

Structured Data Across Wikimedia: A Third-Year Report for the Sloan Foundation

Sloan Grant #G-2020-12665. September 1, 2023. Prepared by Carly Bogen, Amanda Bittaker, and Jonathan Curiel.

Executive Summary

Three years ago, we outlined a plan to continue our work with structured data on the Wikimedia projects. Designed to advance the progress we'd just made on Wikimedia Commons, our new project had a name — Structured Data Across Wikimedia — that spoke to a far-reaching vision: We'd use structured data to give users a better way to read, edit, and access knowledge across the Wikimedia projects, and we'd give users a more inviting, more efficient way to search and find content across the same projects. In this third-year report to the Sloan Foundation, we detail a year where we achieved even more of the vision we had in 2020. And we're happy to say we achieved it in the best way possible: By working even more closely with Wikimedia's global community of users and contributors, who've told us directly how important our new work is to their day-to-day doings on the projects.

In Year 3, our vision of Structured Data Across Wikimedia has come to fruition with practical tools that Wikimedia users around the world are employing in different languages to add knowledge to the Wikimedia projects and also to garner knowledge. Highlights of the past year include:

- **We developed Section Topics**, a data pipeline that identifies sections in a Wikipedia article and creates topics for those sections by using an algorithm that detects Wikidata items based on the section's blue links. Section Topics are now generating image suggestions for Wikipedia articles to thousands of users — including newcomers, a cross-section of mobile app users, and Wikipedia's most experienced contributors. The work we've done greatly advanced Year 2 of the project, when we completed Section Topics' initial research and design.
- **We continued to refine the project's image-suggestion features**, which are now helping both newcomers and experienced users add images to Wikipedia articles — as with a new tool that notifies experienced editors of image suggestions for articles on their watchlists. Like Section Topics, this work greatly advanced Year 2 of the project. Overall since January 2022, contributors have added more than 550,000 images to Wikipedia articles.
- **We made significant improvements to the search experience on Wikipedia**, as with a new feature called Search Preview that shows additional context for search results — including an extended article snippet; table-of-contents links that take the user directly to an article's sections; related images from Wikimedia Commons; and links to related content on other Wikipedias.

Search Preview increases the likelihood that users interact with relevant results, and increases the potential for discovery of adjacent relevant content.

- **Active contributors are now planning to take the work we did in Year 3 to improve the Wikimedia projects even more**, as with our image suggestions data pipeline, which may soon advance [a depictor tool](#) that easily adds [depict statements](#) to Wikimedia Commons images; and as with the Digital Public Library of America's work that will generate more items in MediaSearch results on Commons.

All this and more happened because we evolved what we did with Structured Data on Commons into something much bigger — with a broader scope; an extended commitment from the Sloan Foundation; a deeper layer of engagement with volunteer developers and the Wikimedia community; and a more advanced network of people working from across the Wikimedia Foundation. Besides the core Structured Data team, Year 3 incorporated work from the Wikimedia Foundation's Search team, who contributed to such parts as search improvements and image suggestions; the Data & Platform Engineering team, who contributed to such parts as section topics and image suggestions; the Android team and the Growth team, who contributed to image suggestions; the Research team, who contributed to such parts as topic models and edit types; and the GLAM team, who worked with communities to create impact.

Working together, we were able to iterate Structured Data Across Wikimedia into something tangible. As with the previous two years, some of what we iterated led to unpredictable results or even a realization that we shouldn't pursue a particular project part. But overall, what we see at the end of Year 3 is what we thought we'd see when we first proposed Structured Data Across Wikimedia: An advanced state of the Wikimedia projects that uses structured data to better those projects and connect them better to the wider internet.

Some of the work we detail below is fairly technical since it conveys the important step-by-step configurations that went into our project's success. But please know this: Without those technical machinations, Structured Data Across Wikimedia wouldn't be where it is today. Another reality: The project wouldn't be what it is without your support— so thank you. Thanks for your backing not just over the past three years but for the previous three years that we worked so hard on Structured Data on Commons. The intricate work we do behind the scenes at the Wikimedia Foundation — including the three years we've spent with Structured Data Across Wikimedia — impacts millions of people around the world. In Structured Data Across Wikimedia's third year, we heard from people representing those millions — including an editor greatly benefiting from our new image-suggestion features, which let people who are organizing campaigns or events target topics for impact. This editor attended an event that focused on filling knowledge gaps, such as illustrating articles about women and LatinX people, and the editor said the image-suggestion features saved her valuable time — and inspired invaluable work that improved Wikipedia. Their comment embodies our work's success in Year 3. “Finding articles and images takes a LONG time, if you find one at all to edit,” the editor told us, “and to have that part eliminated just made me feel like editing, editing, editing non-stop.”

Table of Contents

Executive Summary 1-2

Year 3: The Project’s Progress 3-35

 Snapshot of our work 3-4

 Adding structured metadata to a longform content 4-8

 Connecting topically related content across languages 8-17

 Redesign of the search experience 17-24

 Design and launch community-based pilot projects 24-33

 Experiment with additional features to suggest relevant content 33-37

Conclusion 37-38

Budget Report 38

Year 3: The Project’s Progress

A Snapshot of Our Work from April 1, 2022 to June 30, 2023

What We Envisioned for Year 3	What We Did
Add structured metadata to long-form content on one Wikimedia project	<ul style="list-style-type: none"> • We developed Section Topics, a data pipeline that identifies sections in an article and creates topics for those sections, using an algorithm that detects Wikidata items based on the section’s blue links. • We experimented with using Section Topics data on Wikipedia pages to improve Google Search Engine Optimization (SEO) results.
Experiment with connecting topically related content across languages as part of the contribution process	<ul style="list-style-type: none"> • We continued refining the image suggestions features, both for newcomers and for experienced users. We took advantage of the Section Topics infrastructure to recommend images that could illustrate sections of Wikipedia articles, increasing illustration and readability.
Redesign of the search experience for new users on projects beyond Commons	<ul style="list-style-type: none"> • We made impactful improvements to the search experience on Wikipedia, and partnered with community members who built tools that take advantage of our Structured Data Across Wikimedia infrastructure.
Design and launch 2–5 community-based pilot projects focused on experimenting with and encouraging adoption of new features	<ul style="list-style-type: none"> • We worked with community members to launch View It!, a tool that shows Wikipedia users relevant Wikimedia Commons media depicting—or otherwise related to—the article they are reading. • We partnered with the Digital Public Library of America

	(DPLA) and their partners and communities to get more images described using structured data, which will allow us to have more images available in the image suggestions pipeline and generally drive the reuse of described and attributed images.
Experiment with additional features to suggest relevant content to readers and editors during the article edit process	<ul style="list-style-type: none"> We launched Wikistories in Beta Indonesian Wikipedia, a mobile web-based tool that empowers editors to create a new content format: short, visual and reliable knowledge for quick consumption and easy sharing.

Add structured metadata to long-form content on at least one Wikimedia project

Deliverable: Build infrastructure and tools to allow structured metadata from Wikipedia Commons to be added to other content across Wikimedia projects, including Wikipedia itself.

The need: Users from a larger global audience can read Wikimedia content through improved support for new devices and platforms (hovercards, feature phones, chat bots, etc.), especially in emerging markets and on mobile.

The impact: An increase in usage on new platforms, along with a measurable increase in new users reached.

In Year 3, we completed the work underway in Year 2 on adding structured metadata to long-form Wikipedia content – the Section Topics project. Section Topics identifies sections in an article and creates topics for those sections, using an algorithm that detects Wikidata items based on the section’s blue links.

As explained in the Year 2 report, the Section Topics project has the following elements:

- An algorithm that detects the Wikidata topics in an article based on blue links – which are the links that take readers to other Wikipedia articles;
- The ability to automatically identify where sections break in an article;
- Section-level image suggestions, which use the blue-links algorithm and section identification infrastructure above, and are delivered both via the newcomer experience and via notifications for experienced contributors. This built upon the prior image suggestions work, which we detailed in the previous years’ reports;

- Potentially using section topics to improve our SEO reach with outside search engines such as Google, increasing people's access to knowledge within their existing digital environment.

