VoldemortKG: Mapping schema.org and Web Entities to Linked Open Data

Alberto Tonon(✉), Victor Felder, Djellel Eddine Difallah,
and Philippe Cudré-Mauroux

eXascale Infolab, University of Fribourg, Fribourg, Switzerland
{alberto,victor,ded,pcm}@exascale.info, philippe.cudre-mauroux@unifr.ch

Abstract. Increasingly, webpages mix entities coming from various sources and represented in different ways. It can thus happen that the same entity is both described by using schema.org annotations and by creating a text anchor pointing to its Wikipedia page. Often, those representations provide complementary information which is not exploited since those entities are disjoint. We explored the extent to which entities represented in different ways repeat on the Web, how they are related, and how they complement (or link) to each other. Our initial experiments showed that we can unveil a previously unexploited knowledge graph by applying simple instance matching techniques on a large collection of schema.org annotations and Wikipedia. The resulting knowledge graph aggregates entities (often tail entities) scattered across several webpages, and complements existing Wikipedia entities with new facts and properties. In order to facilitate further investigation in how to mine such information, we are releasing (i) an excerpt of all Common Crawl webpages containing both Wikipedia and schema.org annotations, (ii) the toolset to extract this information and perform knowledge graph construction and mapping onto DBpedia, as well as (iii) the resulting knowledge graph (VoldemortKG) obtained via label matching techniques.

Keywords: Knowledge graphs · schema.org · Instance matching · Data integration · Dataset

1 Introduction

Annotating webpages with structured data allows webmasters to enrich their HTML pages by including machine-readable content describing what we call *Web Entities*, along with their properties and the relationships that might exist among them. Such machine-readable content is embodied into the HTML markup by using specific formats like microdata or RDFa, and vocabularies coming from different ontologies. According to Bizer et al. [1], in 2013 the ontologies that were most widely used to describe Web Entities were: schema.org, a schema designed and promoted by several technology companies including Google, Microsoft, Pinterest, Yahoo! and Yandex; the Facebook Open Graph Protocol (OGP), which helps web editors integrating their content to the social networking platform; and

P. Groth et al. (Eds.): ISWC 2016, Part II, LNCS 9982, pp. 220–228, 2016.
DOI: 10.1007/978-3-319-46547-0_23

the GoodRelation vocabulary, which defines classes and properties to describe e-commerce concepts. As a result, the Web is now a prime source of structured data describing self-defined entities.

We argue that there is an underlying unexploited knowledge graph formed by such data, which overlaps and possibly complements other knowledge graphs in the Linked Open Data (LOD) cloud[1]. More specifically, we are interested in identifying connections between entities represented through annotations in webpages and entities belonging to further datasets, as well as discovering new entities that are potentially missing from well-known knowledge bases.

To extract such information, some challenges must first be overcome:

- Due to the decentralized nature of the Web, this knowledge graph is scattered across billions of webpages, with no central authority governing the creation and indexing of Web Entities;
- The markups are added by a crowd of non-experts driven by Search Engine Optimization (SEO) goals, hence the quality of the data is generally-speaking questionable;
- In order to produce high-quality links, one needs to extract supporting evidence from the annotated webpages, track provenance, clean and parse text and identify additional named entities.

In this context, we propose to help the Semantic Web research community tackle the open research problem of mapping Web Entities across webpages and finding their counterparts in other knowledge bases. To that end, we construct and release a dataset containing all webpages extracted from the Common Crawl dump[2] containing both Web Entities and links to Wikipedia. This data structure is designed to disseminate enough contextual information (full HTML content) and prior ground (Wikipedia links) to effectively perform the task of instance matching. In addition to the raw dataset of webpages and triples that we publish, we also showcase the generation of a proof-of-concept knowledge graph (VoldemortKG[3]). Our technique performs instance matching of microdata triples to their DBpedia counterparts via simple label matching. The resulting graph is also available as a downloadable (and browsable) resource and can serve as a baseline for more advanced methods.

