Scaling Wikidata Query Service

unlimited access to all the world's
knowledge for everyone is hard

WIKIMANIA SINGAPORE | Lydia Pintscher - @nightrose

Wikidata Query Service?!

The basics

The Wikidata Query Service is...

- **A critical part of Wikidata**
- **Querying relations that unlock the true power of Wikidata**
- **Running Blazegraph**

Who uses the Query Service?

Wikidata Editors

- **Understand and maintain parts of Wikidata**
- **Advocacy work and workshops**
- **Show the world what they worked on**

WIKIMANIA
SINGAPORE

Knowledge seekers/sharers

- **Use queries to satisfy curiosity**
- **Use queries to share something curious with the world (e.g. via social networks)**

WikidataFacts
@WikidataFacts

popes who were children of other popes:
query.wikidata.org/embed.html#%23...

parent	Sergius III
child	John XI
parent	Anastasius I
child	Innocent I
parent	Hormisdas
child	Silverius

Larissa Borck
@Larissa_Borck

Thanks to @wikidata, here comes an overview on all Swedish citizens who dies in 1949 - and whose works will enter the #PublicDomain on 1 January 2020. #PublicDomainDay (h/t @JolanWuyts)
query.wikidata.org/#%23%20Swedish...

Small and medium sized re-users

- **Use queries to power their applications and services**
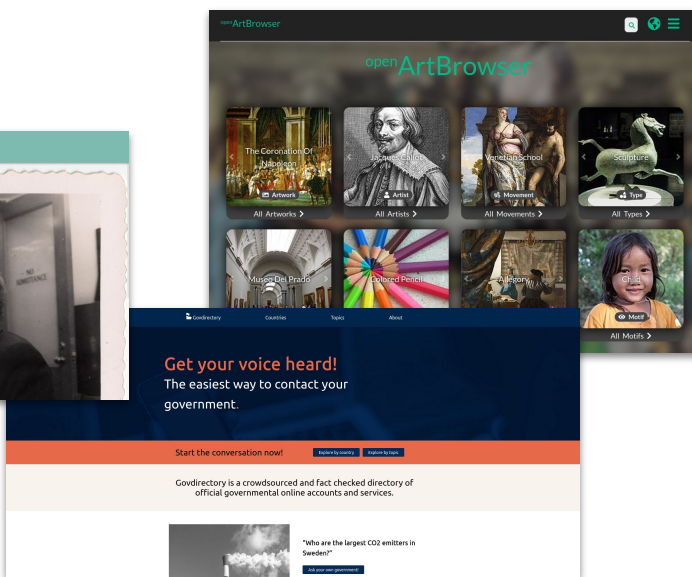
Wikimedia projects

- **Structure work in wiki projects and power campaigns, especially around missing content (e.g. Women in Red)**
- **Better understand the content of the project**

WIKIMANIA
SINGAPORE

Wikimedia development teams and tool builders

- **Queries power tools (Listeria, Item Quality Evaluator, Integraality), often by getting lists of Items**

WIKIMANIA
SINGAPORE

The scale

- **One of the largest SPARQL endpoints on the Internet**
- **15 Billion triples from 105 Million Items with almost 1.5 Billion statements + smaller number of Properties and Lexemes**
- **700k edits per day on Wikidata**
- **About 5000 requests per minute**

Current (interconnected) challenges

WIKIMANIA
SINGAPORE

Keeping up with data size

- **Wikidata keeps growing**
- **Blazegraph has no sharding support -> larger disks and memory size required**
- **Internal Blazegraph limitations for number of allocators**

Keeping up with query and write load

- **The query load can overload the system, leading to high response times and timeouts**
- **WDQS write loads are lagging behind Wikidata**

Keeping it stable and secure

- **Servers crashing leading to limited capacity and overload**
- **Blazegraph no longer actively developed / maintained**
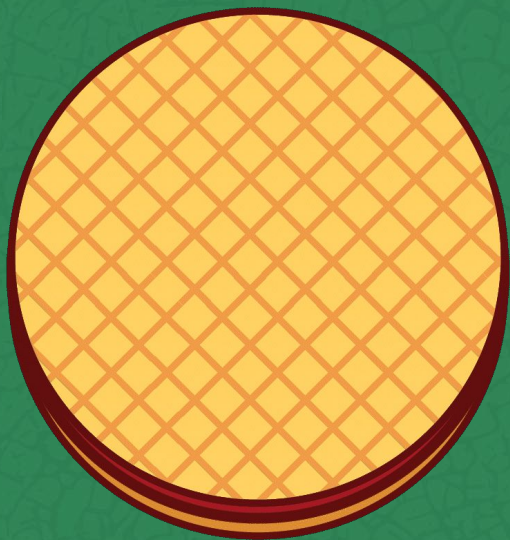
Consequences

