Wikipedia Revision Toolkit

Efficiently Accessing Wikipedia’s Edit History

TECHNISCHE UNIVERSITÄT DARMSTADT

UBIQUITOUS KNOWLEDGE PROCESSING

Oliver Ferschke, Torsten Zesch, Iryna Gurevych

http://www.ukp.tu-darmstadt.de

Motivation

Access to Wikipedia‘s Edit History

- Article revisions constitute a novel knowledge source for NLP. The sequence of article edits can be used as training data for data-driven NLP algorithms.
- Efficient access to this resource is limited by the immense size of the data. Most of the previous work using revisions only regards small samples of the available data.
- Demand for easy programmatic access to the revision data and reduction of the required storage space.

Reconstruction of Wikipedia Dumps

- Most Wikipedia-based NLP algorithms work on single static snapshots of Wikipedia.
- This does not pay respect to the fact that Wikipedia is a dynamic resource which is constantly changed by its millions of editors.
- The rapid change is bound to have an influence on the performance of NLP algorithms using Wikipedia data.
- Older snapshots eventually become unavailable, as there is no official backup server. As a consequence, older experimental results cannot be reproduced anymore.

Revision Storage

- Dedicated revision storage format – only the changes between two adjacent revisions are stored
- Reduction of the demand for disk space by 98% compared to the original dump
- Every nth revision is stored in full to increase reconstruction speed.
- Example

