Materializing the editing history of Wikipedia as linked data in DBpedia

Fabien Gandon, Raphael Boyer, Olivier Corby, Alexandre Monnin

Materializing the editing history of Wikipedia as linked data in DBpedia

Fabien Gandon[1], Raphael Boyer[1], Olivier Corby[1], Alexandre Monnin[1]

[1] Université Côte d'Azur, Inria, CNRS, I3S, France
Wimmics, Sophia Antipolis, France
{fabien.gandon, raphael.boyer, olivier.corby, alexandre.monnin}@inria.fr

Abstract. We describe a DBpedia extractor materializing the editing history of Wikipedia pages as linked data to support queries and indicators on the history.

The different instances of the DBpedia platform typically extract RDF from Wikipedia using up to 16 extractors. The extraction focuses on structured content including infoboxes, categories, links, etc. As an example, the French chapter[1], of which we are responsible, extracted 185 million triples in 2015. The resulting RDF graph is then published and supports up to 2.5 million SPARQL queries per day and an average of 70,000 SPARQL queries per day in 2015. But Wikipedia is a social media that produces more data than the actual content of its pages. The activity of the epistemic communities of Wikipedia produces a huge amount of traces showing, for instance, the evolution, conflicts, trends, and variety of opinions of the users. In fact, the different projects of the Wikipedia Foundation develop at a rate of over ten edits per second, performed by users from all over the world[2]. And this activity is performed on broad collection of topics: the English chapter of Wikipedia alone has over 5 million articles and the combined Wikipedias for all other languages exceed the English chapter in size with more than 27 billion words in 40 million articles in 293 languages[3]. As a result the history of the editing actions captures the peaks and shifts of interests of the contributors and indirectly reflects the unfolding of events all around the world and in every domain.

Providing means to monitor the editing activity has always been important for Wikipedians to follow the changes. These means include APIs such as the recent changes API, the IRC streams per languages, the WebSockets streams, the Server-Sent Events Streams, etc. [6]. Previous works also suggested to monitor real-time editing activity of Wikipedia to detect events such as natural disasters [1]. In [3] a resource versioning mechanism inspired from the Memento protocol (RFC7089) is applied but only to DBpedia dumps. In [4] historical versions of resources are regenerated for a given timestamp with some revision data but through a RESTful API. In [5] the preservation of the history of linked datasets is tested but only on a sample of 100,000 resources. We do not mention here works on formats, vocabularies or algorithms to detect and describe updates to RDF datasets since at this stage we are focusing on editing acts on Wikipedia.

---

[1] French DBpedia Chapter http://fr.dbpedia.org/

[2] Wikipedia Statistics - https://en.wikipedia.org/wiki/Wikipedia:Statistics

[3] Comparisons accessed 23/08/16 https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

Data about the activity provide historical indicators of interest, attention, over the set of resources they cover. They have been also used, for instance, to assess the currency of the data [7], to study conflict resolution [8], to temporally anchor data, to attribute changes and to identify vandalism [9] or to precisely attribute the authorship of content [10]. Inversely, using statements of other datasets (e.g. typing) one can filter and analyze the editing history considering chosen dimensions (e.g. focus on events about artists). But none of the previous contributions support public SPARQL querying of the full editing history. The potential of these linked data is even greater when combined with other linked data sources and this is not easily done with an API approach e.g. "give me the 10 most edited populated places in July 2012".

For this reason we designed and provide a new DBpedia extractor producing a linked data representation of the editing history of Wikipedia pages. Instead of real-time monitoring we capture the history as linked data to be able to query it, mine it and combine it with other sources to augment the dimensions we can exploit when querying linked data in general and DBpedia in particular.

A history dump of a Wikipedia chapter contains all the modifications dating back from the inception of this linguistic chapter along with some information for each and every modification. As an example, the French editing history dump represents 2TB of uncompressed data. The data extraction is performed though streams in Node.js with a MongoDB instance. It took 4 days to extract 55 GB of RDF in turtle on 8 Intel(R) Xeon(R) CPU E5-1630 v3 @ 3.70GHz with 68GB or RAM and using SSD disks. The result is then published through a SPARQL end-point with the DBpedia chapter[4].