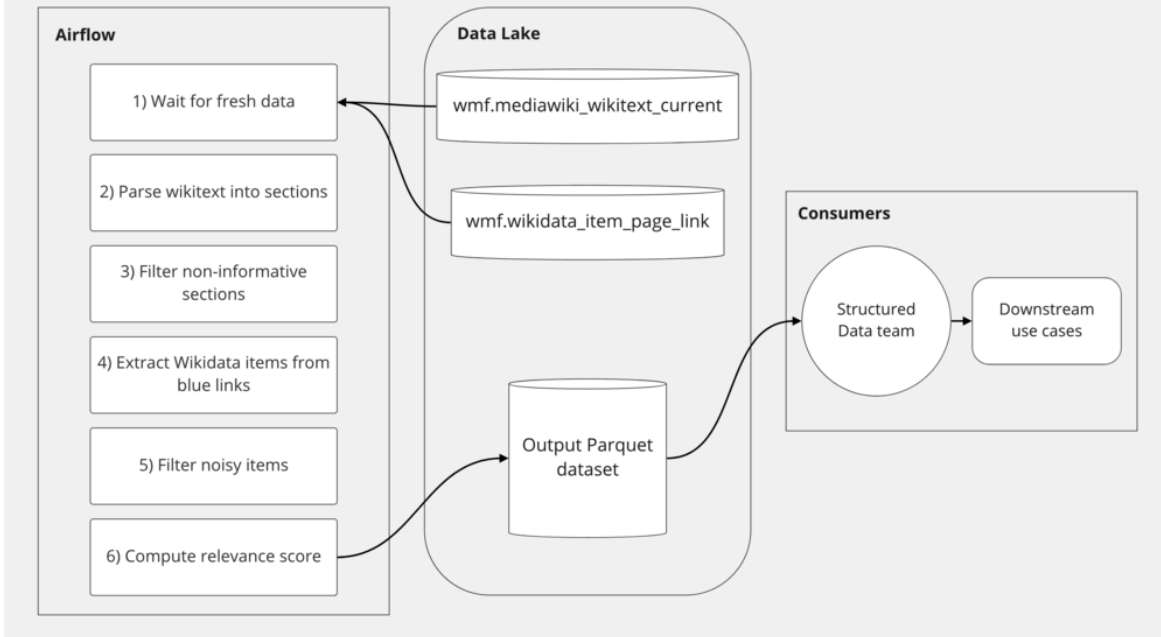
In Year 2, we completed initial research and design for creating structured Section Topics; we received valuable feedback from the communities on their moderation needs for Section Topics; and in a related project we added structure to talk pages to enable more intuitive on-wiki communication for newcomers and experienced contributors alike. In Year 3, we released Section Topics to production, and began using it to power new features.

The Section Topics data pipeline is implemented as an [Apache Airflow](#) job. Airflow is an open source workflow job scheduler, which automatically runs code in a sequence. This takes the burden off of individual software developers to run a script each time it's needed, allowing for much greater automation. For Section Topics, we implemented the following data processing flow:

1. gather the content of top-level sections, lead section included;
2. filter out sections that don't convey relevant topics, such as *External links*;
3. extract Wikidata items from wikilinks: the section topics themselves;
4. filter out noisy topics, such as dates;
5. Resolve redirect pages to their intended destination;
6. compute topics relevance score.

This flow is visualized below:

Section Topics Data Pipeline Architecture



A row of data contains the following:

Date	Wiki	Revision ID	Page Wikidata ID	Page Title	Section Title	Section Topic Wikidata ID	Section Topic Title	Section Topic Score
2023-01-16	enwiki	1127523670	Q36724	Attila	Solitary kingship	Q3623581	Arnegisclus	1.13

Section Topic score is a relevance measure that measures to what extent a given topic helps summarize and understand a given piece of Wikipedia content. This enables topic ranking and is computed as a [term frequency-inverse document frequency](#) (TF-IDF) weight based on the distribution of topics. This is used to ensure that features powered by Section Topics only use those section topics that provide a strong enough match.

As is all our code, the Section Topics code is open source and can be found at <https://gitlab.wikimedia.org/repos/structured-data/section-topics> and https://gitlab.wikimedia.org/repos/data-engineering/airflow-dags/-/blob/main/platform_eng/dags/section_topics_dag.py.

Uses of Section Topics in Production

Section Topics are now used to generate section-level image suggestions for articles via the newcomer experience, mobile app structured tasks, and notifications for experienced contributors, and to experiment with using section topics to improve our SEO reach with outside search engines. The impact of Section Topics is made through these uses, which are detailed in the sections below.

Using Section Topics to improve Search Engine Optimization (SEO)

In the Year 2 report, we said that we may be able to use Section Topics data to the schema.org metadata of our pages, further improving SEO in Google Search results. As we reported, in Year 2 we did some initial experiments with this, which proved promising. We created an example report to provide an example of the metadata that would be produced by the addition of the following metadata on the page:

- Defined a main article
- Defines sub heading using “hasPart”
- Define thumbnail and images using "image" or "thumbnailURL"
- Defined reference link using "isBasedOnUrl"

Once the Section Topics pipeline was complete and in production, we decided to explore this further with an additional experiment. We created a small test Wikipedia on Toolforge with a selection of static html pages that were copied from English Wikipedia and that had section topic data added to the schema.org “hasPart” data on the page to indicate sections. The goal was to learn whether Google would use Section Topics data added to Schema.org data to enhance the search results, so we then submitted that test Wikipedia to Google for indexing. The data produced on the test Wikipedia for the Glastonbury Festival page looked like this:

```
"hasPart": [  
  {  
    "@type": "Article",  
    "name": "Glastonbury Festival - History",  
    "url": "https://en.wikipedia.org/wiki/Glastonbury_Festival#History",  
    "about": [  
      {  
        "@type": "article",  
        "name": "National Jazz and Blues Festival",  
        "url": "https://en.wikipedia.org/wiki/National_Jazz_and_Blues_Festival",  
        "mainEntity": "https://www.wikidata.org/wiki/Q3336901"  
      }  
    ]  
  }  
]
```

```
    },
    { ... }
  ]
}
]
```

After submitting the test wiki, we hoped to see some of the following changes in Google search results for a sample of modified articles/sections:

- When Searchers on Google type in a topic related to a section of an article, they see the article with a link to that section come up in the search results.
- Searchers on Google can see visual representations of headings of sections when searching for a section related topic.

Unfortunately, Google never indexed the test Wikipedia that we created for the experiment, so we were unable to confirm the results. In the future, we plan to experiment with adding this data to a live Wikipedia page, which will be much more likely to be indexed by Google.

Exploration of integration of concept metadata with anti-vandalism and quality control systems

In the Year 2 report, we also said we would explore an integration of concept metadata with anti-vandalism and quality control systems. We deprioritized this work in favor of the image suggestions projects described below, which we believed would have a bigger impact. We hope to return to it in the near future as another potential use of the Section Topics infrastructure.

Experiment with connecting topically related content across languages as part of the contribution process

Deliverables:

- **Build infrastructure and tools to allow structured metadata from Wikipedia Commons to be added to other content across Wikimedia projects, including Wikipedia itself.**
- **More, higher quality media added to content pages.**

The need: Users in emerging markets and on mobile can contribute using related content suggestions when editing and other computer-aided editing improvements.

The impact: An increase in editing in emerging markets and on mobile and an increase in effectiveness of editing in those places. An increase in image contributions from newcomers and users in emerging

communities. More illustrated Wikipedia pages, which we know receive 4x more views than unillustrated ones.

What changed?

In the Year 2 grant report, we reported that we changed the 3-year target of adding images to 5 million Wikipedia content pages. Based on analysis of editor engagement and articles, we determined there are barriers to reach this goal and that it will be more impactful to add fewer, higher-quality images within the grant timeline. We therefore aimed to increase content numbers more slowly over a longer time period, using the infrastructure and tools built during the grant period.

