BERD @NFDI

2nd Workshop on Wikibase in Knowledge Graph based Research Data Management (NFDI) Projects

RaiseWikibase: Towards fast data import into Wikibase

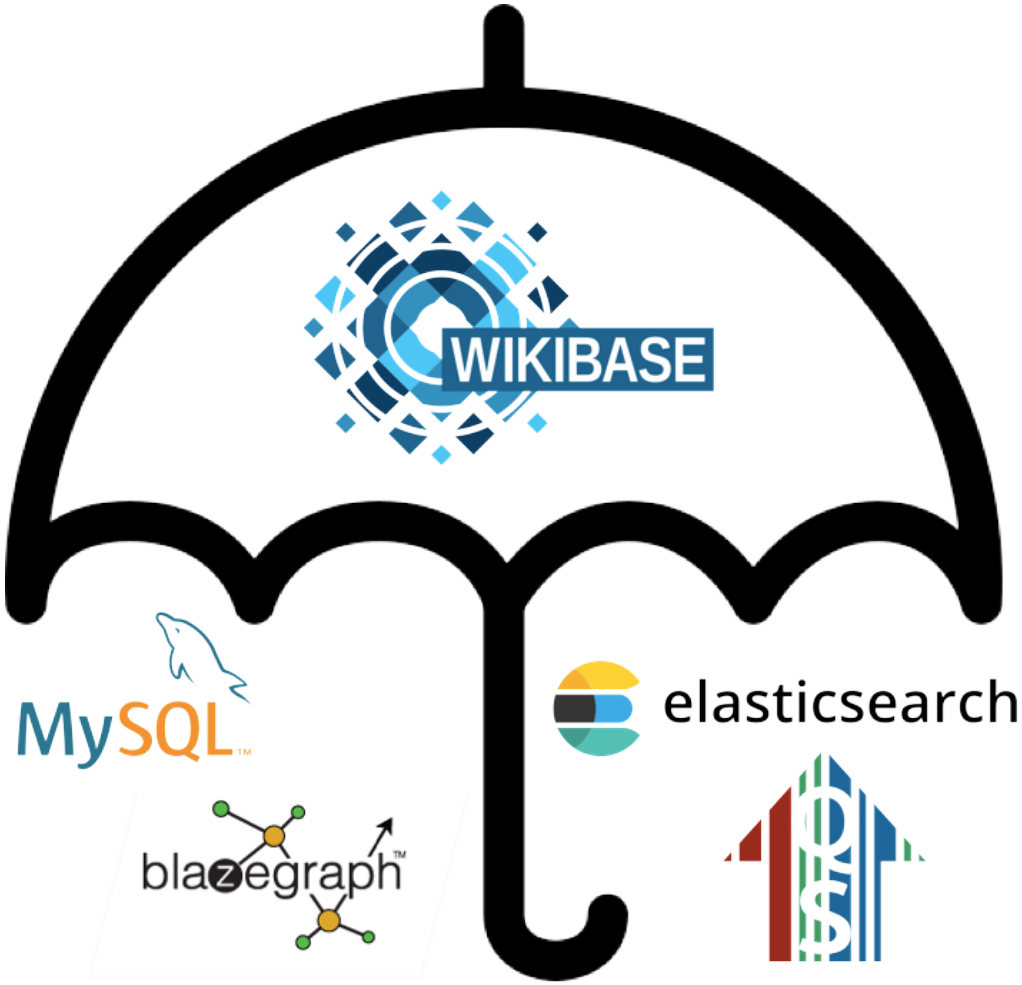
Dr. Renat Shigapov, UB Mannheim

29.07.2021

https://github.com/UB-Mannheim/RaiseWikibase

UNIVERSITÄT MANNHEIM · UB MANNHEIM · UNIT MANNHEIM · ZEW · BERD@BW

Definitions & performance of data import into Wikibase



gadgets, extensions and third-party tools

Definition:

Wikibase is a software for collaborative multilingual knowledge graph construction

Data import:

Humans interact with UI (JS + the Wikibase API) and bots send requests to the Wikibase API

Performance of data import:

The Wikibase API & its wrappers create roughly 1-6 entities per second

Overview of data import tools for Wikibase

BERD@NFDI

	transaction per entity	transaction per many entities	
the Wikibase frontend	**the Wikibase API**	**MariaDB**	
manual page creation	WikidataIntegrator	wikibase-insert	Java
	WikibaseIntegrator	RaiseWikibase	Python
	wikibase-cli		
	WikidataToolkit		
	Pywikibot		
	QuickStatements		
less than 1 page per second	**1-6 pages per second**	**100-300 pages per second**	

