a moderate positive relationship ($PCC > 0.65$) between the mailing lists and the publications.

Table 1: Correlation between reuse metrics in publications and mailing lists.

	Full text	*semantic-web*	*public-lod*	Non-eval.	Evaluation
Full text	1.	0.659814	0.682735	0.837973	0.838151
semantic-web	0.659814	1.	0.860982	0.762222	0.720168
public-lod	0.682735	0.860982	1.	0.766012	0.745704
Non-eval.	0.837973	0.762222	0.766012	1.	0.955192
Evaluation	0.838151	0.720168	0.745704	0.955192	1.

4.2 Reuse of resources in datasets

In our collection, most of the datasets are associated with a namespace specified in *datahub.io*. To estimate reuse of resources, we consider the datasets for which we have both—namespace definition and dumps—available from the LOD Laundromat. This sub-collection contains 393 datasets in total. To compute reuse of the resources, we measure the interlinking among them, extracting references from one dataset to others in our sub-collection by streaming and parsing these datasets from the LOD Laundromat. After the extraction we removed duplicate references and computed PageRank on the resulting dataset graph, obtaining overall 261 datasets that were referenced at least once.

Table 2 presents the correlation results between PageRank and dataset popularity in publications and mailing lists for the 261 datasets in our collection that have $PageRank > 0$. For these datasets we observe a moderate positive correlation between the PageRank values and mentions in publications, and a strong positive correlation between PageRank and mentions in mailing lists.

Table 2: Correlation between PageRank and dataset popularity for 261 datasets that have $PageRank > 0$.

	Full text	*semantic-web*	*public-lod*	Non-eval.	Evaluation
PageRank	0.643109	0.734210	0.732594595	0.559696	0.622233

4.2.1 Evolution of reuse over time. We compute the correlation between the creation year of a dataset with its popularity in each channel. The results are shown in Table 3. First, we use the year the dataset entry was created in *datahub.io*. The result shows that there is no correlation between the metadata creation year with the popularity of the dataset in other channels. We then calculate an estimate of the dataset creation year by looking into the year the dataset

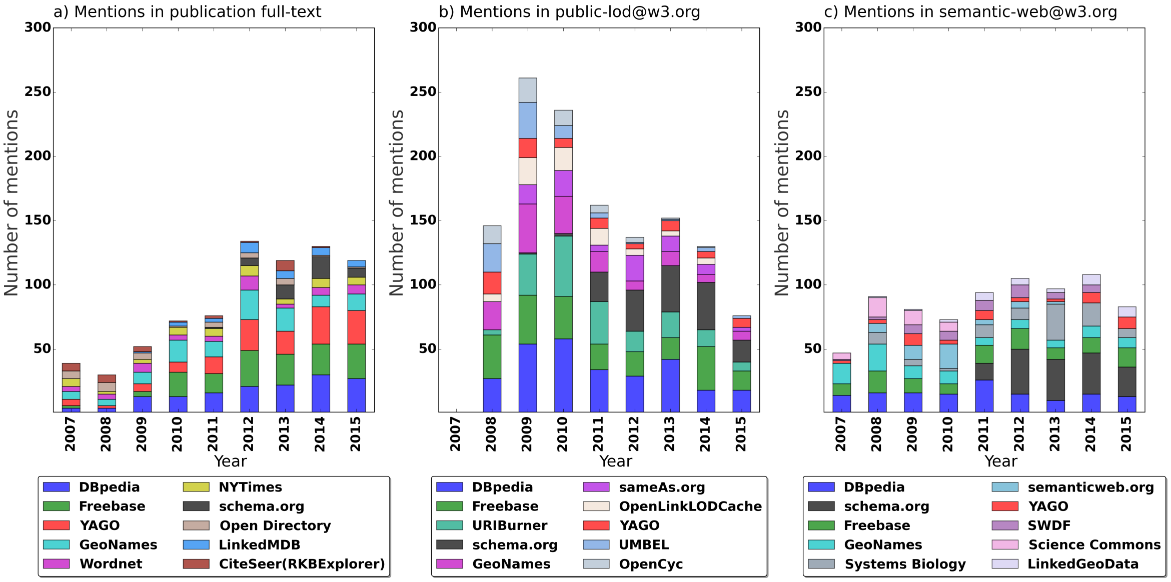

Figure 2: Top-10 datasets mentioned from 2007 to 2015.

