

Meta <https://w.wiki/3DXc>
MediaWiki https://w.wiki/_zpWQ



Lane Becker
Business Development



Ryan Brounley
Product Management

Technology companies across the planet **rely extensively on Wikimedia data throughout their multi-billion dollar product suites**

- Wikimedia data is widely used to populate many organizations' knowledge graphs, which contribute to billions of dollars of revenues across multiple industries
- Search and digital assistants are common consumers of these graphs
- Other platforms that consume these graphs — maps, ads, etc. — are also \$1B+ businesses



WIKIMEDIA
FOUNDATION

**But! Maintaining
Wikimedia data
access and quality
is very expensive
due to commercial
limitations**

- Large enterprises expend significant resources to maintain their data ingestion pipelines
- Smaller enterprises can't afford to
- Data quality is never guaranteed
- Data is inconsistent between APIs
- Delivery methods lack clear indicators of the data's actual value to the consumer
- Multiple engineers per company spend significant time managing this data as a result



Companies will pay for a **commercial data service** that serves their needs directly, consistently, and reliably

- We will provide a commercial platform that explicitly addresses the problems and requests of our largest commercial data consumers
- Operating as a separate platform and company allows us to tune functionality and services to the specific needs of commercial consumers without interfering with the function of the Foundation or the community



Wikimedia



**Wikimedia
Enterprise**

Movement Strategy Recommendations

Increase the sustainability of our movement:

"Explore new opportunities for both revenue generation and free knowledge dissemination through partnerships and earned income—for example...Building enterprise-level APIs"

Improve user experience:

"Make the Wikimedia API suite more comprehensive, reliable, secure and fast, in partnership with large scale users.... and improve awareness of and ease of attribution and verifiability for content reusers."



WIKIMEDIA
FOUNDATION

A data delivery platform that delivers what commercial consumers of our data need most

- An enterprise-level product to be released in 2021
- Baselined against existing methods of data access to make transitioning (relatively) seamless
- Maintaining the already established value of Wikimedia data to these customers
- Providing additional value by delivering data with higher frequency, guarantees of reliability, and better context

The Wikimedia Enterprise **service** **offering** will include:

- Dedicated, on-demand customer support available via phone, email, and chat
- A Service Level Agreement (SLA) guaranteeing a reliable percentage of uptime
- Engineering consulting services to help customers ensure they're getting the most value out of their use of Wikimedia data



Wikimedia Enterprise (WME) APIs - Q2/Q3 2021 Release

Type	Name	What is it?	What's New?
Realtime	WME Activity Firehose API	A stable, push HTTP stream of real time activity across “text-based” Foundation Projects	<ul style="list-style-type: none">• <i>Push</i> changes to client with stable connection• <i>Filter</i> by Project and Page-Type• <i>Machine Readable and Consistent</i> JSON schema• <i>Guaranteed</i> uptime, no rate-limiting
	WME Structured Content API	Recent, machine readable content from all “text-based” Foundation Projects	<ul style="list-style-type: none">• <i>Machine Readable and Consistent</i> JSON schema• <i>Guaranteed</i> uptime, no rate-limiting
Bulk	WME Bulk Content API	Recent, compressed Foundation data exports for bulk content ingestion	<ul style="list-style-type: none">• <i>Machine Readable and Consistent</i> JSON schema• <i>Daily</i> “Entire Corpus” exports• <i>Hourly</i> “Activity” exports• <i>Guaranteed</i> delivery• <i>Historical</i> Downloads

WME 2021 Roadmap Considerations

Theme	Feature	Details
Machine Readability	Parsed Wikipedia Content	Break out the HTML and Wikitext content into clear sections that customers can use when processing our content into their external data structures
	Optimized Wikidata Ontology	Wikidata entries mapped into a commercially consistent ontology
	Wikimedia-Wide Schema	Combine Wikimedia project data together to create “single-view” for multiple projects around topics
	Topic Specific Exports	Segment corpus into distinct groupings for more targeted consumption
Content Integrity	Anomaly Signals	Update schema with information guiding customers to understand the context of an edit. Examples: page view / edit data
	Credibility Signals	Packaged data from the community useful to detect larger industry trends in disinfo, misinfo, or bad actors
	Improved Wikimedia Commons license access	More machine readable licensing on Commons media
	Content Quality Scoring (Vandalism detection, “best last revision”)	Packaged data used to understand the editorial decision-making of how communities catch vandalism

Enterprise API to download daily exports, hourly diffs, or single articles all in JSON.

GET /v1/diffs/json/{date}/{db_name} Returns zip file with a specified date's project revisions in JSON

Hourly updated bundle of revised pages starting at 00:00 UTC each day.

Parameters

Name	Description
date * required string (path)	Date of the diff in YYYY-MM-DD <input type="text" value="2021-04-20"/>
db_name * required string (path)	Project DB name <input type="text" value="enwiki"/>

Responses Response content type





















Curl

```
curl -X GET "https://api.wikimediaenterprise.org/v1/diffs/json/2021-04-20/enwiki" -H "accept: application/json"
```

Request URL

```
https://st-okapi-data-bk.s3.us-east-2.amazonaws.com/diff/enwiki/2021-04-20/enwiki_json.tar.gz?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAMU4DBC-east-2%2Fs3%2Faws4_request&X-Amz-Date=20210421T014857Z&X-Amz-Expires=60&X-Amz-SignedHeaders=host&X-Amz-Signature=207bd100b1f06f6325c9445e46ab9154f62589dd7
```