2 Related Work

Extracting and leveraging online structured data has been of interest to many companies and was the core of a number of Web services. Sindice [7], was a search engine that indexed LOD data and provided a keyword search interface over RDF. The Sig.ma project [12] was an application built on top of Sindice that allowed browsing the Web of data by providing tools to query and mashup

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

[2] http://commoncrawl.org/.

[3] The knowledge graph that everyone knows exists, but no one talks about.

the retrieved data[4]. While potential applications of VoldemortKG could overlap with these projects, our present endeavor aims at providing key building blocks to perform data integration on the Web of data.

The Web Data Commons (WDC) initiative [5] extracts and publishes structured data available on the Web. The project makes available two important resources (i) Datasets, namely: RDFa, Microdata and Microformat, Web tables, Web hyperlinks, and IsA relations extracted from webpages, and (ii) the toolset for processing the Common Crawl dataset. Similarly, we build on top of the WDC Framework, and in addition extract and organize both structured data and HTML contents encompassing links pointing to Wikipedia. In contrast to the Web Data Commons, our objective is not only to collect and distribute the triples, but also the context in which they appear.

Instance Matching and Ontology Alignment. The process of matching entities across knowledge graphs is usually referred to as *Instance Matching*. The challenge is to automatically identify the same real world object described in different vocabularies, with slightly different labels and partially overlapping properties. Advanced techniques for instance matching compare groups of records to find matches [6] or use semantic information in the form of dependency graphs [3]. A task that often goes hand in hand with instance matching is *Ontology Alignment*. This task requires to map concepts belonging to an ontology to concepts of another ontology; for example, one can align the schema.org classes to their equivalent classes in the DBpedia ontology. The Ontology Alignment Evaluation Initiative (OAEI)[5] aims at stimulating and comparing research on ontology alignment. We point the reader to the work by Otero-Cerdeira et al. [8] for a detailed survey of the state of the art on the subject.

Our dataset will pose new challenges for both the instance matching and the ontology alignment community given the complexity of automatically mapping embedded structured data onto other LOD datasets. New methods need to be investigated in order to leverage the webpage contents for these tasks.

Entity Linking/Typing and AOR. Another relevant task in our context is *Entity Linking*, where the goal is to detect named entities appearing in text, and identify their corresponding entities in a knowledge base. Similarly, the dataset we release can be exploited for designing new methods for Ad-hoc Object Retrieval (AOR) [9], that is, building a ranked list of entities to answer keyword queries. Recent approaches for AOR make use of the literals connected to the entities in some knowledge base in order to use language modeling techniques to retrieve an initial ranked list of results that can be then refined by exploiting different kinds of connections among entities [2,11]. Lastly, Entity Typing (ET) is the task of finding the types relevant to a named entity. For instance, some ET systems focus on the following types Organization, Person and Location [4]. More recent work try to map named entities to fine-grained types [10]. Our dataset will challenge

[4] We note that both http://sindice.com and http://sig.ma are defunct as of April 2016.

[5] http://oaei.ontologymatching.org/.

these tasks by providing novel use-cases, where the extracted entities together with their types will then be used to verify and match against structured data embedded in the document.

3 The Dataset

As pointed out above in the introduction, it is worth exploring multiple representations of entities and connections among them. To foster investigations on this subject, we gathered a dataset that guarantees the presence of at least two sources of entities: i) DBpedia (as wikipedia anchors), and ii) structured data describing Web Entities. The dataset is created starting from the Common Crawl dated from November 2015[6], a collection of more than 1.8 billion pages crawled from the World Wide Web.

Data Extraction. We slightly modified the Web Data Commons Framework [5] to extract both the semantic annotations contained in the pages and the source code of all pages containing anchors pointing to any Wikipedia page. To lower the computational complexity during the extraction, we first test for the presence of Wikipedia anchors by matching against a set of simple regular expressions.

Even though we designed these regular expressions to achieve high recall—thus accepting the possibility of having many false positive pages—this simple filtering process significantly reduced the number of pages that we had to parse in order to extract the triples.