The extractor reuses as many existing vocabularies from the LOV directory[5] as possible in order to facilitate integration and reuse. Figure 1 is a sample of the output of the edition history extractor for the page describing the author "Victor Hugo" in the DBpedia French chapter. The history data for such an entry contains one section of general information about the article history (lines 1-15) along with as many additional sections as there are previous revisions to capture each change (e.g. two revisions at lines 16-24). The general information about the article includes: the number of revisions (line 3), the date of creation and last modification (lines 4-5), the number of unique contributors (line 6), the number of revisions per year and per month (e.g. lines 7-8) and the average sizes of revisions per year and per month (e.g. lines 9-10). In addition each individual revision description includes: the date and time of the modification (e.g. lines 17), the size of the revision as a number of characters, (e.g. lines 18) the size of the modification as a number of characters (e.g. lines 19), the optional comment of the contributor (e.g. lines 20), the username or IP address of the contributor and if the contributor is a human or a bot (e.g. line 21 or 24) and a link to the previous revision (e.g. line 22).

By construction the data are fully linked to the DBpedia resources and the vocabularies used include: PROV-O, Dublin Core, the Semantic Web Publishing Vocabulary, DBpedia ontologies, FOAF and SIOC. As a result the produced linked data are well integrated to the LOD cloud. Every time we were missing a predicate we added it to DBpedia FR ontology. As shown in Figure 2 these data support very

[4] History endpoint http://dbpedia-historique.inria.fr/sparql

[5] Linked Open Vocabularies (LOV) http://lov.okfn.org/ as accessed in June 2016

arbitrary queries such as, in this example, the ability to request most modified pages grouped by pairs of pages modified the same day.

```
1. <https://fr.wikipedia.org/wiki/Victor_Hugo> a prov:Revision ;
2.  dc:subject <http://fr.dbpedia.org/resource/Victor_Hugo> ;
3.  swp:isVersion "3496"^^xsd:integer ;
4.  dc:created "2002-06-06T08:48:32"^^xsd:dateTime ;
5.  dc:modified "2015-10-15T14:17:02"^^xsd:dateTime ;
6.  dbfr:uniqueContributorNb 1295 ;
   (...)
7.  dbfr:revPerYear [ dc:date "2015"^^xsd:gYear ; rdf:value "79"^^xsd:integer ] ;
8.  dbfr:revPerMonth [ dc:date "06/2002"^^xsd:gYearMonth ; rdf:value
        "3"^^xsd:integer ] ;
   (...)
9.  dbfr:averageSizePerYear [ dc:date "2015"^^xsd:gYear ; rdf:value
        "154110.18"^^xsd:float ] ;
10. dbfr:averageSizePerMonth [ dc:date "06/2002"^^xsd:gYearMonth ; rdf:value
        "2610.66"^^xsd:float ] ;
   (...)
11.  dbfr:size "159049"^^xsd:integer ;
12.  dc:creator [ foaf:nick "Rinaldum" ] ;
13.  sioc:note "wikification"^^xsd:string ;
14.  prov:wasRevisionOf <https:// … 119074391> ;
15.  prov:wasAttributedTo [ foaf:name "Rémih" ; a  prov:Person, foaf:Person ] .

16.  <https:// … 119074391>  a prov:Revision ;
17.   dc:created "2015-09-29T19:35:34"^^xsd:dateTime ;
18.   dbfr:size "159034"^^xsd:integer ;
19.   dbfr:sizeNewDifference "-5"^^xsd:integer ;
20.   sioc:note "/*Années théâtre*/ neutralisation"^^xsd:string ;
21.   prov:wasAttributedTo [ foaf:name "Thouny" ; a  prov:Person, foaf:Person ] ;
22.   prov:wasRevisionOf <https://... 118903583> .
   (...)
23.  <https:// … oldid=118201419> a prov:Revision ;
24.   prov:wasAttributedTo [ foaf:name "OrlodrimBot" ; a prov:SoftwareAgent ] ;
   (...)
```

Fig. 1. Extract of the output of the edition history extractor for Victor Hugo

```
1. PREFIX dc: <http://purl.org/dc/element/1.1/>
2. PREFIX prov: <http://www.w3.org/ns/prov#>
3. PREFIX swp: <http://www.w3.org/2004/03/trix/swp-2/>
4. select distinct ?x ?y ?d where
5. { ?x a prov:Revision .
6.   ?y a prov:Revision .
7.   ?x dc:modified ?d .
8.   ?y dc:modified ?d .
9.   ?x swp:isVersion ?v .
10. FILTER (?v>1000 && ?x<?y) } LIMIT 10
```