We explained our new plan, which had two target user groups with separate strategies and features:

- We will increase contributions from newcomers (especially newcomers with topical expertise) via a structured task that walks newcomers through the process, with a goal to increase the number of users who add imagery to articles.
- We also plan to increase contributions from users in emerging communities by inviting experienced users to the image-suggestions functionality on Wikipedias in languages where there is less illustration. Experienced users with over 500 edits are asked to evaluate an existing Commons image that is suggested for an unillustrated article and add it to the article if they determine it improves the content. This allows us to fill the imagery gap for editors and readers, and because this strategy has a higher human touch, it won't run afoul of community concerns about image quality or article over-illustration.

In Year 2, we released the first version of the newcomer structured task, and began building the infrastructure that allows experienced users to evaluate suggested images. In Year 3, we further developed the infrastructure for these tools, released article-level image suggestions for experienced users to production, and built the algorithms that power new section-level image suggestions tools for both newcomers and experienced users. We also ran several events to work with newcomers in underserved communities and teach them how to use the tools.

Image Suggestions Data Pipelines

The first element of that infrastructure is the image suggestions data pipeline, which is now running in production, both for article-level and section-level image suggestions. This pipeline delivers the data to the user from several custom built algorithms that power the suggestions. In Year 3, we expanded the pipeline built in Year 2 to make article level image suggestions to

include the Section Topics and Section-Level Image Suggestions algorithms, which now allows us to make accurate image suggestions at the section-level of Wikipedia articles.

The section-level image suggestions tools leverage two principal algorithms to generate suggestions: section alignment and [section topics](#). Given a language and a Wikipedia article section:

- the former retrieves images that already exist in the corresponding section of other languages;
- the latter, as described above, takes the section's wikilinks and looks up images that are connected to them via several properties, typically Wikidata ones.

The Section Alignment algorithm builds on top of the section topics data pipeline and aims at constructing a visual representation of the blue links available in Wikipedia article sections.


To achieve this, it follows two kinds of paths that connect a given blue link to a Commons image, namely:

1. Blue link → Wikidata item → Wikidata [image property \(P18\)](#) and [Commons category \(P373\)](#) → Commons image
2. Blue link → Wikipedia article's lead image

First, the algorithm identifies the target section of an article in a specific language, and then searches for similar sections in other languages that contain images, which we use as a source for our recommendation. The section alignment model then takes Section Topics as input and creates a language-agnostic representation of a section. This representation is an embedding, which is essentially a vector that allows us to measure the semantic distance between two sections. Using manually annotated data, we determine a threshold that indicates high similarity between sections. Finally, we sort all the images found in similar sections across languages based on the number of languages in which they appear. We use this as a score for our section-level image suggestions system. The work to develop this algorithm by our Research team was detailed in the paper "[Crosslingual Section Title Alignment in Wikipedia](#)," by D. Difallah, D. Saez-Trumper, E. Augustine, R. West, and L. Zia, presented at the 2022 IEEE International Conference on Big Data.

The following table displays a section-level image suggestion example that stems from both section alignment and section topics:

page title	section title	image
------------	---------------	-------

Boombox	Design	 <p data-bbox="805 558 1073 590">Ghettoblaster-family</p>
---------	--------	---

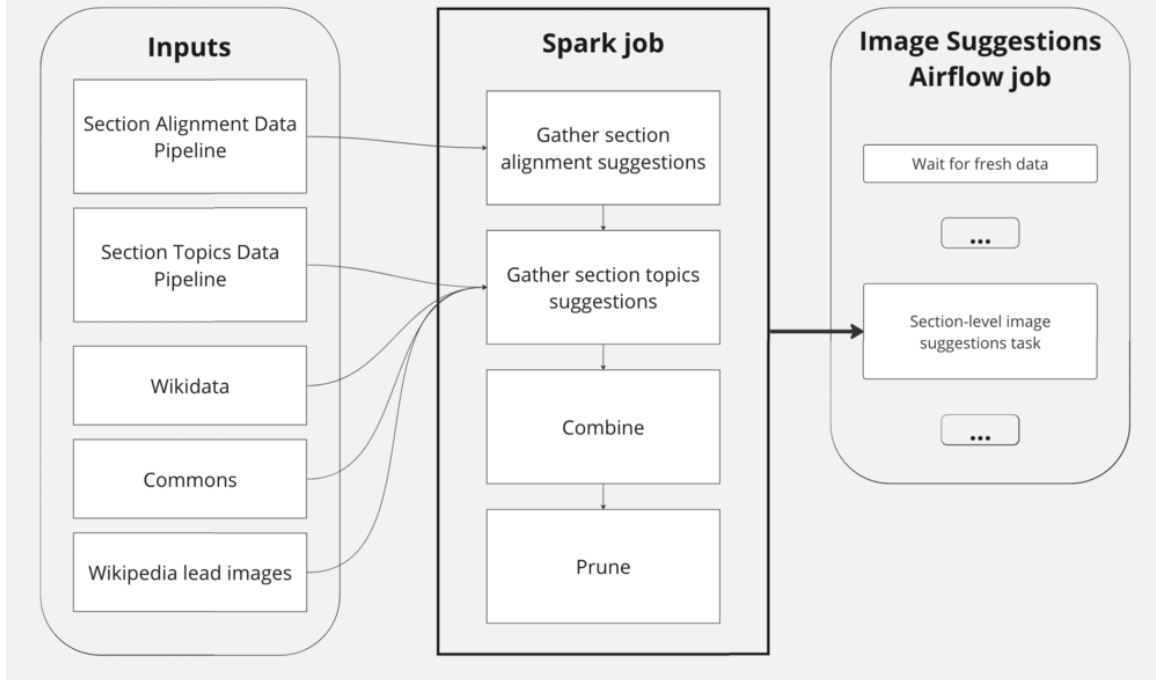
The section alignment algorithm found the image in Japanese Wikipedia's equivalent section, and the section topics algorithm obtained it through the following path:

[Sharp Corporation](#) section wikilink → [Sharp Corporation \(Q53227\)](#) Wikidata item → [Commons category \(P373\)](#) Wikidata property → [Sharp boomboxes](#) Commons category.

Lastly, every suggestion includes a confidence score, which is computed upon the following idea: the more sources agree on the same suggestion, the greater the confidence. This is supported by evidence: when we elicited human judgments over a random sample, results indicated a significant correlation between suggestions rated as good and their confidence scores.

A diagram of the algorithms and methods used to generate section-level image suggestions

Section-level Image Suggestions Data Pipeline Architecture



Algorithm Evaluation

Because the algorithms are used to provide suggestions to both experienced users and newcomers, we aimed to ensure that good suggestions were provided at least 70% of the time. We conducted several rounds of evaluation to review the accuracy of the image suggestion algorithm. Ambassadors from several language Wikipedias – community members we hired as contractors to help communicate about our goals with their communities – worked with us to perform this evaluation.

In the initial evaluation, suggestions did not meet our 70% threshold, averaging about a 62% accuracy rate. Many images were suggested in sections that shouldn't have images, or the image related to one topic in the section but didn't represent the section as a whole. Based on feedback from this evaluation, we continued to work on logic and filtering improvements to ensure suggestions were more accurate, and ultimately surpassed our goal, averaging about a 78% accuracy rate.

Remediating Bias

We also spent significant time doing our best to ensure that our data and algorithms will not cause harm or perpetuate inherent biases. There are several ways that our tools had the potential to perpetuate bias:

First, articles that are candidates for suggestions can bear existing Wikipedia selection biases, e.g., towards men and English-speaking countries. The system may also propagate existing bias in the articles that users select for their watchlists, as notifications for experienced users are based on those watchlists. We mitigate these types of bias by surfacing suggestions to end users for articles that do not represent those selection biases. Helpful examples include:

- topical filters and edit tags, selected for articles about underrepresented groups;
- incorporating [category-based notifications](#) into events run by local groups, such as the ones described in the “Image Suggestions Events” section below.