The whole process ran on 100 c3.4xlarge Amazon AWS spot instances featuring 16 cores and 30 GiB of memory each. The instances ran for about 30 h and produced 752 GiB of data, out of which 407 GiB contained compressed raw webpages and 345 GiB contained compressed semantic annotations in the form of 5-ple (subject, predicate, object, page url, markup format). In the rest of this document, we use the word "triple" to refer to the first three components of such extracted 5-ple. We release our modified version of WDC together with the dataset.[7]

Data Processing. In this step we process the webpages we extracted previously to build the final datasets we release. To create the datasets we used Apache Spark[8] and stored the pages, the 5-ples, and the anchors using the Parquet storage format[9] combined with the Snappy compression library:[10] this allows selective and fast reads of the data. We then used SparkSQL methods to discard semantic annotations extracted from pages not containing Wikipedia anchors, to determine the Pay Level Domains of the pages, to compute statistics, and to generate the final data we release. Together with the data, we also provide a

[6] http://commoncrawl.org/2015/12/.
[7] https://github.com/XI-lab/WDCFramework.
[8] https://spark.apache.org/.
[9] https://parquet.apache.org/.
[10] https://github.com/google/snappy.

Table 1. (left) Markup formats for including structured data in webpages and their popularity in number of annotations and number of webpages. (right) Top-10 vocabularies used to denote properties of structured data.

Format	N. Triples	N. Pages
μformats-hcard	317,636,734	4,190,649
μdata	84,073,194	2,539,539
RDFa	17,451,754	1,675,747
μformats-xfn	9,567,666	396,102
μformats-adr	4,333,580	165,601
μformats-geo	778,343	114,092
μformats-hcalendar	4,802,572	70,174
μformats-hreview	234,221	17,210
μformats-hrecipe	144,836	4,237
μformats-species	18,190	2,282
μformats-hresume	1,385	75
μformats-hlisting	2,610	20

Vocabulary	N. Triples	N. Pages
www.w3.org	363'103'085	7'113'775
schema.org	47'211'476	2'202'504
vocab.sindice.com	7'886'539	396'102
data-vocabulary.org	6'213'512	243'531
purl.org	5'408'960	2'625'015
ogp.me	2'231'434	398'927
opengraphprotocol.org	1'927'743	313'734
historical-data.org	1'213'531	43'026
rdfs.org	1'041'408	28'943
www.facebook.com	1'005'237	665'205

framework written in Scala allowing researchers to easily run their own instance matching methods on each webpage of the provided dataset.[11]

Key Statistics. Out of the 21,104,756 pages with Wikipedia anchors, 7,818,341 contain structured data. Table 1 (left) shows the distribution of the markup formats used to include structured data on webpages. As can be seen, 54 % of the webpages of our dataset are annotated by using some type of Microformats, 28 % of the pages contain Microdata, and 18 % contain RDFa annotations. This gives an idea of the diversity of sources one could tap into in order to extract entities and, possibly, connect them to other knowledge bases. In addition, we notice that more than one million pages contained in the dataset feature more than one type of markup format; detecting when the same entity is represented using different formats is an interesting open topic.

Table 1 (right) lists the top-10 vocabularies used in our dataset. The most wildely used is "www.w3.org" since the tool we used to extract structured data uses properties defined in that domain to encode is-a relations (e.g., all the `itemtype` Microdata annotations are translated into triples featuring the http://www.w3.org/1999/02/22-rdf-syntax-ns#type predicate). We observe that more than 3.3 million pages feature properties coming from more than one vocabulary and, more interestingly, almost 2.5 million pages feature properties selected from more than three vocabularies.

Distribution of the datasets. The dataset and the tool-chain used throughout this project is duly described on our website, for which we created a permanent URL: https://w3id.org/voldemortkg/. The extracted data is provided according the same terms of use, disclaimer of warranties and limitation of liabilities that apply to the Common Crawl corpus.[12]

[11] https://github.com/XI-lab/WDCTools.

[12] http://commoncrawl.org/terms-of-use/.