was mentioned for the first time in a publication or a mailing list, and select the earliest date. The correlation result shows that there is a weak positive correlation between the estimated creation year with the dataset popularity in the full text of the publications. This illustrates that popularity of datasets in publications can grow with their age. However, we did not observe any significant correlation between the age of the dataset and its popularity in the mailing lists. For example, OpenCyc was often discussed in the *public-lod* mailing list, but become less popular with time. DBpedia had its discussion peak in the *public-lod* mailing list in 2009–2010; later it was mainly mentioned in the publications.

Table 3: Correlation between the popularity of a dataset and its age in *datahub.io* or an estimated dataset creation date.

	Full text	*public-lod*	*semantic-web*
Metadata created	0.128290	0.191551	0.143659
Dataset created (est.)	0.422133	0.225004	0.230189

4.3 Dataset papers in SWJ dataset track

According to [3], the most cited dataset published in the dedicated track of Semantic Web Journal (SWJ) in 2016 was *AGROVOC* with 39 citations of its SWJ paper. Among all SWJ datasets under consideration, this dataset has also the highest values with respect to the reuse indications in publications (13 mentions) and mailing lists (24 times in *public-lod@w3.org*). Although the absolute numbers of mentions are not very high, that can be explained by the relatively young age of the SWJ dataset track. Table 4 presents correlation results between the citations of the dataset papers published in the SWJ track and reuse metrics of these datasets as indicated in publications and mailing lists. The correlation results show that there is a positive correlation between the number of citations of the dataset papers and mentions in publications, that is most prominent in the *public-lod@w3.org*.

5 CONCLUSIONS

In this paper we analysed indications of dataset reuse within the Semantic Web community over nine years on a large-scale collection of LOD datasets.

Table 4: Correlation of reuse metrics for datasets in SWJ.

	Full text	*public-lod*	*semantic-web*
Paper citations	0.270327	0.555587	0.403551

First, we observed a positive correlation of reuse metrics across the different communication channels such as mailing lists and publications, indicating that these channels generally agree on the datasets they discuss. Datasets mentioned in scientific articles are typically mentioned within the evaluation section indicating their use for evaluation purposes. Second, our results demonstrate that dataset discussions in different channels and interlinking of the actual data are positively correlated. This correlation is stronger in case of the mailing lists. This confirms that the datasets discussed by the community are also typically reused in terms of data interlinking. In future research we would like to better understand reuse influence factors.

Acknowledgement. This work was partially funded by H2020-MSCA-ITN-2014 WDAqua (64279), project ANR-14-CE24-0029 OpenSensingCity and Data4UrbanMobility (BMBF, 02K15A040).

REFERENCES

[1] K. Boland, D. Ritze, K. Eckert, and B. Mathiak. 2012. Identifying References to Datasets in Publications. In *Proc. of the TPDL 2012.*

[2] M. B. Ellefi and et al. 2017. RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. *Semantic Web* (2017).

[3] A. Hogan, P. Hitzler, and K. Janowicz. 2016. Linked Dataset Description Papers at the Semantic Web Journal: A Critical Assessment. *Semantic Web* 7 (2016).

[4] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. 2012. An empirical survey of Linked Data conformance. *J. Web Sem.* 14 (2012), 14–44.

[5] J. Lehmann and et al. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *SWJ* 6, 2 (2015).

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank citation ranking: bringing order to the web.* Technical Report. Stanford InfoLab.

[7] H. A. Piwowar and T. J. Vision. 2013. Data reuse and the open data citation advantage. *PeerJ* (2013).

[8] M. Schmachtenberg, C. Bizer, and H. Paulheim. 2014. Adoption of the Linked Data Best Practices in Different Topical Domains. In *Proc. of the ISWC 2014.*

[9] F. M. Suchanek, G. Kasneci, and G. Weikum. 2007. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Prof. of the WWW 2007.*

[10] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. 2016. Quality assessment for Linked Data: A Survey. *Semantic Web* 7, 1 (2016), 63–93.