Another potential source of bias could be caused by the algorithm’s mapping of images from an article or section in one language to a similar article or section in another language. In this framework, acceptable images from one language community are also assumed to be culturally appropriate for others. We mitigated this by having users from different communities evaluate suggestions to see if they could be considered offensive, and the evaluation data showed that only 0.07% of suggested images were judged as offensive.

Lastly, section alignment's machine-learned component may suffer from language bias: multilingual language models are known to work better on languages with large presence on the Internet compared to under-resourced ones. For instance, the English-Spanish pair's measured precision is 95%, while the English-Japanese one is 83%. This translates into more confident suggestions for larger Wikipedias. We mitigate this by having the model learn section alignments for every Wikipedia project pair, thus enabling eventual expansion to new languages.

Further uses of the data

We have seen significant interest in using the data in new ways. For example, [User:Husky](#) expressed interest in a dump of our suggestions pipeline to feed into his [depictor tool](#) which has thus far been used to add 1M+ depicts statements to images on Wikimedia Commons. This tool makes suggestions for what images depict and lets users judge whether the suggestions are a good match. The tool then adds the depict statement to the Structured Data of the image if it is deemed a good match. The first iteration of the depictor tool uses Commons categories to make the suggestions. Our image suggestions data pipelines can also make these types of suggestions, from additional sources, which could help to expand the depictor tool. To help meet this request and others like it, we have made snapshots of the [section topics](#), [article](#) and [section-level](#) image suggestions data available publicly.

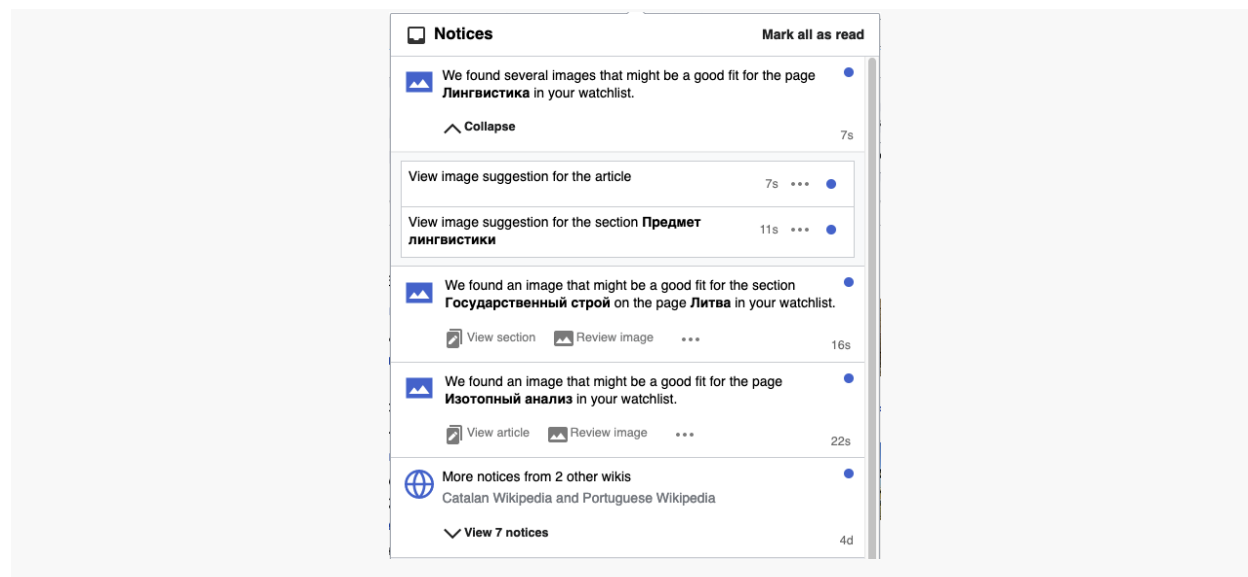
Image suggestions notifications for experienced users

As described in the Year 2 report, the second fork of our image-addition strategy helps experienced editors contribute more effectively. To this end, we’ve developed a tool that notifies experienced editors of image suggestions for articles on their watchlists, and covers a

broader array of images. This tool uses the algorithms and data pipelines described above to generate suggestions for both articles and sections of articles. For image suggestions notifications, we target users who have edited or watched a particular article or set of articles, since they are likely to be experts in the topic and to have interest in seeing that article(s) improve. By embedding the suggestions in the user’s existing Wikipedia activities through weekly notifications, we increase the likelihood they will review such suggestions and add selected images as part of their current editing workflows.

Notifications are currently sent weekly to all users on Wikipedias with the feature who have at least 500 edits. These notifications include a link to user’s preferences to allow users to opt out of the notifications. Suggestions are selected randomly from the list of matched recent images to unillustrated articles or sections, using the algorithms explained above. Notifications for a particular article-image match notification are only be shown once to a particular user, but the same match can be sent to multiple users to review, unless or until the image is inserted into an article. Based on the information provided in the notification, the user can click on “Review image” in the notification, which will redirect the user to the image on Commons; click on “Review article” in the notification, which will redirect the user to the article on Wikipedia; and go through their normal image addition workflow (e.g. choose to insert the image with wikitext or Visual Editor insert flow).

Article and section level image suggestions notifications in Russian Wikipedia



In July 2023, 66,286 notifications were sent to experienced contributors for articles on their watch list across Catalan, Finnish, Hungarian, Norwegian, Portuguese, Indonesian, and Russian Wikipedias. An average of 19.93% of those were acted upon, with a below average revert rate under 1%, and a very low opt out rate of 2.09%. This includes article level and section level suggestions.

Feedback on the tool has been very positive, such as this comment from User:AlbertRA:

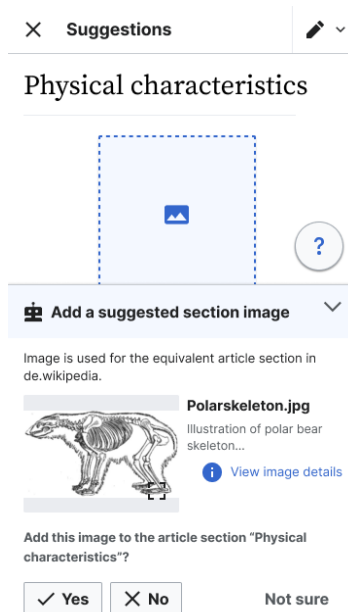
“Thanks for the work...it's a useful tool that I often use to illustrate infoboxes.”

Add an Image structured task for newcomers

In the Year 2 report, we detailed the add an image structured task for newcomers that allows them to easily add suggested images to Wikipedia articles based on their topics of interest. In Year 3, this feature was extended to 14 wikis. Since the feature’s release, over 55,800 images have been added to articles using the tool.

In addition, we created a new section-level image suggestions newcomer task, which is currently available in Spanish, Arabic, Bengali, and Czech Wikipedias. This task is considered a more difficult task, so we don’t suggest it for newcomers until they “level up” and are successful with article-level “add an image” tasks. Since the tool’s release in June 2023, over 480 images have been added to sections of articles using the tool. As more newcomers “level up” to this task, we expect to see usage continue to increase.

An example of the section-level image suggestions newcomer task



To ensure that this task would not be too difficult for relative newcomers, initial user testing of designs for the section-level task was completed in April 2023. Six testers were given instructions, asked to experiment with a prototype, and evaluate the easiness and enjoyableness of the task. Testers ranged in age from 18 to 55, were from five different countries, and most had not previously edited Wikipedia. Three of the testers were male, and three were female.

Some key take-aways from the user testing:

- The tool was understood by all participants: “*Clear, easy to understand, straightforward.*”
- Participants seemed to understand the task and that they needed to focus on the section when making their decision.

We also asked Ambassadors to review these tasks to better understand revert rates and issues newcomers might be facing. Results were positive: of the 60 edits reviewed, all of the images added were technically correct.

