Wikidata’s data access methods

What we found out and what we’re going to do with it

Ifrah Khanyaree
Product Manager
ifrah.khanyaree@wikimedia.de

Daria Ammalainen
UX Researcher
daria.ammalainen@wikimedia.de

Lydia Pintscher
Wikidata Portfolio Lead
lydia.pintscher@wikimedia.de

This session is recorded: Please mute your microphone and camera when you’re not speaking.

DATA REUSE DAYS February 18-27_2025

The problem

DATA REUSE DAYS February 18-27_2025

Wikidata’s data is there to be used!

We want more people to use Wikidata’s data
to build meaningful applications and services.

DATA REUSE DAYS February 18-27_2025

- Many ways to get to Wikidata's data for developers - all with benefits and drawbacks for different use cases
- Confusion about which to use when
- Wikidata Query Service is overloaded and we need to take load off of it

-> We need to understand how to improve the other access methods and how to move people to the most appropriate one for their use case

What we did

DATA REUSE DAYS February 18-27_2025

- Classified a sample of queries
- Gathered data about usage of REST and action API
- Survey (70 responses)
- Interviews (12 sessions)

What we found out

DATA REUSE DAYS February 18-27_2025

General data access

- Most developers are using a **combination of access methods** to build their application
- Developers are using the Query Service for simple requests instead of other methods because WDQS allows them to do it in **one request instead of several** or **gives a more specific response** (e.g. some statements instead vs whole Item)

General data access

The missing middle

We have systems to cover:

- **Direct access** to all or partial data of an entity
- **Querying** more complex connections in the graph
- Large-scale **analysis**

A lot of use cases need (a little bit) more than the current APIs provide but less than the full power of SPARQL -> Build out existing APIs and offer additional ones if really needed

General data access

- Documentation and discoverability of documentation is still a problem
  - Findability of the new developer portal
  - Query examples diversity and complexity

General data access

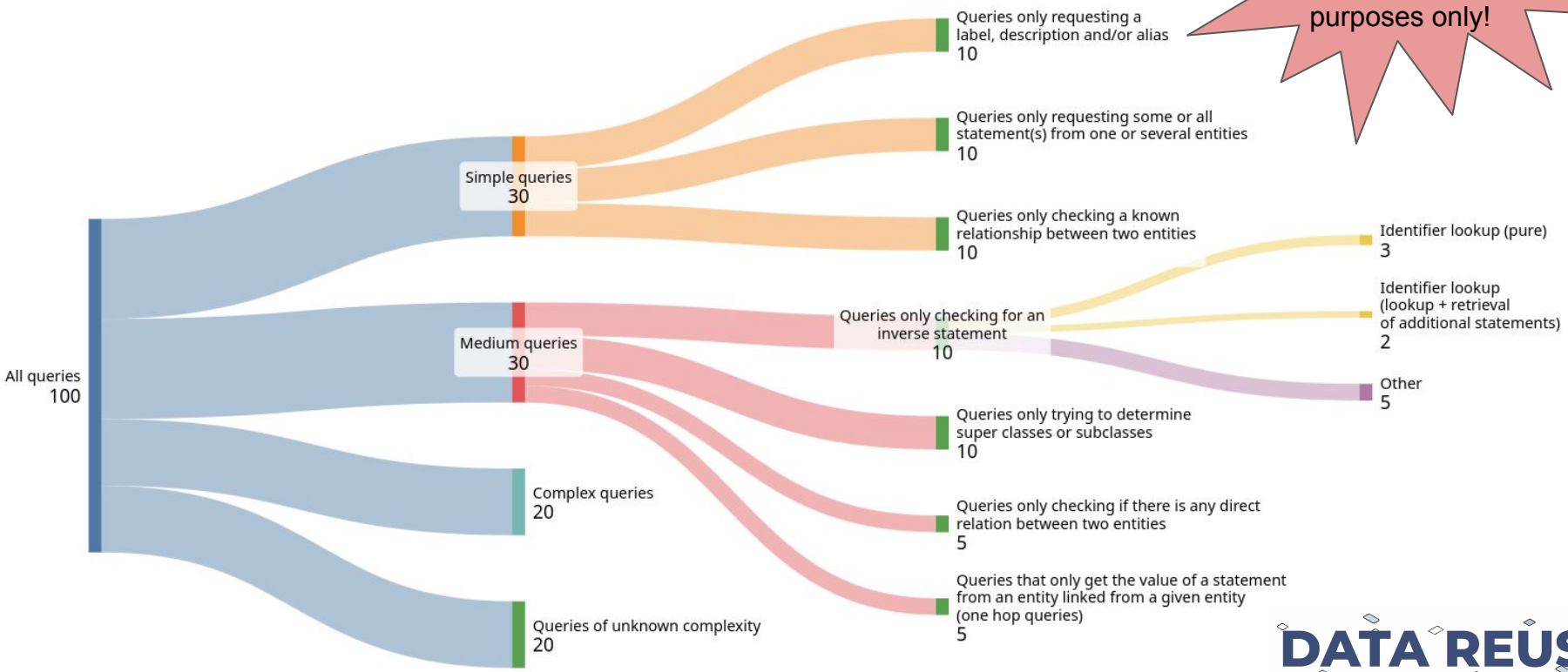
- Surprisingly few developers in the interviews thought about it as **using “Wikimedia content”** in combination with for example Wikipedia articles
- We do not have good enough **communication channels** to reach developers
- Developers are often not setting meaningful user agents, making it harder to understand what is happening or contact them in case of issues
- Many developers we interviewed are **giving back** - not representative of the wider reuser community!

