The DBpedia Events Dataset

Magnus Knuth[1], Jens Lehmann[2], Dimitris Kontokostas[2],
Thomas Steiner[3], and Harald Sack[1]

[1] Hasso Plattner Institute, University of Potsdam, Germany
{firstname.lastname}@hpi.uni-potsdam.de
[2] Universität Leipzig, Institut für Informatik, AKSW, Germany
{lastname}@informatik.uni-leipzig.de
[3] CNRS, Université de Lyon, LIRIS – UMR5205, Université Lyon 1, France
tsteiner@liris.cnrs.fr

Abstract. Wikipedia is the largest encyclopedia worldwide and is frequently updated by thousands of collaborators. A large part of the knowledge in Wikipedia is not static, but frequently updated, e. g., political events or new movies. This makes Wikipedia an extremely rich, crowdsourced information hub for events. However, currently there is no structured and standardised way to access information on those events and it is cumbersome to filter and enrich them manually. We have created a dataset based on a live extraction of Wikipedia, which performs this task via rules for filtering and ranking updates in DBpedia Live.

1 Introduction

Since 2007, the DBpedia project has been extracting metadata and structured data from Wikipedia and made it publicly available as RDF triples [3]. DBpedia also offers a live synchronized version of extracted data – DBpedia Live [4]. The English Wikipedia alone has hundreds of updates per minute [6] that are processed via the Live framework. Changes in Wikipedia articles are often connected to real life events, such as news related events from politics, cultural life, or sports. Due to the large user base of Wikipedia, these events are often quickly updated – in many cases quicker than in other Web sources [7,1].

However, currently there is no structured and standardised way to access information about these events and it is cumbersome to filter and enrich them manually. While there are previous efforts to extract events from Wikipedia such as [2,5,7,8], associated data about these events is not always available as RDF or even archived. Providing an RDF dataset has the benefit of being able to rely on standards for accessing and querying information. Furthermore, events can be readily combined with background knowledge from DBpedia and other sources, which enables mashups of events with further structured data.

The most important challenges when extracting events from DBpedia are (i) detecting events, (ii) providing context, and (iii) ranking events according to their importance. Since by far not all changes in Wikipedia are events, we need a mechanism to detect those. In our case, we rely on a semi-automatic approach

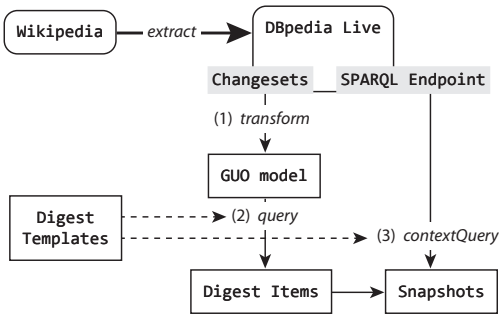

Fig. 1. The extraction process

based on extensible rule sets, which are executed over DBpedia Live changesets. If a rule fires, it triggers another query obtaining contextual information. The detected event is ranked according to the resource's PageRank and the edit activity of the Wikipedia page. The output of the processing pipeline is stored in RDF preserving all provenance information.

2 Conversion Process

Figure 1 shows the underlying workflow, which has three major steps: (1) DBpedia Live changesets are transformed to an RDF representation, (2) relevant changes are retrieved according to queries defined in *Digest Templates*, and (3) *Digest Items* are made up with contextual information, e. g., snapshots taken from the DBpedia Live SPARQL endpoint.

DBpedia Live constantly monitors updates to Wikipedia articles and re-extracts corresponding resources using the DBpedia extraction framework. The resource descriptions are diffed with their current revision and changed RDF triples are published in form of changeset files, i. e., gzipped N-triples dumps of added and removed triples[4]. These changeset files are primarily intended for synchronization of RDF stores. In order to make these changesets queryable, they are transformed to an RDF representation using the Graph Update Ontology (GUO)[5]. For each re-extracted resource a `guo:UpdateInstruction` is created, that contains the added and removed subgraphs, aggregated for a given time-span.

Relevant changes are extracted from this model by executing SPARQL queries, which are defined in so-called *Digest Templates*. These queries can exclusively select patterns from inserted and deleted triples. The structure of a digest template is shown exemplary in Listing 1.1. The context query is executed on the DBpedia Live SPARQL endpoint to validate the result based on context information that is not available in the changesets. This allows to also consider unchanged statements about the resource for the event selection, e. g. the `dig:PRESIDENT` template in Listing 1.1 only updates of resources are allowed, which are typed as

[4] `http://live.dbpedia.org/changesets/`
[5] `http://purl.org/hpi/guo#`

dbo:Organisation and having a label. More complex restrictions that can be expressed with SPARQL may apply. The dig:descriptionTemplate is used to generate a natural language headline for the detected event by substitution of the placeholders (enclosed in %%) with the respective resource labels.