Daily Exports of all text-based Wikimedia Foundation Projects

Name	Size	Last Modified
 !!!_json	14 KB	4/19/21 11:12:43 PM EDT
 "Litoria"_castanea.json	51 KB	4/19/21 11:13:09 PM EDT
 "Weird_AI"_Yankovic.json	67 KB	4/19/21 11:13:12 PM EDT
 \$pringfield_(Or,_How_I_Learned_to_Stop_Worrying_and_Love_Legalized_...	22 KB	4/19/21 11:13:03 PM EDT
 '03_Bonnie_&_Clyde.json	42 KB	4/19/21 11:13:14 PM EDT
 'Amran_Governorate.json	41 KB	4/19/21 11:13:12 PM EDT
 'N'_Dey_Say.json	13 KB	4/19/21 11:13:09 PM EDT
 'N_Sync_(album).json	63 KB	4/19/21 11:13:01 PM EDT
 'Round_Springfield.json	150 KB	4/19/21 11:12:52 PM EDT
 'Splosion_Man.json	15 KB	4/19/21 11:12:50 PM EDT
 's-Gravendeel.json	7 KB	4/19/21 11:13:05 PM EDT
 's-Hertogenbosch.json	92 KB	4/19/21 11:13:07 PM EDT
 (307261)_2002_MS4.json	16 KB	4/19/21 11:12:54 PM EDT
 (524522)_2002_VE68.json	24 KB	4/19/21 11:13:01 PM EDT
 (55565)_2002_AW197.json	24 KB	4/19/21 11:12:52 PM EDT
 (Everything_I_Do)_I_Do_It_for_You.json	47 KB	4/19/21 11:13:06 PM EDT
 (I've_Had)_The_Time_of_My_Life.json	83 KB	4/19/21 11:12:42 PM EDT
 (Shake,_Shake,_Shake)_Shake_Your_Booty.json	6 KB	4/19/21 11:12:43 PM EDT
 (There's)_Always_Something_There_to_Remind_Me.json	19 KB	4/19/21 11:12:53 PM EDT
 (What's_the_Story)_Morning_Glory?.json	75 KB	4/19/21 11:12:54 PM EDT
 (You_Drive_Me)_Crazy.json	34 KB	4/19/21 11:12:50 PM EDT
 +44_(band).json	19 KB	4/19/21 11:13:11 PM EDT
 +_(album).json	20 KB	4/19/21 11:12:57 PM EDT
 -hou.json	20 KB	4/19/21 11:13:04 PM EDT
 ...And_Justice_for_All_(album).json	36 KB	4/19/21 11:13:02 PM EDT
 ...Baby_One_More_Time_(song).json	38 KB	4/19/21 11:12:45 PM EDT
 .223_Remington.json	5 KB	4/19/21 11:13:10 PM EDT
 .30-06_Springfield.json	9 KB	4/19/21 11:13:10 PM EDT
 .30_Carbine.json	24 KB	4/19/21 11:12:46 PM EDT
 .357_Magnum.json	25 KB	4/19/21 11:12:49 PM EDT
 .44_Magnum.json	9 KB	4/19/21 11:13:05 PM EDT

The structure of a page.... for now.

title:	"Douglas_Adams"
pid:	8038
qid:	"Q42"
revision:	7085101
dbName:	"simplewiki"
inLanguage:	"simple"
▼ url:	
canonical:	"https://simple.wikipedia.org/wiki/Douglas_Adams"
dateModified:	"2020-08-27T23:01:16Z"
▼ articleBody:	
▶ html:	"<!DOCTYPE html><html pre...</section></body></html>"
▶ wikitext:	"{{Infobox person honori...ry:Writers from London}}"
▼ license:	
0:	"CC BY-SA"



...mapped right
over to
Wikidata

Item [Discussion](#) [Read](#) [View history](#)

Douglas Adams (Q42)


English writer and humorist
Douglas Noel Adams | Douglas Noël Adams | Douglas N. Adams

[▼ In more languages](#)
[Configure](#)

Language	Label	Description	Also known as
English	Douglas Adams	English writer and humorist	Douglas Noel Adams Douglas Noël Adams Douglas N. Adams
Spanish	Douglas Adams	escritor y humorista británico	Douglas Noel Adams Douglas Noël Adams
Traditional Chinese	道格拉斯·亞當斯	英國作家	
Chinese	道格拉斯·亞當斯	英國作家	亞當斯

[All entered languages](#)

Statements

instance of	human ► 2 references
image	

Sample WME Structured Content Page Schema

BASELINE SCHEMA: *Benjamin Franklin*

```
{
  "qid": Q34969,
  "pid": 12345,
  "title": "Benjamin Franklin",
  "dbName": "enwiki",
  "inLanguage": "en",
  "revision": 94302,
  "namespace": 0,
  "dateModified": ...,
  "url": {
    "canonical": "https://en.wikipedia.org/wiki/Benjamin\_Franklin"
  },
  "content": {
    "license": "CC BY-SA",
    "html": "STRING BLOCK",
    "wikitext": "STRING BLOCK"
  }
}
```

Ways to make this schema more useful:

- Parsed content
- Anomaly Detection
- Page Trending Information
- Credibility Signals
- Wikimedia Projects of similar topics
- Commons links in articles
- Quality Scoring
- Topic information
- More elaborate edit information
- And more....

Enterprise.Wikimedia.com

[[Wikimedia Enterprise]]

Meta: FAQ, Principles, Essay

MediaWiki.org: Documentation, Roadmap, Updates

Phabricator.wikimedia.org: Workboard

Github.com/wikimedia/OKAPI: Website, Service

"Office hours": third Friday of each month @ 15:00 UTC



WIKIMEDIA
FOUNDATION

CC BY-SA 4.0