Wikipedia editing history in DBpedia

Fabien Gandon, Raphael Boyer, Olivier Corby, Alexandre Monnin

Wikipedia editing history in DBpedia

extracting and publishing the encyclopedia editing activity as linked data

Fabien Gandon, Raphael Boyer, Olivier Corby, Alexandre Monnin
Université Côte d'Azur, Inria, CNRS, I3S, France
Wimmics, Sophia Antipolis, France
Firstname.Name@inria.fr

***Abstract*— DBpedia is a huge dataset essentially extracted from the content and structure of Wikipedia. We present a new extraction producing a linked data representation of the editing history of Wikipedia pages. This supports custom querying and combining with other data providing new indicators and insights. We explain the architecture, representation and an immediate application to monitoring events.**

Keywords— DBpedia, editing activity, linked data

I. EXTRACTING EDITION HISTORY

DBpedia is a central source in the Linked Open Data cloud[1] and one of the largest in terms of topics coverage since it is based on the cross-domain encyclopedic knowledge of Wikipedia. Today, the DBpedia platform extracts structured knowledge from Wikipedia using up to 16 extractors that mostly focus on page content (infoboxes, categories, links, etc.). The French chapter[2] for which we are responsible currently extracts 185 million triples that are then published and support up to 2.5 million SPARQL queries per day for an average of 68 700 SPARQL queries per day in 2015.

However, besides the content, the editing activity on Wikipedia also produces traces that are highly representative of Wikipedians' variety of focus at a given time. "Wikipedia and its sister projects develop at a rate of over 10 edits per second, performed by editors from all over the world"[3]. This happens while the "English Wikipedia alone has over 5,093,940 articles of any length, and the combined Wikipedias for all other languages greatly exceed the English Wikipedia in size, giving more than 23 billion words in 35 million articles in 291 languages"[4] . As a side effect, logs of the page changes capture the current focus and shift of interests of the contributors. Since the encyclopedia has a fast and broad coverage, this editing reflects the unfolding of events all around the world and in every do-main. Previous works thus proposed to monitor real-time editing activity of Wikipedia as a way to detect natural disaster [1].

To support even more applications of the analysis of the edition activity, we designed a new DBpedia extractor that produces linked data from the history of each Wikipedia page. A historic dump of Wikipedia contains all the modifications dating back from the inception of a linguistic chapter along with some information for each and every modification. As an example, the French editing history dump represents 2TB of uncompressed data. This data extraction is performed by stream in Node.js with a MongoDB instance. It takes 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 currently published as a DBpedia test chapter with its own SPARQL end-point for beta-testing[5].

II. RELATED WORK

In the Wikipedia community there is a long history of providing means to monitor the editing activity they include: the recent changes API, IRC streams per languages, WebSockets streams, Server-Sent Events Streams, etc. [6]

Authors of [3] proposed a resource versioning mechanism based content negotiation on time as proposed in the Memento protocol (HTTP Framework for Time-Based Access to Resource States, RFC7089[6]). They applied it to DBpedia but their approach is only archiving released dumps. In fact they are limited to five versions of the English-language of DBpedia (2.0 through 3.3) while in our work we will cover each and every single editing act since the creation of the French Wikipedia chapter.

The DBpedia Wayback machine [4] allows to retrieve historical versions of resources at a given timestamp and to be able to reconstruct the full history of RDF documents. They offer this service through a Web user interface and a RESTful API. The major difference with what we propose here is that we focus on capturing and materializing the editing activity that led to each version (time, frequency, sizes, editors, etc.) as linked data to support any possible query. The Wayback machines provides the RDF version of a Wikipedia article at a certain time and some revision data but only through and API and for a given resource. It is not possible to have arbitrary queries such as "give me the 20 most edited actors last month". Moreover our extraction reuses as many existing vocabularies from the LOV[7]directory as possible in order to facilitate integration and reuse.