4 The VoldemortKG Knowledge Graph

To demonstrate the potential of the dataset we release, we built a proof of concept knowledge graph called VoldemortKG. VoldemortKG integrates schema.org annotations and DBpedia entities by exploiting the Wikipedia links embedded in the webpages we share. Equivalence between schema.org entities and DBpedia entities in VoldemortKG is based on string matching between the name of the former and the labels of the latter. Specifically, given a webpage P containing a DBpedia entity w, and a schema.org entity s, we say that w and s denote the same entity if the name of s, extracted from p by using the http://schema.org/name property, is also a label of w. A string s is a label of a DBpedia entity w if there is either a triple (w, rdfs:label, s) in DBpedia, or if in some webpage there is an anchor enclosing the text s and pointing to the Wikipedia page of w. We also exploit transitivity to generate equivalences among entities. Figure 1 shows our matching algorithm applied to a simple example.

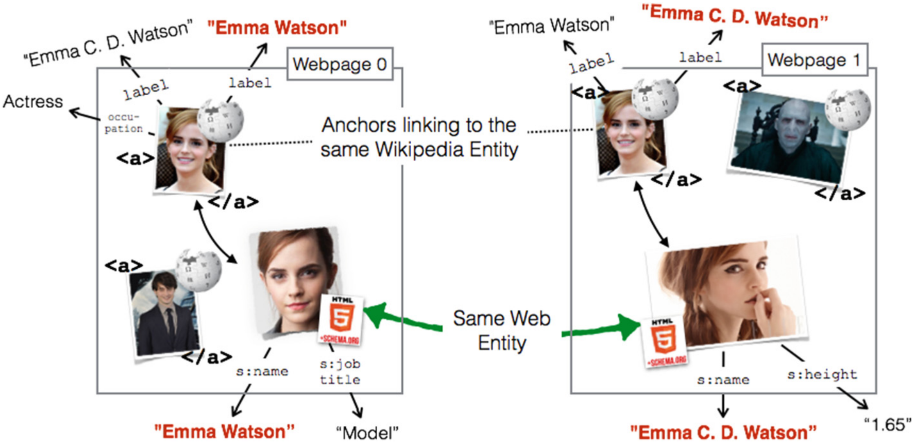

Fig. 1. Instance matching method used to build VoldemortKG. On the left-hand side, the DBpedia entry for Emma Watson is considered equivalent to a schema.org entity as its name is also a label of Emma Watson in DBpedia. On the right-hand side, a similar situation takes place for the same DBpedia entry and another Web entity. We thus conclude that all the mentioned entities refer to the same actress.

VoldemortKG is composed by 2.8 millions triples and contains information about 55,869 entities of 134 different types extracted from 202,923 webpages. Table 2 shows the top-15 entity types ordered by the number of instances (left), number of pages (center), and number of Pay Level Domains (PLDs, right) in which one of their instances appear. It is interesting to observe how top types change: notice that the top ranked type is different depending on the statistics taken into consideration. For example, the top ranked type in the right table is WebSite, with a count that is much higher than the number of VoldemortKG entities. This is due to the fact that entity E13418[13] appears in 132,616 webpages. This shows how one can get compelling results by leveraging simple string matching techniques in order to connect schema.org entities mentioned in different pages. Nevertheless, relying on such a simple method may result in many false positives, such as entity E13140[14], which is a person in VoldemortKG but an organization in DBpedia. This calls for further research on the topic.

Entity Fragmentation. It often happens that information about the same entity is scattered across several webpages. During the construction of VoldemortKG we extracted data about entities from 4 pages on average per entity (min. 1, max 132,616). As expected, there were cases in which the same (entity, property) pair was found in more than one webpage. For example, the properties `s:alternateName` and `owl:sameAs` appear, on average, in 367 and 11 pages per entity. Deciding which values of the property should be assigned to the entity taken into consideration is out of the scope of this work and is an interesting subject for future research.

Table 2. Top-15 entity types ordered by the number of instances (left), number of pages (center), and number of Pay Level Domains (PLDs, right) in which one of their instances appear. The prefix “hd” refers to http://historical-data.org.