Image suggestions to support events and campaigns

In addition to the notifications sent to experienced users, we also created a custom category-based notification process that allows users or organizations running campaigns or events to produce notifications that specifically target topics they are trying to impact. We then used this feature to hold several focused events. In collaboration with the [Museo de los Museos](#), Wikimedia Chile, Wikimedia Portugal and Wiki Editorias LatinX, we held events teaching users about the new image suggestions features – both the experienced user notifications and the newcomer tool. These events focused on filling knowledge gaps, such as illustrating articles about women and LatinX people.

Social media advertisement for an event with Wikimedia Chile around image suggestions



During the event advertised above, 18 users added 60 images, and then added an additional 80 images in the days after the event. In another event focusing on women represented on Portuguese Wikipedia, 62 new articles were created, 466 articles were edited, and 247 images were uploaded to Wikimedia Commons. These events also boosted use of the image suggestions tools, with a 105% increase in image additions on Spanish Wikipedia after the events, and a 560% increase in image additions on Portuguese Wikipedia.

Feedback on the events was very positive, including comments such as:

“Participants were so happy about the fact that they received a lot of notifications...because now they would have a lot to do. My analysis is that one of the main difficulties is finding an article and image to work with and, to have that "part" already shared with you, it just makes things easier. Also, participants were so glad to receive so many women to edit. That was shared a lot. It felt like we were really making a difference in that regard, helping the gender gap.”

“My personal feedback using the tool is that it felt like a game, which was pretty fun. It also made the editing process easier for me and I wish I had even more suggestions because, like that, I would definitely edit much more. Finding articles and images takes a LONG time, if you find one at all to edit, and to have that part eliminated just made me feel like editing, editing, editing non-stop.”

In an anonymous follow up survey, 83.3% of participants agreed that the image suggestions tool suggested good images for their edits. When participants were asked “what did you think of the image suggestion tool” in the same survey, we got the following responses:

“Good, it allows not only to insert images, but also other elements in the articles and to complete the entries in the commons.”

“For me, 34 image suggestions appeared...several I added to Wikipedia.”

“I liked it, it was useful in all cases.”

“I use the notifications that the tool gave as a starting point and take the opportunity to show how images are inserted in articles, infoboxes (manuals and wikidata) and how entries in commons are completed.”

What's next for Image Suggestions?

The add an image structured task, both for article and section-level image suggestions, is currently in development to be released on our Android and iOS mobile Wikipedia apps, with a goal release date of late 2023. We plan to start rolling out on the mobile apps for Spanish, Portuguese, Persian, and Hindi Wikipedias. We will also be expanding both the structured task and the notifications to more language wikis later this year, with a goal of releasing on all Wikipedias by the end of 2023. In addition, the teams are working to continue to streamline and enhance the data pipelines that power this tool, and take the learnings from this work to apply it to other machine learning tools across Wikipedia. Lastly, we aim to publish the full image suggestions and section topics data sets on a regular cadence so that volunteer developers can use the data for new experimentation and tools.

Redesign of the search experience for new users on projects beyond Commons

Deliverable: Redesign and improve the search experience using structured data.

The need: This project will use structured content to give users a more inviting and efficient way to search and find content on the Wikipedias.

The impact: Search is demonstrably better, including updated user interfaces that are easier to use for new visitors. Users are able to find the information they are looking for, or that they may not have noticed, or previously came across through existing search – especially in those language wikis that have fewer articles.

In the Year 2 report, we detailed the impact that [MediaSearch](#) has had on Wikimedia Commons since it became the default search experience in May 2021. We explained that searches had increased approximately 50%, from approximately 95,000 per day before the release of MediaSearch, to approximately 140,000 per day in March 2022. We noted that Search teams get excited when they increase search sessions by 1%, so this indicates a very significant increase in search engagement on Commons and is evidence of the success of the new search experience. Since the last report, searches have increased another 10%. Additionally, in May 2022, [Portuguese Wikinews](#) became the first non-Commons wiki to make MediaSearch their default interface.

MediaSearch as the default search interface on Portuguese Wikinews

Pesquisar

[Mudar para Especial:Pesquisa](#) | [Ajuda](#)

Ferramentas ▾

Q birds

Busca

[Categorias e páginas](#) [Imagens](#) [Vídeo](#) [Áudio](#) [Outros resultados multimídia](#)

Tipo ▾ Tamanho da imagem ▾ Ordenar por: Relevância ▾

619 749 resultados



In the Year 2 report we also detailed our plans to improve the search experience on the Wikipedias, sharing our initial designs. Specifically, we said that we would use structured data to give users a more inviting and efficient way to search and find content on the Wikipedias. By improving the Search experience, we want to enable users to find the information they are looking for, or that they may not have noticed, or previously come across through existing search – especially in those language wikis that have fewer articles. Previously existing search functionalities on the Wikipedias mainly focus on article matches. This is a problem for Wikipedias in emerging languages, because there is a higher risk for readers to end up on the Search page without finding relevant articles or content. Linguistic factors and inexperience in using search are also among the main obstacles in finding content. Moreover, it is likely that the information that the user is seeking might be “hidden” somewhere on the projects – i.e., under sections of existing articles, or inside articles with a different name, or in sister projects and other sources in our ecosystem. We decided to solve this by building features that:

- enable readers to easily find what they are looking for when the exact article match is not found;
- surface relevant information from articles for better discoverability on the search page;
- help casual readers in emerging language Wikipedias assess the relevance of results;
- increase awareness of relevant information on other wiki projects such as Commons and Wikiquote; and
- Use the Section Topics pipeline described above to show relevant section information about articles.

Search improvements on all Wikipedias

Early in the year, we released some basic user interface changes to the search interface on all Wikipedias that make it easier for users to find what they're looking for. Because these represented a user interface change and not an entirely new feature, communities were receptive and we were able to release quickly across all Wikipedias. Differences include adding thumbnails to allow better scanning of the content, using the same thumbnails that appear in the Go bar, rearranging the sister search section and adding a header for better context and awareness of these projects, and restyling metadata to put more focus on the content itself.

Old Search Interface, before the changes

The screenshot shows the Wikipedia search interface from before recent updates. At the top, there is a search bar with the text "Search Wikipedia" and a "Create account" link. Below the search bar, the search results for "Ross Ice Shelf Project" are displayed. The results list several articles, each with a brief description and metadata. On the right side, there is a section for "Results from sister projects" which includes links to "Temp Image Testing", "The Twilight Zone (1959 TV series)", and "Introduction to Software Engineering/Authors".

WIKIPEDIA The Free Encyclopedia

Search Wikipedia

Create account ...

Special page

Search results Help

Q Ross Ice Shelf Project Search Results 1 – 20 of 408

Advanced search: (Sort by relevance X)

Search in: (Article X)

The page "*Ross Ice Shelf Project*" does not exist. You can *ask for it to be created*, but consider checking the search results below to see whether the topic is already covered.