Other contributions like [5] focus on preserving the history of linked datasets over time, providing access to prior states of contained resources and a versioning platform for linked data in general. Again a sample of 100,000 resources in releases 3.2 through 3.9 of DBpedia is used for the evaluation dataset but the goal of the authors is not to provide and support a

[1] LOD Cloud Diagram http://lod-cloud.net/
[2] French DBpedia Chapter http://fr.dbpedia.org/
[3] Wikipedia Statistics - https://en.wikipedia.org/wiki/Wikipedia:Statistics
[4] Comparisons https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons
[5] History endpoint http://dbpedia-test-fr.inria.fr/historique/sparql
[6] RFC 7089 https://tools.ietf.org/html/rfc7089
[7] Linked Open Vocabularies (LOV) http://lov.okfn.org/

comprehensive view of the editing activity of Wikipedia supporting open-queries as we do.

None of these previous contributions support custom querying of editing history while it is known now that the availability as linked data supports unforeseen innovative usages far beyond what is possible with constrained APIs. Moreover the potential of this data is even greater when combined with other linked data sources which again is not easily done with an API approach. These data about the activity provide indicators of interest, attention, over the set of resources they cover. Inversely, using statements of other datasets (e.g. typing) one can filter and analyze the editing history considering many dimensions. Finally we do not include here works on formats, vocabularies or algorithms to detect and describe updates to RDF datasets since at this stage what we are capturing are editing acts on Wikipedia.

III. Modeling the edition history as linked data

Figure 1 is a sample of the output of the edition history extractor for the page describing "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 many additional sections corresponding to the previous revisions made to capture each change (e.g. two revisions at lines 16-24). The general information about the article includes: a) the number of revisions, b) the date of creation and last modification, c) the number of unique contributors, d) the number of revisions per year and per month and e) the average size of revisions per year and per month. Each revision section includes: a) the date and time of the modification, b) the size of the modification as a number of characters, c) the size of this previous revision as a number of characters, d) the username or IP address of the contributor, e) the optional comment of the contributor and f) if the contributor is a human or a bot. 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. Therefore 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 exemplified in Figure 2 these data support very expressive queries such as, in this example, the ability to request most modified pages grouped by pairs of pages modified the same day.

IV. First applications for validation: news and event detection

We prototyped two applications to demonstrate the effectiveness of the historical data captured. To do so we used the STTL template language which allows us to develop proofs of concept very quickly [2]. The first application designed (Figure 2) is a visual history browser that displays images of the 50 most edited topics for every month[8]. For example, French elections, as a topic, appear in the top five of edited pages in September 2015. Then we demonstrated the ability to cross this new dataset with other linked data sources starting with DBpedia itself. We built a focus portal generator that reduces the monitoring activity to specific DBpedia categories of resources (e.g. companies, actors, countries, etc.). In Figure 3 the portal is focused on the artist category[9] and gives visibility to the death of Christopher Lee, an event which happened in June 2015. Many applications of the editing activity already exist [6] and these two portals are only examples of what can be done with simple SPARQL queries when the editing activity is available as linked data.

[8] Images of editing history http://corese.inria.fr/srv/template?profile=st:dbedit

V. Next step: going live

The extractor is being integrated to the DBpedia open-source code and in parallel the running service will migrate from the test server to the production server of the French chapter. We are now studying the adaptation of the new live mechanism dedicated both to the chapter and history extraction in order to reflect real-time changes to the content and edition history. Finally we are considering new applications, such as new linking methods [7], that could be supported by this dataset.

Acknowledgment

We thank the French Ministry of Culture for supporting the French DBpedia chapter.

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, Italy, October 19, 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. In Proc. of 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), Sebastian Hellmann, Josiane Xavier Parreira, and Axel Polleres (Eds.). ACM, New York, NY, USA, 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), Sebastian Hellmann, Josiane Xavier Parreira, and Axel Polleres (Eds.). ACM, New York, NY, USA, 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] Ramine Tinati, Markus Luczak-Roesch, Wendy Hall, Nigel Shadbolt, (2016) More than an edit: using transcendental information cascades to capture hidden structure in Wikipedia. At 25th International World Wide Web Conference, Montreal, Canada, 11 - 15 Apr 2016.

