What We Have Learned from Large Commercial Users

Understanding how and why Wikimedia data is re-used in third party environments

or

What the Wikimedia Enterprise team has learned from talking to big tech companies (so that you don't have to)

WIKIMANIA SINGAPORE

Lane Becker / LBecker (WMF)
Liam Wyatt / Lwyatt (WMF) / Wittylama

But first, some helpful context

Who and what is Wikimedia Enterprise?

The Wikimedia Enterprise team is a WMF team that that has **built an API platform specifically designed for large-scale commercial re-users of Wikimedia content**, which transfers high volumes of content at high speeds. Users operating at this scale **pay for access to the platform** as well as for contractual guarantees of stability, uptime, and support services. **The content served via Enterprise is no different than what's available across our public APIs.**

You can learn more about this work — technical, legal, financial, and strategic — via the homepage on Meta. See: [[m:Wikimedia Enterprise]].

Lessons learned & caveats applied

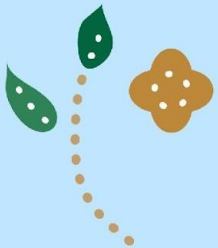
We're going to get detailed, but not too specific

In order to develop a set of APIs for large commercial re-users, we met with many of them, multiple times, in order to learn:

- What they are **currently** doing with Wikimedia content;
- What they are **unable** to do with Wikimedia content;
- What they think is **important** about Wikimedia content; and
- What they **want** to do with Wikimedia content.

In this presentation, we will summarize these lessons, but we won't:

- Name specific companies or describe their infrastructure; or
- Make value judgements about them (or Wikimedia). Just the facts.



Our starting point

The Wikimedia Enterprise grew out of a **dual requirement** put forth in the *Movement Strategy*:

“Increase the sustainability of our movement” and **“Improve user experience.”**

Today we’re talking about the latter, although it’s intimately intertwined with the former.

WIKIMANIA
SINGAPORE

- **All large tech companies use Wikimedia content extensively.**
- **Commercial use of Wikimedia data is already incredibly widespread** thanks to our public APIs and permissive licenses.
- **Most large tech companies vacuum up all the data they can get**, store it, and have large internal teams that normalize it for use across their services.
- **None of these tech companies fully understand the environment in which the content is created**, which limits their ability to make the best use of it.

Who are these customers, anyway?

Three broad categories of commercial re-users of Wikimedia content:

1. **Big Tech**
    - GAFA? MAGMA? FAANG? MANGA?
    - Have a **core use case around real-time search**
    - Numerous other secondary uses, from **bios to images to maps**
2. Everyone else
    - From Fortune 100 down to small commercial tech shops
    - Use Wikimedia data for **search results, reference tools, and topic lookups**
3. Artificial Intelligence (AI) companies
    - The newest and least developed/understood market
    - Have a different use case, around **Large Language Model (LLM) training data**

What big tech companies are currently doing with Wikimedia data

WIKIMANIA
SINGAPORE

Already...

All of this takes *a lot* of work on their side

WIKIMANIA SINGAPORE

1. **Scraping, copying, and querying multiple Wikimedia APIs**; normalizing the data sourced from these different methods; and **storing this data in their own proprietary knowledge graph**, alongside data gathered from other sources, also stored in the same format.

2. **Particularly need to understand what has changed in the world recently — also know as "news."** It's how they ensure that when one of their users asks about a topic, they're able to understand the context of the ask.

3. **Holding back publication of anything they suspect might be vandalism**, using a combination of internally developed heuristics (sometimes based on flawed assumptions about how Wikimedia operates) as well as data provided by WMF.

4. When different language editions disagree on facts, companies **generally prioritize based on pageviews.** Lacking native pageviews, Wikidata gets deranked.

What big tech companies are unable to do with Wikimedia data

WIKIMANIA
SINGAPORE

Unfortunately...

The way Wikimedia data is structured, it's very hard for most companies to use

WIKIMANIA SINGAPORE

1. **Wikimedia API content is almost entirely unstructured, which means it's not very machine readable**. Articles are presented as one big text field in Wikitext or HTML, which requires significant resources to parse on the customer side in order to reuse, and even the biggest companies don't get it right 100% of the time.

2. **Our public APIs weren't designed to work together, and so they don't work together.** Data formats are inconsistent and different APIs provide different types of data. Retrieving an entire article and associated metadata takes three to five API calls to our public APIs.

3. **Infoboxes and tables, where editors put significant effort into structuring content for public consumption, are, ironically, the hardest things for machines to read.** Most third party re-users who have invested in the tech necessary to parse our content **have given up on using the data contained in infoboxes and tables.**

What big tech companies think is important about Wikimedia data

WIKIMANIA
SINGAPORE

Importantly...

Wikimedia data has immense value to these organizations

WIKIMANIA SINGAPORE

1. **It is their biggest source for real-time information about what is happening in the world.** It is also in some cases their only source for real-time information about the world, particularly for some of the smaller language editions.

2. **It has more information in more languages than any other source of data they have access to**, and the quality and accuracy of the data is quite high.

3. **It has licenses that are permissive and give them a level of control over the content we provide that is unlike most of their other data sources**, which have much more restrictive data usage and storage requirements.

4. **The flexibility and permissiveness of the content licenses contributes significantly to their innovation capacity.** AI/LLM training models are the most recent example of this, as our Creative Commons licensing allowed for this novel usage of the content.

WIKIMANIA
SINGAPORE

But! Even while acknowledging how valuable Wikimedia data is to them, big tech companies consider the *actual* value of the data to be nothing. $0.00.

Because that is how much it costs them to obtain it. Additional costs to store, process, or maintain the data don't factor in. The definition of value is cost to acquire the data — full stop.

What big tech companies want more of from Wikimedia content

WIKIMANIA
SINGAPORE

If only...

Wikimedia APIs could improve data access in these specific ways

WIKIMANIA SINGAPORE

1. **Better data usability.** Better design of all WMF APIs to make it easier to search, sort, and filter against the data, in whatever format, with fewer API calls. Documentation to help customers understand how the content was created, in order to make better use of it in other contexts.

2. **Improved machine readability.** Content extracted from the full article and provided in a machine-readable format; integration of Wikidata and Wikipedia for a fuller picture of content available on a particular topic.

3. **Clarity regarding content integrity.** Metadata to help companies understand which edits are credible and which are vandalism, critical when trying to integrate information about changes in the world (“news”) in as close to real-time as possible.

4. **Content formatted for AI training use.** Content structured in a format that’s easiest for LLMs to consume, with no personally identifiable information (PII) included.

WIKIMANIA
SINGAPORE

Our product roadmap has been designed around these needs.

- **“Breaking news”** to identify notable activity in any particular time frame;
- **“Credibility signals”** to support real-time decision-making around credible content, based on multiple community sourced data points;
- **Parsing of infoboxes and tables** for improved machine readability; and
- **Integration of Wikidata** alongside other text-based projects in Enterprise APIs.

All data points sourced from community activity, and none imply value judgments in any way. Content re-users can take or leave as much of this metadata as they see fit to inform their own decision-making processes.

Thank you for listening!

Questions?

WIKIMANIA
SINGAPORE