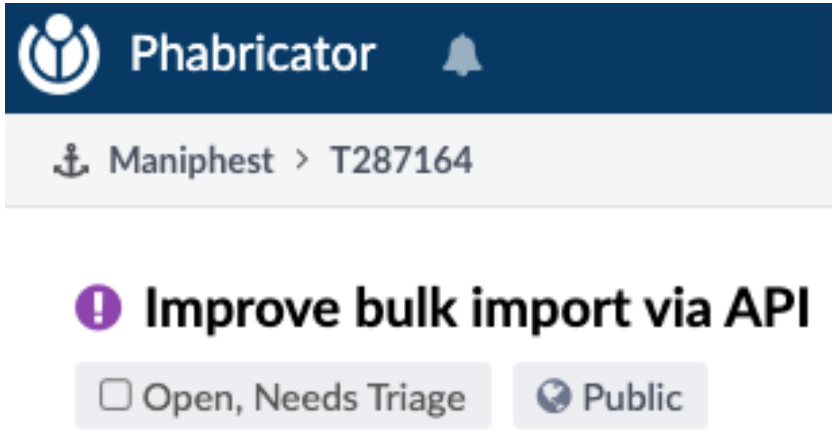
Why is bulk import via API slow?

The Wikibase API saves entities one by one.
During each entity saving request the API does:

- parameter validation
- creation of an empty item
- transforming the JSON representation
- permission checks
- inserts into many tables of MariaDB
- reporting back to the user

Current activity

https://phabricator.wikimedia.org/T287164



Please join our discussion on bulk data import via API

See the post https://addshore.com/2021/07/what-happens-in-wikibase-when-you-make-a-new-item by Adam Shorland

RaiseWikibase: QuickStart

1. Basic info

- Open-source Python tool.
- Adapted to Wikibase Docker Image "1.35".
- Connects to MariaDB via the mysqlclient library.

2. Installation

Clone RaiseWikibase and install it via `pip3`:

```
git clone https://github.com/UB-Mannheim/RaiseWikibase
cd RaiseWikibase/
pip3 install -e .
```

3. How to use

```
from RaiseWikibase.raiser import batch
batch(content_model='wikibase-item', texts=[item for i in range(1000)])
```

where `item` is the JSON representation of an item created using the entity function.

RaiseWikibase: functions for the Wikibase Data Model

BERD@NFDI

```
from RaiseWikibase.datamodel import label, alias, description, snak, claim, entity
```

```
labels = {**label('en', 'organization'), **label('de', 'Organisation')}
```

```
aliases = alias('en', ['organisation', 'org']) | alias('de', ['Org', 'Orga'])
```

```
descriptions = description('en', 'social entity (not necessarily commercial)')
descriptions.update(description('de', 'soziale Struktur mit einem gemeinsamen Ziel'))
```

```
mainsnak = snak(datatype='external-id', value='Q43229', prop='P1', snaktype='value')
```