- **Legitimate queries time out**
- **As the graph grows, queries that worked before, now no longer work**
- **Editors are restricting their editing work**
- **Editors and reusers are not getting useful new functionality**

How have we addressed the problem?

WIKIMANIA
SINGAPORE

What we've done

- Introduced a new streaming updater to cope with more edits/min
- Made a disaster mitigation plan
- Got an overview of alternative backends
- Took pressure off the system
  - Built out the Wikibase Ecosystem and especially Wikibase Cloud
  - Developed the Wikibase REST API
  - Improved documentation for the different ways to access Wikidata's data to help developers chose the right one for their usecase

WIKIMANIA SINGAPORE

What we are doing

- **Thinking through splitting off a part of the graph into a separate Blazegraph instance while keeping the data in Wikidata**
- **Discussing the future of the scientific article corpus in Wikidata with the WikiCite community**
- **Reducing redundant data by introducing a new language code for multilingual content**

And in the future?

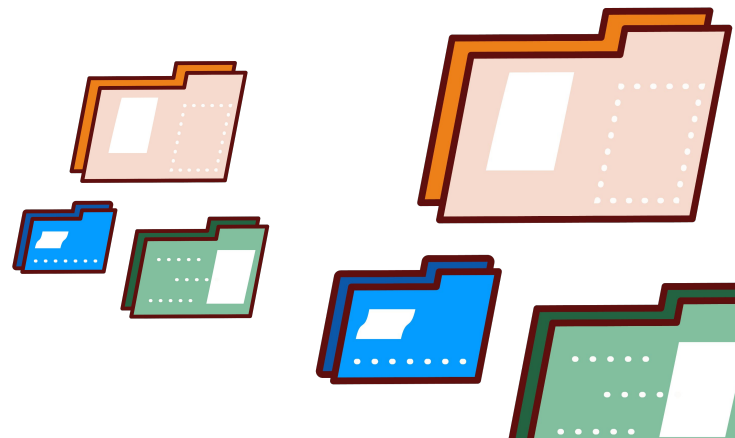# We need to continue addressing the different aspects of the problem

- A lot of data
- A lot of queries
- A lot of edits
- Unmaintained Blazegraph

A lot of data

- **Move large specialized data out of Wikidata (e.g. Wikibase Ecosystem)**
- **Continue to reduce redundant data (e.g. mul language code, automated descriptions, Lua modules for inverse relations)**
- **Split the graph**

A lot of queries

- **Move people to more appropriate access methods to reduce the load**
  - **Make it easier to work with other existing access methods (e.g. subset dumps, Query Service on cloud providers)**
  - **Provide additional access methods**
  - **Automatically reroute queries to more appropriate systems**
  - **Automatically rewrite inefficient queries**
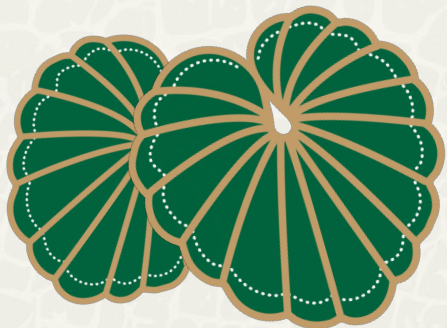 - **Increase the incentives and pressure to move to other access methods**

WIKIMANIA
SINGAPORE

A lot of edits

- **Reduce redundant data and the unnecessary edits that come with them**

Unmaintained Blazegraph

- **Evaluate alternatives and move away from Blazegraph**
- **Evangelize for the development of new graph backends**

More questions? Want to stay up-to-date?

Wikidata's weekly newsletter

Search Platform team office hours

lydia.pintscher@wikimedia.de

@nightrose

User:Lydia Pintscher (WMDE)

WIKIMANIA SINGAPORE

WIKIMANIA
SINGAPORE