Wikidata Query Service

- **Used by almost all** interviewees and survey participants
- A number of queries that don’t need the full power of SPARQL are for getting **labels of linked entities**
- **Missing real-time updates** are a barrier to switching to 3rd-party endpoints - lag in the order of minutes or hours are ok, but not days

Classification buckets

Numbers are made up for illustration purposes only!



Example classification of queries

Here's three types of queries that ideally shouldn't be run using the WDQS:

```
SELECT ?itemLabel
WHERE {
wd:Q1530 rdfs:label ?itemLabel.
FILTER (lang(?itemLabel) = "en")
}
```

Simple query asking for an Item label in a specific language

Simple query asking for a statement, label etc of a Property

```
SELECT (STRAFTER(STR(?property), 'entity/') as ?id) ?property ?propertyType ?propertyLabel ?propertyDescription
?propertyAltLabel (STRAFTER(STR(?propertyType), '#') as ?value_type) ?formatter_url WHERE { VALUES (?property) {
(wd:P11683) } ?property wikibase:propertyType ?propertyType . OPTIONAL { ?property wdt:P1630 ?formatter_url. } SERVICE
wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } ORDER BY
ASC(xsd:integer(STRAFTER(STR(?property), 'P')))
```

Example classification of queries

```
SELECT DISTINCT ?wd WHERE {
VALUES ?wd { wd:Q12144080 wd:Q12145367 wd:Q12147464 wd:Q12148979 wd:Q12150515 wd:Q12151833 wd:Q12153482 wd:Q12154846
wd:Q12157500 wd:Q12159187 wd:Q12160375 wd:Q12161784 wd:Q12163007 wd:Q12164732 wd:Q12166022 wd:Q12166889 wd:Q12167780
wd:Q12168597 wd:Q12169822 wd:Q12170570 wd:Q12172924 wd:Q12174661 wd:Q12177364 wd:Q12181489 wd:Q12184956 }
?wd ?p ?o.
FILTER EXISTS { ?wd ?p ?o }
}
```

Asking for a list of Items that meets a simple filtering criteria

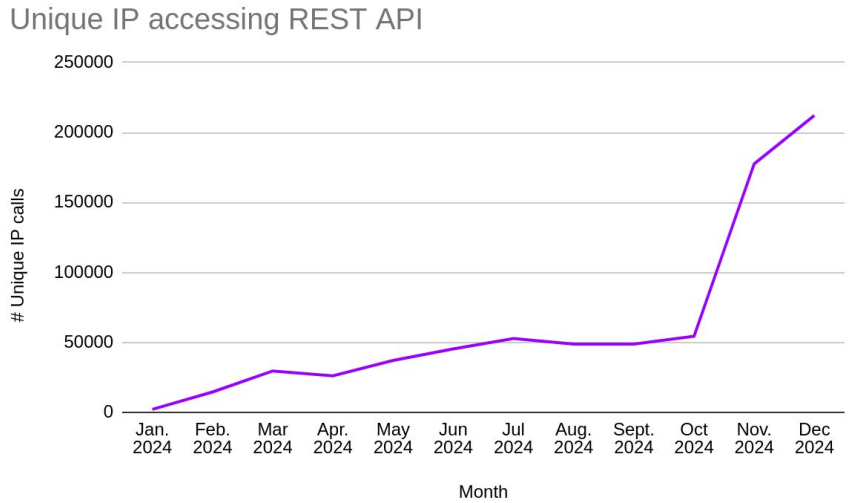
Dumps

- Dumps would be an alternative to WDQS in some cases but are **too large** and **not generated frequent enough**
    - Subset dumps were a frequent request

Some quantitative data on the APIs

REST API

- Over the last 3 months, there have been approximately **130 million requests** to the REST API
- As compared to a year ago, the API gets **90 times as many** requests!
- The most used endpoints are **GET item,** followed by **GET item label** and **GET property label**



Some quantitative data on the APIs

Action API

- Over the last 3 months, there have been approximately **3-4 billion** requests to the Action API
- The most used endpoints are **query**, **parse** and **wbgetentities**. If filtered by Wikibase-specific endpoints, then following **wbgetentities** is **wbsearchentities** and **wbformatvalue**

What’s next

DATA REUSE DAYS February 18-27_2025

Upcoming topics for discovery or development

In development

- Search in REST API
- WDQS Graph split
- Embeddings for ML use

Ongoing

- Developer advocacy

To be picked for discovery

- Alternatives to simple & medium complexity queries in WDQS, such as Labels of linked entities or Bulk GET
- Subset dumps
- Action API scenario mapping

Thanks for your attention!

Get in touch with us:

Ifrah Khanyaree
ifrah.khanyaree@wikimedia.de

Daria Ammalainen
daria.ammalainen@wikimedia.de

Lydia Pintscher
@nightrose
lydia.pintscher@wikimedia.de

DATA REUSE DAYS February 18-27_2025