Ross Ice Shelf
175°00′W﻿ / ﻿81.500°S 175.000°W﻿ / -81.500; -175.000 The **Ross Ice Shelf** is the largest **ice shelf** of Antarctica (as of 2013[update] an area of roughly 500 20 KB (2,634 words) - 05:39, 14 February 2022

Crary Ice Rise
southernmost ice rise. The feature was investigated by the USARP **Ross Ice Shelf Project** in the 1970s. The name came into use among USARP workers and honors 686 bytes (83 words) - 18:42, 29 January 2022

Rand Peak
Engineering Laboratory (CRREL), who drilled **ice** core at site J-11 (82°22′S, 168°40′W) during the **Ross Ice Shelf Project**, austral summers 1974-75 and 1976-77 584 bytes (124 words) - 23:40, 21 July 2020

Eilers Peak
States Antarctic Research Program glaciological party during the **Ross Ice Shelf Project**, 1974–75 field season. "Eilers Peak". Geographic Names Information 604 bytes (147 words) - 23:13, 21 November 2015

Marie Byrd Land (section **Glaciers, ice streams, and ice shelves**)
century. The territory lies in West Antarctica, east of the **Ross Ice Shelf** and the **Ross Sea** and south of the Pacific Ocean portion of the Southern Ocean 36 KB (4,501 words) - 03:48, 5 February 2022

Ross Sea
Peninsula in Marie Byrd Land, while the southernmost part is covered by the **Ross Ice Shelf**, and is about 200 miles (320 km) from the South Pole. Its boundaries 41 KB (4,749 words) - 21:25, 10 February 2022

Results from sister projects

Temp Image Testing
largest icebergs recorded have been calved, or broken off, from the **Ross Ice Shelf** of Antarctica. Iceberg B-15, photographed by satellite in 2000, measured Texts from Wikisource

The Twilight Zone (1959 TV series)
Hayley: Except you **Ross**: Except me. Lucky, I guess, huh? Hayley: Very lucky but... **Ross**: But what? Hayley: You're not even wet. **Ross**: Wet? What is wet? Quotes from Wikiquote

Introduction to Software Engineering/Authors
ClueBot; ClueBot NG; Conan; Cybercobra; Dalvizu; Daniel.Cardenas; Derek **Ross**; Dmcq; DRogers; DSParillo; Dwchin; Ed Poor; Edward; Edward Z. Yang; Erkan Textbooks from Wikibooks

New Search Interface, after the changes


Search results

Search: Ross Ice Shelf Project Search Results 1 – 20 of 430

Advanced search: Sort by relevance

Search in: (Article)

The page "Ross Ice Shelf Project" does not exist. You can [create a draft and submit it for review](#), but consider checking the search results below to see whether the topic is already covered.




Ross Ice Shelf

175°00′W / 81.500°S﻿ / ﻿175.000°W﻿ / -81.500; -175.000

The **Ross Ice Shelf** is the largest **ice shelf** of Antarctica (as of 2013[update]) an area of roughly 500...


20 KB (2,650 words) - 15:28, 19 August 2022



Cray Ice Rise

southernmost ice rise. The feature was investigated by the USARP **Ross Ice Shelf Project** in the 1970s. The name came into use among USARP workers and honors...


686 bytes (83 words) - 18:42, 29 January 2022



Ross Sea

Peninsula in Marie Byrd Land, while the southernmost part is covered by the **Ross Ice Shelf**, and is about 200 miles (320 km) from the South Pole. Its boundaries...

41 KB (4,754 words) - 14:40, 30 September 2022



Marie Byrd Land (section **Glaciers, ice streams, and ice shelves**)

century. The territory lies in West Antarctica, east of the **Ross Ice Shelf** and the **Ross Sea** and south of the Pacific Ocean portion of

Screenshot Southern Ocean...

36 KB (4,519 words) - 02:24, 3 October 2022

Texts from Wikisource

[The American Practical Navigator/Chapter 33](#)
and by seismic means at a number of locations along the edge of the **Ross Ice Shelf** near Little America Station. It was also substantiated by density measurements
[See all results](#)

Quotes from Wikiquote

[The Twilight Zone \(1959 TV series\)](#)
Hayley: Except you **Ross**: Except me. Lucky,I guess, huh? Hayley: Very lucky but... **Ross**: But what? Hayley: You're not even wet. **Ross**: Wet? What is wet?
[See all results](#)

Textbooks from Wikibooks

[Introduction to Software Engineering/Authors](#)
ClueBot; ClueBot NG; Conan; Cybercobra; Dalvizu; Daniel.Cardenas; Derek **Ross**; Dmcq; DRogers; DSParillo; Dwchin; Ed Poor; Edward; Edward Z. Yang; Erkan

Search Preview

Building upon the success of these user interface changes, and on the structured data infrastructure we'd created throughout the grant period, we created a feature called Search Preview that would allow us to accomplish our more ambitious project goals. Search Preview is a panel on the search page that shows additional context for each result, allowing users to discover relevant content and go directly to a desired section of the article. Search Preview appears after a user has executed a search, and allows them to preview any of the search results before clicking through to the content itself. Search Previews are intended to lower the risk of the user feeling like they have "failed" a search by clicking through to an irrelevant result, only to exit the site. Instead, we hope that being able to preview results will increase the likelihood they interact with relevant results, as well as increase the potential for discovery of adjacent relevant content that encourages further site interaction.

Search Previews contain the following elements: a header image depicting the article topic; an extended article content snippet to give context; table of contents links that take the user directly to sections of the article; related images from Wikimedia Commons; and links to related content on other Wikipedias. Search Preview takes advantage of the structured data

infrastructure we developed in previous phases of the work to display these connections between content. For example, showing table of contents links takes advantage of the Section Identification infrastructure work we did in the Section Topics project, and related images on Commons uses Structured Data on Commons and the MediaSearch backend to find good image matches for the article.

To build Search Preview, we extended the front end technology we built using the [vue.js framework](#) used to create MediaSearch to the Wikipedias. By installing this new framework on the Wikipedias, we've made future improvements to the front-end interface on the Wikipedias significantly simpler and more modern.

Search Preview on Portuguese Wikipedia. The user can see more information about the topic they're interested in.

The screenshot shows a search interface on Portuguese Wikipedia. At the top, the search results are titled "Resultados da pesquisa". Below this is a search bar containing the text "Aves". To the right of the search bar is a grid of 12 small images of various birds. Below the search bar are two filter boxes: "Pesquisa avançada: Ordenar por relevância" and "Pesquisar em: (Principal)". Below these filters are radio buttons for search engines: "Wikipédia" (selected), "Wikiwix", "Google", and "Yahoo!". Below the search bar and filters, there is a text block: "Há uma página com o nome 'Aves' na wiki Wikipédia. Veja também os outros resultados encontrados. Ver (20 anteriores | 20 posteriores) (20 | 50 | 100 | 250 | 500)". Below this text block are two search results. The first result is for "Aves", with a small grid of bird images to its left. The text for this result reads: "oferecem às **aves** a capacidade de voar, embora a especiação tenha produzido **aves** não voadoras, como as avestruzes, pinguins e diversas **aves** endêmicas insulares...". Below this text is the file size and date: "181 kB (19 777 palavras) - 22h11min de 14 de maio de 2023". The second result is for "Passeriformes (redirecionamento de Aves passeriformes)", with a small grid of bird images to its left. The text for this result reads: "ordem da classe **aves**, que compreendem a mais numer...". To the right of the search results is a large text block titled "Aves" in blue. The text reads: "classe de vertebrados que possuem penas, bico sem dentes e oviparidade de casca rígida". Below this text is a paragraph: "...as moas e as **aves**-elefante, ambos extintos. As asas, que evoluíram a partir dos membros anteriores, oferecem às **aves** a capacidade de voar, embora a especiação tenha produzido **aves** não voadoras, como as avestruzes, pinguins e".

After scrolling down in the Search Preview interface, the user sees a list of article sections that they can navigate directly to. This takes advantage of the Section Identification infrastructure work we did in the Section Topics project. Users can also see related images on Commons, which uses Structured Data on Commons and the MediaSearch backend to find good image matches for the article. Lastly, the user can see information about the same topic in other Wikipedias, such as Wikispecies.



To prepare for the Search Preview work, we completed a series of usability tests, with 16 participants ranging from 28-61 years old, half male and half female, representing France, Indonesia, Mexico, Singapore, Italy, Brazil, Switzerland, Portugal, Venezuela, the United Kingdom, and the United States. The participants reflected a range of Wikipedia experience, with some readers and some editors represented. We learned through the tests that all participants understood the purpose of Search Preview, as well as how to open and use it. Feedback from the participants included the following:

“Super easy, especially with the help of a tooltip.”

“Overall pretty easy. Better than expected. The whole experience seems seamless.”

“If this could be the next Wikipedia search engine, it would be very good.”

“Very easy [and] self explanatory.”

“It’s a great way to showcase different types of things.”