Entity Type	N. Instances	Entity Type	N. Pages	Entity Type	N. PLDs
s:Person	36,370	s:WebSite	132,666	s:Article	295
s:RadioStation	3,524	s:Person	33,457	s:Person	203
hd:HistoricalPerson	1,963	s:Store	5,709	s:SiteNavigationElement	72
s:Movie	1,734	s:RadioStation	4,393	s:WebPage	70
hd:Person	1,496	s:AccountablePerson	3,645	s:Recipe	56
s:CollegeOrUniversity	1,308	s:CollegeOrUniversity	2,591	s:ListItem	53
s:AdministrativeArea	1,230	s:Movie	2,456	s:Product	40
s:School	1,118	s:Recipe	2,313	s:BlogPosting	36
s:Article	1,020	hd:Person	2,070	s:Organization	33
s:TVSeries	906	hd:HistoricalPerson	1,997	s:Place	25
s:AccountablePerson	844	s:Article	1,451	s:LocalBusiness	25
s:Blog	578	s:AdministrativeArea	1,252	s:Blog	24
s:Place	569	s:School	1,156	s:Thing	20
s:Landmarks...	518	s:TVSeries	948	s:MusicGroup	19
s:Book	428	s:QAPage	677	s:Book	17

5 Conclusions and Open Challenges

Taking advantage of the growing amount of structured data produced on the Web is critical for a number of tasks, from identifying tail entities to enriching existing knowledge bases with new properties as they emerge on the Web. While this information has been essentially exploited by commercial companies, it remains

[13] http://voldemort.exascale.info/resource/E13418.

[14] http://voldemort.exascale.info/resource/E13140.

an under-explored ground for the research community where several fundamental research challenges arise.

In this paper, we proposed a new dataset composed of webpages containing both Web Entities and Wikipedia links. Our goal was to extract and match structured pieces of data with high confidence in addition to provenance data, which constitutes a playground for researchers interested in a number of tasks including entity disambiguation & linking, entity typing, ad-hoc object retrieval or provenance management.

To demonstrate the usefulness of this dataset, we built a proof-of-concept knowledge graph (VoldemortKG) by label-matching triples to corresponding Wikipedia entities found on the same webpage. The resulting data was also made available in a browsable and downloadable format, and can be used as a baseline for further extraction and linking efforts.

Acknowledgements. This work was supported by the Swiss National Science Foundation under grant number PP00P2 153023. The authors would like to thank DAPLAB (http://daplab.ch/) for providing the resources necessary to create the datasets and, in particular, Benoit Perroud and Christophe Bovigny for their precious help.

References

1. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of RDFa, microdata, and microformats on the web - a quantitative analysis. In: ISWC 2013, pp. 17–32 (2013)
2. Bron, M., Balog, K., Rijke, M.: Example based entity search in the web of data. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 392–403. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36973-5_33
3. Dong, X., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: SIGMOD 2005, pp. 85–96 (2005)
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: ACL 2005, pp. 363–370 (2005)
5. Meusel, R., Petrovski, P., Bizer, C.: The Webdatacommons microdata, RDFa and microformat dataset series. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 277–292. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11964-9_18
6. On, B., Koudas, N., Lee, D., Srivastava, D.: Group linkage. In: ICDE 2007, pp. 496–505 (2007)
7. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: a document-oriented lookup index for open linked data. IJMSO **3**(1), 37–52 (2008)
8. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: a literature review. Expert Syst. Appl. **42**(2), 949–971 (2015)
9. Pound, J., Mika, P., Zaragoza, H.: Ad-hoc object retrieval in the web of data. In: WWW 2010, pp. 771–780 (2010)

10. Tonon, A., Catasta, M., Demartini, G., Cudré-Mauroux, P., Aberer, K.: Trank: ranking entity types using the web of data. In: ISWC 2013, pp. 640–656 (2013)
11. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Combining inverted indices and structured search for ad-hoc object retrieval. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 125–134. ACM, New York (2012)
12. Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., Decker, S.: Sig.ma: live views on the web of data. JWS **8**(4), 355–364 (2010)