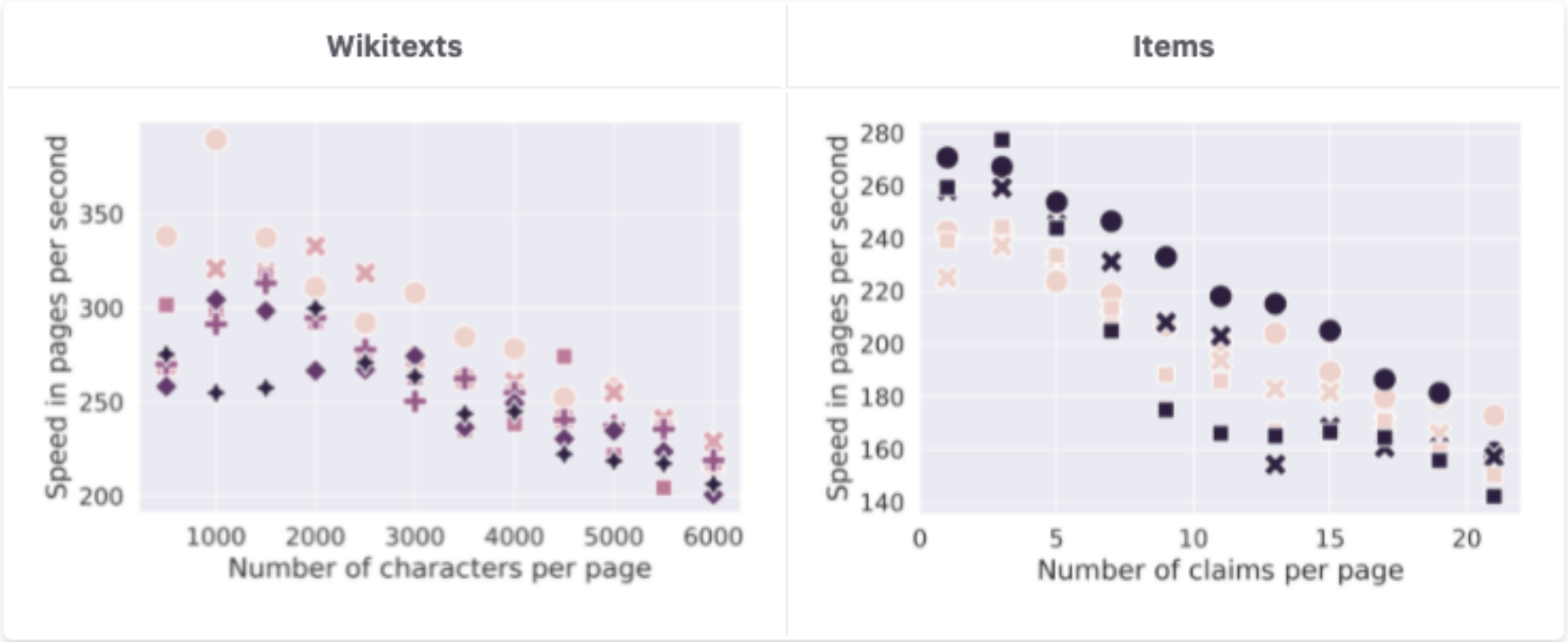
```
qualifiers = [snak(datatype='external-id', value='Q43229', prop='P1', snaktype='value')]
references = [snak(datatype='external-id', value='Q43229', prop='P1', snaktype='value')]
```

```
claims = claim(prop='P1', mainsnak=mainsnak, qualifiers=qualifiers, references=references)
```

```
item = entity(labels=labels, aliases=aliases, descriptions=descriptions, claims=claims, etype='item')
```

RaiseWikibase: Performance analysis

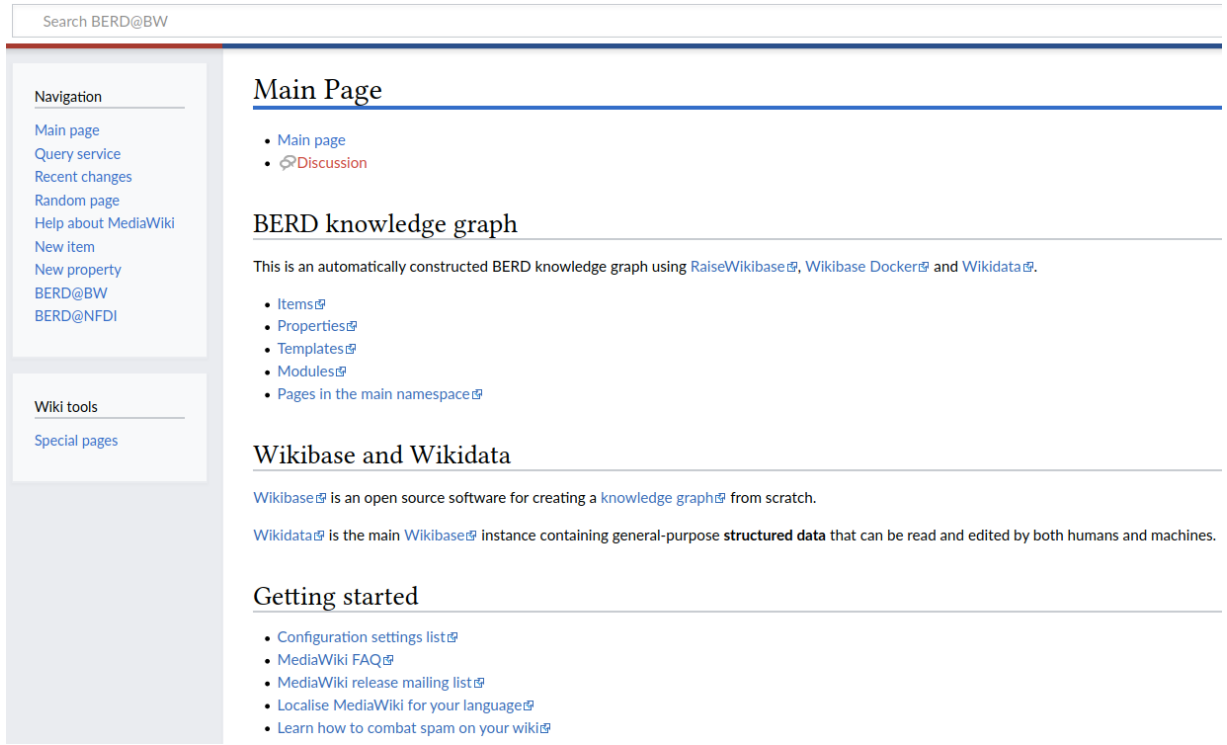
- The insert rate decreases approximately linearly with increasing number of characters per wikitext and with increasing number of claims per item.
- Small pages are uploaded at rates of 250-350 wikitexts per second and 220-280 items per second.