“I think it’s useful because it lets you search in a more efficient way so you don’t have to go back and forth to see if that’s exactly what you are looking for.. It is a nice change.”

“Once you see there is additional information... it was pretty easy and intuitive.”

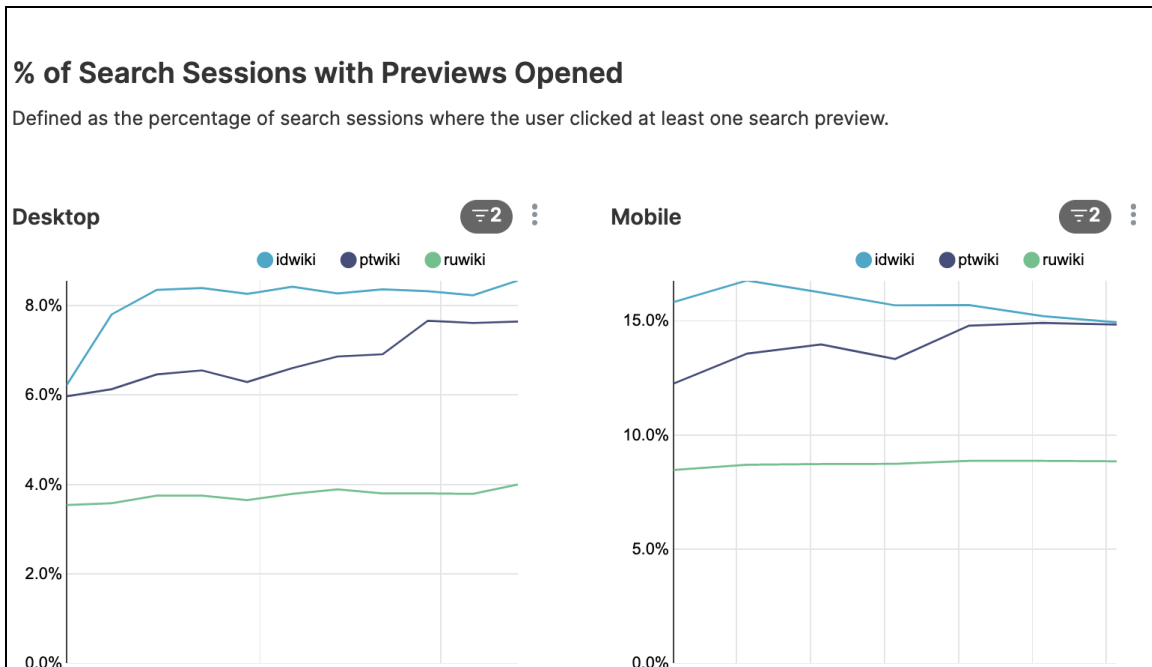
We also got the following feedback from Wikipedia users:

From User:Veracious (Indonesian Wikipedia): The features are really cool!

From User:Utilisateur (Indonesian Wikipedia): Makes it easier for us to check articles without having to move browser tabs.

From User:EpicPupper (English Wikipedia): These improvements look fantastic and greatly needed...thanks for all of your work!

The Search Preview interface is now live on Portuguese, Indonesian, Russian, Norwegian, Hungarian, Catalan, Dutch, and Ukrainian Wikipedias, with plans to extend it to more Wikipedias later this year. The data shows that very few users opt out of Search Preview: across all wikis, the opt-out rates are all below 0.01%. We also learned that these changes are especially useful for users on mobile, where the usage of the feature is more than double that on desktop interfaces. This is especially impactful, as [previous search user research](#) has shown that the majority of Wikipedia searchers are on mobile devices.



Design and launch 2–5 community-based pilot projects focused on experimenting with and encouraging adoption of new features.

Deliverable: Experiment with and encourage the use of new features.

The need: Getting the community involved is critical to the success of any new feature in the Wikimedia environment, and will help increase the sustainability of the infrastructure.

The impact: Increase adoption of reading, editing, or reuse features in emerging or diverse user segments, and increase community engagement with structured data.

Throughout the course of the grant period, we made sure to communicate regularly with our communities, including the 45+ community members who are subscribed to our [Structured Data Across Wikimedia newsletter](#), of which we published 12 newsletters. This created a community of interested participants that we were able to reach out to for further experimentation with our new infrastructure.

In May 2022, we sent out a Request for Proposal looking for interested community members to ideate and develop a tool or gadget that experiments with using structured data to improve the search experience on one or more Wikipedias. The RFP had the following requirements:

- The project must be search or discovery related.

- The project must use structured metadata tags. Metadata tags are statements that describe detailed characteristics of content and consist of a property and a value, such as Wikidata or Structured Data on Commons.
- The project must involve at least one Wikipedia. It can use data from other projects like Wikidata, but must help users search or discover Wikipedia content.
- The project should work on mobile devices.

After reviewing several proposals, we decided to move forward with [View It!](#), a tool that shows Wikipedia users relevant Wikimedia Commons media depicting—or otherwise related to—the article they are reading. View It! was developed by the following team:

- Project Manager: User: Dominic (Dominic Byrd-McDevitt)
- Lead Developer: User: SuperHamster (Kevin Payravi)
- Community Outreach Manager: User: JamieF (Jamie Flood)

View It! was such a promising proposal because it met multiple Structured Data Across Wikimedia goals – improving search and discovery, increasing illustration of articles, and increasing usage in underserved communities – all using the infrastructure built for structured data during this grant period and the previous one.

View It! enriches Wikipedia content by offering an illustration of a given subject. It increases the discovery of Wikimedia Commons uploads and encourages contributors to utilize Commons and structured data. While the number of images displayed in a Wikipedia article is finite and highly curated by editors, View It! allows readers to access the full catalog of images available on Wikimedia Commons, and helps editors easily add relevant items to a given article. In turn, the tool aids editing by surfacing new images that editors can use in an article. A copy-to-clipboard function allows users in edit mode to select an image from the View it! gallery to place in the article. In this way, View it! indirectly benefits even logged-out readers of Wikipedia who do not have it installed, since it can be used to assist the editors in finding media to improve the visual content of the pages they are reading.

View It! also allows users to update an article with better pictures even if an article or section is already illustrated, which our other image suggestions tools don't do, since they focus on unillustrated articles or sections. Importantly, View It! also allows users who aren't on Commons to improve the data quality of the Structured Data on Commons. View It! also offers an advanced search interface, which allows users to search by different property constraints, free text, specific resolutions, quality assessments, and Commons categories and sub-categories. The tool works on all language Wikipedias as an opt-in interface, and has multilingual support. It is also available on other Wikipedias, such as Wikispecies and Wikivoyage.

Once opting-in to View It!, users can access the interface by clicking on the View tab of an article.

The screenshot shows the Wikipedia article for Mary Cassatt. At the top, the user 'JamieF' is logged in. The navigation tabs include 'Article', 'Talk', and 'View', with 'View' circled in red. Below the tabs are options for 'Read', 'Edit', 'Edit source', 'View history', and a search bar. The article title 'Mary Cassatt' is displayed in green. A sub-header indicates it is a 'B-class article'. The main text describes her as an American painter and printmaker, born in Allegheny, Pennsylvania, and lived in France, befriending Edgar Degas. A 'Contents' table of contents is visible, listing sections like 'Early life', 'Impressionism', and 'References'. On the right, there is a gallery of images, with the first image showing Cassatt seated in a chair with an umbrella, captioned 'Cassatt seated in a chair with an umbrella, 1913. Verso reads "The only photograph for which she ever posed."'. Below the image is a structured data table with fields for 'Born', 'Died', 'Nationality', and 'Education'.