Listing 1.1. The PRESIDENT digest template

```
dig:PRESIDENT a dbe:DigestTemplate ;
  dcterms:identifier "PRESIDENT" ;
  dcterms:description """President changed."""@en ;
  dbe:queryString """SELECT ?u ?res ?oldPres ?newPres
    { ?u guo:target_subject ?res ;
         guo:delete [ dbo:president ?oldPres ] ;
         guo:insert [ dbo:president ?newPres ] . }""" ;
  dbe:contextQueryString """SELECT ?label
    { %%res%% a dbo:Organisation ; rdfs:label ?label . }""" ;
  dbe:descriptionTemplate """%%newPres%% succeeds %%oldPres%% as the
      ↪ president of %%res%%.""" .
```

From the validated result the final event or so-called *Digest Item* is created. These items contain all necessary information to understand the change that occurred in DBpedia Live.

Listing 1.2. An event created from the LEADER template

```
item:2015/04/25/Christian_Democrats_(Sweden)-LEADER a dbe:Event ;
  dbe:context snapshot:2015/04/25/Christian_Democrats_(Sweden) ;
  dbe:update update:2015/04/25/Christian_Democrats_(Sweden) ;
  dcterms:description """Ebba Busch succeeded Goran Hagglund as the leader of
      ↪  Christian Democrats (Sweden)."""@en ;
  dbe:rank 1.82421e-06 ;
  prov:generatedAtTime "2015-04-30T13:45:35.798+02:00"^^xsd:dateTime ;
  prov:wasDerivedFrom dig:LEADER ,
    changesets:2015/04/25/14/000201.removed.nt.gz ,
    changesets:2015/04/25/14/000201.added.nt.gz .
```

3 Dataset Description

The dataset consists of daily digest dump files, which contain the descriptions of events (c.f. Listing 1.2) occurring on this day as well as the resource updates related to them. The resource snapshots that are linked in the event descriptions might be relevant for further investigation. Thus, they are kept separately in individual snapshot dumps. The daily generated event dumps can be accessed at `http://events.dbpedia.org/dataset/` and additionally a SPARQL interface is offered at `http://events.dbpedia.org/sparql` for querying the full dataset. The resource snapshots corresponding to the events are published in a separate path at `http://events.dbpedia.org/snapshot/`.

At the current stage 12 digest templates have been defined and fired[6]. Table 1 shows the number of events that matched the templates.

4 Conclusion

This paper presents an automated means to detect events and extract relevant data changes within DBpedia Live on the one hand, and on the other hand make these events available as Linked Data for others to consume and build upon.

[6] Defined digest templates: `http://events.dbpedia.org/data/digests.ttl`.

Count	Template	Count	Template
4509	AWARDED	146	LEADER
3252	HEADHUNTED	118	PODIUM
2539	RELEASED	89	PRESIDENT
1991	JUSTMARRIED	22	VOLCANO
785	DEADPEOPLE	7	EUROPE2015
782	JUSTDIVORCED	1	INTRODUCED

Table 1. Extracted events per template

Potential use cases for our constantly growing dataset include, but are not limited to, (breaking) news detection systems for news agencies, brand monitoring systems for so-called digital war rooms, but also more mundane use cases such as celebrity trackers (who married whom), or mashups in general. The dataset provides a comprehensible overview on usually rather complex data changes and may give valuable insights into dataset dynamics. Having stable identifiers for events further allows for interesting reasoning use cases.

Some information can simply not be deduced from discrete state resource descriptions, e. g. that a person moved from Germany to France can not be extracted from the separated facts that she lived in Germany and lives in France, rather both states need to be regarded and compared. This is what this project makes possible. The application supports an individual selection of changes of interest by the free definition of digest templates, which allows monitoring customized data change events.

References

1. B. Fetahu, A. Anand, and A. Anand. How much is wikipedia lagging behind news? In *Proceedings of the 2015 ACM conference on Web science*. ACM, 2015.
2. M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, pages 254–266. Springer, 2013.
3. J. Lehmann, R. Isele, M. Jakob, et al. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.
4. M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann. DBpedia and the live extraction of structured data from wikipedia. *Program*, 46(2):157–181, 2012.
5. M. Osborne, S. Petrović, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *Proceedings of the SIGIR Workshop on Time-aware Information Access*, 2012.
6. T. Steiner. Bots vs. wikipedians, anons vs. logged-ins (redux): A global study of edit activity on wikipedia and wikidata. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 25:1–25:7. ACM, 2014.
7. T. Steiner et al. MJ no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proceedings of the 22nd Int. Conference on World Wide Web Companion*, pages 791–794, 2013.
8. G. B. Tran and M. Alrifai. Indexing and analyzing wikipedia's current events portal, the daily news summaries by the crowd. In *Proceedings of the the 23rd International Conference on World Wide Web Companion*, pages 511–516, 2014.