RaiseWikibase: Extras

Configuration

- The folder texts contains templates, modules and other unstructured data.
- Modify them and add your own files to texts.
- Run the function fill_texts.



Federated properties

- Federated properties are under development in Wikimedia Germany.
- Therefore, run the script miniWikibase.py.
- It creates all properties from Wikidata in a local Wikibase instance.

Next page (P1309)

- Wikidata ID (P1)
- country (P10)
- ethnic group (P100)
- depicted by (P1000)
- bibcode (P1001)
- number of elevators (P1002)
- primary destinations (P1003)
- instrument (P1004)
- central bank (P1005)
- CTBUH Skyscraper Center building ID (P1006)
- Swiss parliament ID (P1007)
- officeholder (P1008)
- EGAXA ID (P1009)
- performer (P101)
- statement disputed by (P1010)
- lostbridges.org ID (P1011)
- has facet polytope (P1012)
- office held by head of government (P1013)
- number of spans (P1014)
- NLA Trove ID (P1015)
- Swedish Media Database ID (P1016)
- floruit (P1017)
- proved by (P1018)
- earliest date (P1019)
- manufacturer (P102)
- OpenCorporates ID (P1020)
- licensed to broadcast to (P1101)
- Cycling Archives cyclist ID (P1102)
- number of representatives in an organization/legislature or won in elections (P1103)
- nominated for (P1104)
- languages spoken, written or signed (P1105)
- GUI toolkit or framework (P1106)
- Oxford Dictionary of National Biography ID (P1107)
- affiliation (P1108)
- Encyclopædia Britannica Online ID (P1109)
- made from material (P111)
- orbits completed (P1110)
- shape (P1111)
- taxon synonym (P1112)
- GRIN URL (P1113)
- Sandrart.net person ID (P1114)
- template has topic (P1115)
- topic's main template (P1116)
- ecoregion (WWF) (P1117)
- start point (P1118)
- Lost Art ID (P1119)
- location of discovery (P112)
- has pet (P1120)
- OpenPlaques subject ID (P1121)
- executive producer (P1122)
- published in (P1123)

Limitations

- Secondary tables for items and properties are filled via the maintenance scripts
- CirrusSearch indexing is lengthy

TODOs

- Find faster way to fill the secondary tables for items and properties
- Add parallel CirrusSearch indexing

https://lists.wikimedia.org/hyperkitty/list/wikibaseug@lists.wikimedia.org/thread/RBPTOYYMMLIFYRSHEPNEXUGLXTJDJTCI

Summary

- RaiseWikibase is a Python tool for fast data import into Wikibase
- Relatively easy to use, but it still needs further development
- What if the Wikibase API would have the same functionality?

BERD@NFDI funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581

Links

- Ticket “Improving bulk import via API”: https://phabricator.wikimedia.org/T287164
- RaiseWikibase GitHub repo: https://github.com/UB-Mannheim/RaiseWikibase
- RaiseWikibase docs: https://ub-mannheim.github.io/RaiseWikibase/
- Paper “RaiseWikibase: Fast inserts into the BERD instance”: https://doi.org/10.1007/978-3-030-80418-3_11
- Post “What happens in Wikibase when you make a new item” by Adam Shorland: https://addshore.com/2021/07/what-happens-in-wikibase-when-you-make-a-new-item
- Wikibase-Insert GitHub repo: https://github.com/jze/wikibase-insert
- Post “Filling a Wikibase instance with millions of data” by Jesper Zedlitz: https://blog.factgrid.de/archives/2013