[9] Category-filtered view of the History: Focusing on artists http://corese.inria.fr/srv/template?profile=st:dbedit&mode=http://dbpedia.org/ontology/Artist

```
1. <http://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 Paris – can be reproduced by visiting http://bit.ly/vh-revisions

```
1. PREFIX dc: <http://purl.org/dc/terms/>
2. PREFIX prov: <http://www.w3.org/ns/prov#>
3. PREFIX swp: <http://www.w3.org/2004/03/trix/swp-2/>
4. select * 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 pages grouped by pairs of pages modified the same day

Fig. 4. The French election topic in the top 5 of edited pages in September 2015

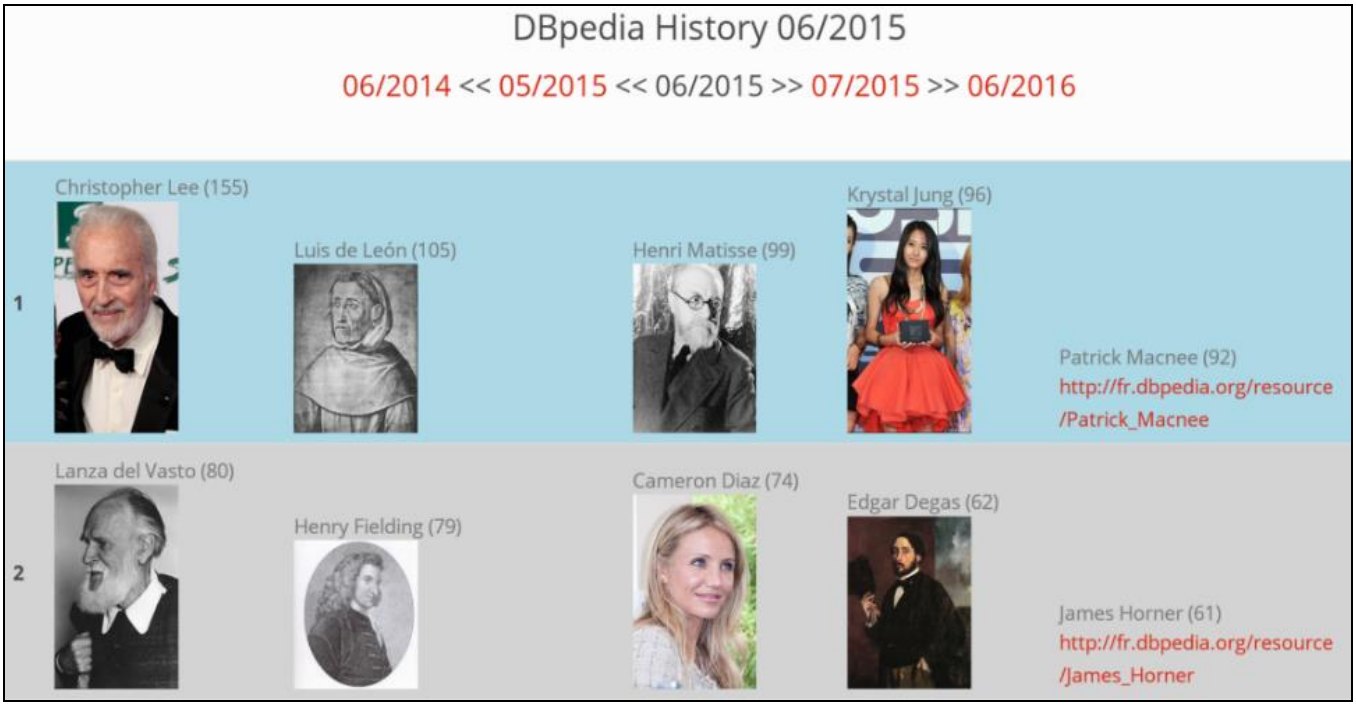

Fig. 3. Crossing DBpedia artist category with edition history data