Fig. 2. Ten of the most modified pairs of pages modified the same day

The STTL template language [2] allows to generate portals in a declarative and fast way. We used it to build two portals to show the richness of the historical data materialized. The first application designed is a visual history browser that displays images of the 50 most edited topics for every month. With the second portal we demonstrate the ability to join this new dataset with other linked data sources starting with DBpedia itself: we built a focused portal generator that reduces the monitoring activity to specific DBpedia categories of resources[6] (e.g. companies, actors, countries,

[6] Category-filtered view of the History: Focusing on artists (mode=dbo:Artist)
http://corese.inria.fr/srv/template?profile=st:dbedit&mode=dbo:Artist
Or focusing on countries (mode=dbo:Country) using the DBpedia ontology
http://corese.inria.fr/srv/template?profile=st:dbedit&mode=dbo:Country

etc.). Figure 3 is a screenshot of the portal focused on countries and shows the events in Ukraine in 2014. Many applications of the editing activity already exist [6] and these two portals are only a proof of concept for what can be done with SPARQL over the linked data of editing activity.

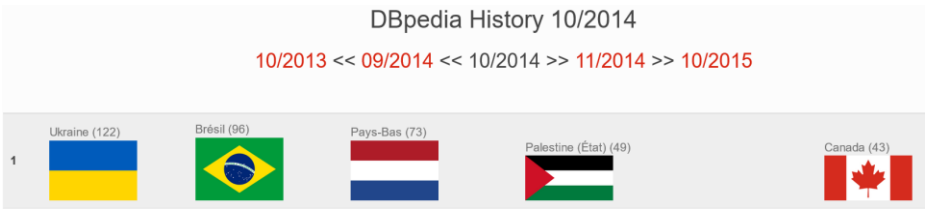

Fig. 3. Portal showing Countries whose page was subject to a maximum activity.

The history extractor is now integrated to the DBpedia open-source code and running on the production server of the French chapter. We are studying the integration of the live change feed for both the chapter and its history in order to reflect real-time changes to the content and editing logs. We are also considering ways to represent more precisely the changes between two revisions.

References

1. T. Steiner, Comprehensive Wikipedia monitoring for global and realtime natural disaster detection. In Proceedings of the ISWC Developers Workshop 2014, at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, 2014. , 86–95.
2. O. Corby, C. Faron-Zucker, F. Gandon, A Generic RDF Transformation Software and its Application to an Online Translation Service for Common Languages of Linked Data. The 14th International Semantic Web Conference, Oct 2015, Bethlehem, United States. 2015
3. H. Van de Sompel, R. Sanderson, M. L. Nelson, L. L., Balakireva, H. Shankar, and S. Ainsworth. An HTTP-based versioning mechanism for linked data. LDOW, 2010.
4. Javier D. Fernández, Patrik Schneider, and Jürgen Umbrich. 2015. The DBpedia wayback machine. In Proceedings of the 11th International Conference on Semantic Systems SEMANTICS '15 ACM, 192-195.
5. Paul Meinhardt, Magnus Knuth, and Harald Sack. 2015. TailR: a platform for preserving history on the web of data. In Proceedings of the 11th International Conference on Semantic Systems SEMANTICS '15, ACM, 57-64
6. Thomas Steiner, The Wiki(pedia|data) Edit Streams Firehose, Invited Talk, Wiki Workshop, April 12, WWW 2016, Montreal, Canada, http://bit.ly/wiki-firehose
7. Anisa Rula, Luca Panziera, Matteo Palmonari, Andrea Maurino: Capturing the Currency of DBpedia Descriptions and Get Insight into their Validity. COLD 2014
8. Volha Bryl and Christian Bizer. Learning conflict resolution strategies for cross-language Wikipedia data fusion. In 4th Workshop on Web Quality Workshop (WebQuality) at WWW 2014, 2014
9. E. Alfonseca, G. Garrido, J.-Y. Delort, and A. Penas. WHAD: Wikipedia historical attributes data – Historical structured data extraction and vandalism detection from the Wikipedia edit history. Language Resources and Evaluation, 47(4):1163–1190, 2013
10. Fabian Flöck and Maribel Acosta. 2014. WikiWho: precise and efficient attribution of authorship of revisioned content. In Proceedings of the 23rd international conference on World wide web WWW '14. ACM, 843-854