  r1 : This is the very first sentence!
  r2 : This is the second sentence

  r2 can be encoded as

```
REPLACE 12 10 ’second’
DELETE 31 1
```

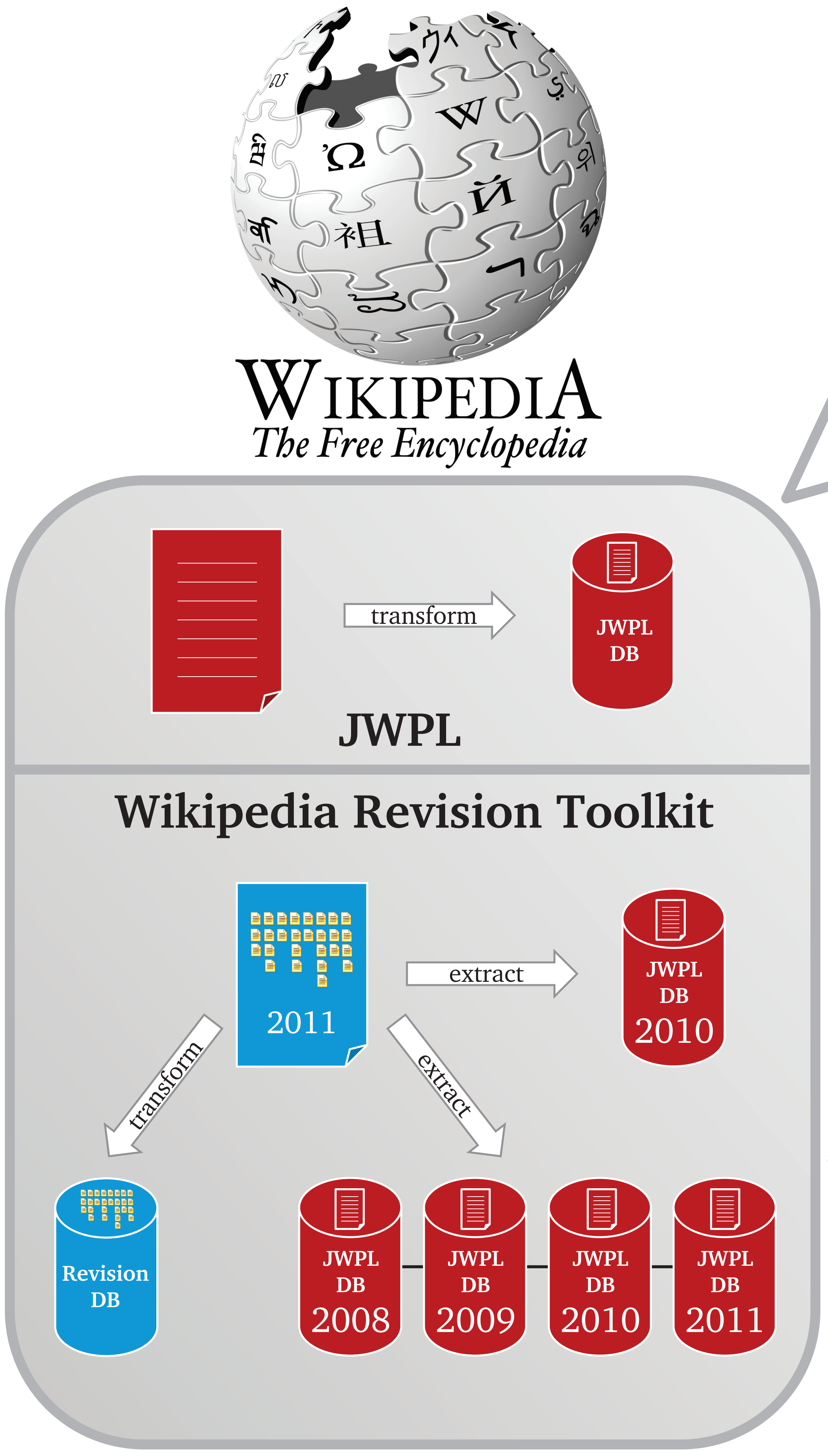
Revision Access

- Iteration over all revisions of all articles via Java Iterator
- In combination with JWPL: Access to revisions of specific articles

```
Page article = wiki.getPage(”Computer”);
int id = article.getPageId( );

// Get all revisions for the article
Collection<Timestamp> revTimeStamps =
 revApi.getRevisionTimestamps(id) ;
for (Timestamp t : revTimeStamps)
{
 Revision rev = revApi.getRevision(id,t);
 // process revision …
}
```

- Access to revision meta data
  - Edit comment
  - Revision author
  - Author is registered (bool)
  - Minor revision (bool)
  - Unique contributors to article

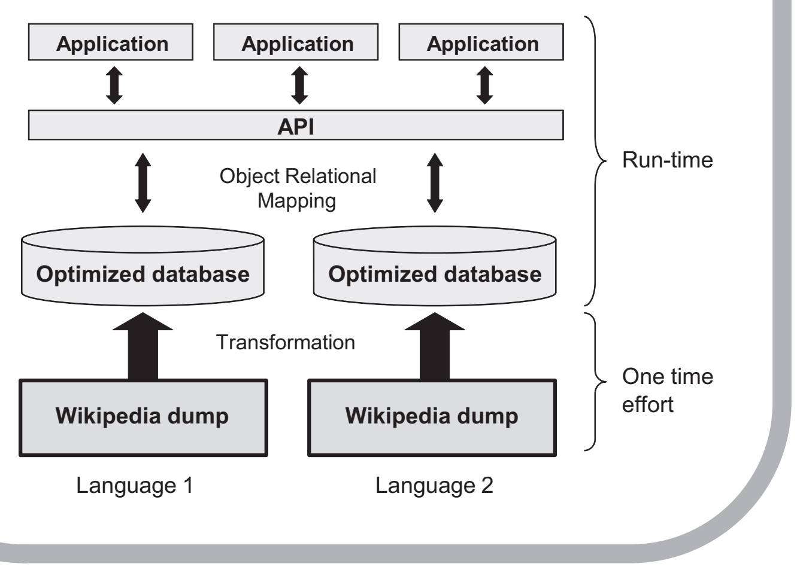


■ Wikipedia dump with/without revisions

● Database with/without revisions

JWPL

- Open source Java-based API
- High performance access to Wikipedia via optimized database
- Articles and Categories as Java Objects
- Access to information nuggets:
  redirects, links, sections, interlanguage links, first paragraphs, etc.
- Language independent
- Parser for the MediaWiki syntax



Dump Reconstruction

Single Snapshot

Reconstruction of any earlier state of Wikipedia from a single revision dump

Snapshot Series

Automatic creation of series of Wikipedia snapshots given the start and end date of the series and the interval between snapshots

References

- T. Zesch, C. Mueller, and I. Gurevych, 2008. **Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary**. *In Proceedings of the Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco.
- O. Ferschke, T. Zesch, and I. Gurevych, 2011. **Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia‘s Edit History**. *In Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies. System Demonstrations.* Portland, OR, USA.

VolkswagenStiftung

http://jwpl.googlecode.com

LOEWE