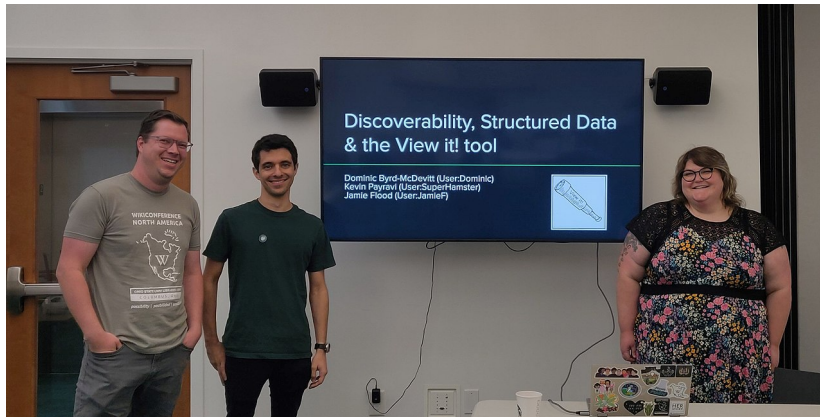
An example of the View It! Interface. Users can see relevant images for an article directly while viewing that article, and can easily copy those images to add to the article. The copy button is highlighted in red. When an image is added, if it has a caption in Structured Data on Commons in the correct language, it will be automatically added as the caption.

The screenshot shows the Wikipedia article for Sybil Thorndike. At the top, there's a search bar and user information for 'JamieF'. The article title is 'Sybil Thorndike' with a language selector for '16 languages'. Below the title, there are tabs for 'Article' and 'Talk', and a toolbar with various editing tools. A 'Good article' badge is visible. The main content area features a gallery of images, with several images circled in red. Below the gallery, there are sections for 'Good article', 'Short description', and 'Use dmy dates' / 'Use British English'. The article text begins with 'Dame Agnes Sybil Thorndike, Lady Casson, CH, DBE (24 October 1882 – 9 June 1976) was an English actress whose stage career lasted from 1904 to 1969.' It continues with details about her training as a concert pianist, her work with Ben Greet, Lewis Casson, and her roles in 'Saint Joan' and 'The Second World War'.

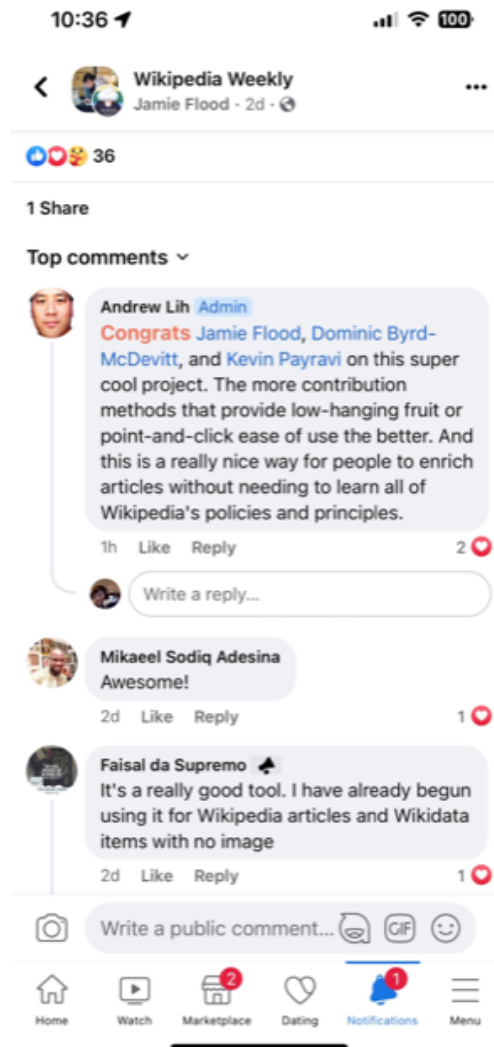
On the backend, View it! uses the [MediaWiki API](#) for Wikimedia Commons to search for images based on properties associated with the article. Every Wikimedia content page (such as a Wikipedia article) has a linked [Wikidata](#) entity, which is used to store information, such as the category of Commons media associated with that topic. Additionally, Wikimedia Commons files can use [Structured Data on Commons](#) to store information about entities [depicted](#) in a media file. Using these two data sources, the default configuration displays a combination of images that are either (1) found in an article's associated Commons category or (2) where the subject is tagged as depicted in an image. A custom View it! API powers the user scripts and Toolforge project, and can also be consumed directly by any app. All of the code is published [on Gitlab](#).

As part of the work on View It!, the project team conducted several outreach events, including talks at Wikimania and at the Wikimedians in Residence Exchange Network. The tool was beta tested by approximately 20 users, and had over 40 users sign up for regular updates during development. The project team shared blog posts about the tool on the [Wikipedia Signpost](#) and the [Diff blog](#).

The View It! Team presenting at the 2022 Wikimania meetup in Pittsburgh



The tool received overwhelmingly positive [feedback](#), including these messages from the Wikipedia Weekly Facebook group:



At the conclusion of the project, 192 users had installed View it! across global, Commons, and various language specific projects, and 94 users had used the tool to add P18 (image) statements to Wikidata. View it! has also been modified and added to other projects and utilized in other search queries, such as [ITN Syndication](#), [Template:Item documentation on Wikidata](#), and the Commons search in [Conzept](#).

DPLA Partnership

As part of our efforts to increase adoption of new features, we partnered with the Digital Public Library of America (DPLA) and their partners and communities to get more images described using structured data, which will allow us to have more images available in the image suggestions pipeline and generally drive the reuse of described and attributed images. This work is a follow on to DPLA the work funded by the Wikimedia Foundation in 2021-2022 to develop automatic updates of their contributing institutions' media files to Wikimedia Commons, with structured data.

As part of the preceding project, DPLA created over 27 million structured data statements for 2,334,599 items, including statements about copyright status, copyright license, RightsStatements.org statement, creators, subjects, identifiers, contributing institutions, description, title, and collection. DPLA also helped to specify a new “references” feature for structured data on Commons and redesigned the file info box for DPLA items in Wikimedia Commons to draw from Structured Data Statements rather than duplicative wikitext. Finally, DPLA successfully prototyped image citations for Wikipedia that draw from structured data statements and shared the implementation with the community. This project included coordination with volunteers that helped implement structured data querying from Commons into Commons' interface, so the prototype could be shown in a sandbox in the Commons interface.

For this next phase of our partnership with DPLA, we recognized that this record-breaking contribution of media files to Commons will have the most impact when the images are put in new contexts on Wikipedia and accessed by a large global audience, which is in line with the overall goals of Structured Data Across Wikimedia.

We set forth the following goals in our partnership with DPLA:

1. DPLA will develop a path for one or more contributing institutions or hubs to share more of their descriptive metadata with DPLA and Wikimedia that could be reconciled with Wikidata entities. This will allow for more descriptive structured data, which means more items will appear in MediaSearch results on Commons. By the end of the

project, DPLA aimed to update at least 1 million DPLA-contributed images on Commons updated with subject, creator, or other reconciled entities.

In order to accomplish this goal, DPLA built infrastructure that allows them to integrate these updates with their regular data synchronization process. They maintain a public JSON file that can be edited to add or modify subject reconciliations. Additionally, the Wikidata URIs for these subject matches are added into the DPLA aggregation data itself. At latest count, over 3.5 million subject statements have been added.

When DPLA introduced these changes to the DPLA data model, it was the first time institutions could provide subject term URIs (using SKOS exactMatch) rather than simple text strings. DPLA now ingests all of its subjects from the US National Archives as URIs for the NARA authority file, which allowed them to add these as SDC statements for the terms that have already been matched to their items on Wikidata.

DPLA learned over the course of the project that aggregation works very well with Commons and Structured Data, and their resulting approach to data synchronization has allowed them to make iterative changes that can be implemented quickly and have high impact. For example, at the start of this project, there were no DPLA-added subject (P921) statements in DPLA uploads—and by the end, there were millions. DPLA accomplished this by utilizing their data synchronization script, previously developed, to make updates to past uploads. They implemented new logic in the script to add the statements when the subject is one that was already identified. As part of this project, DPLA also began [keeping a database](#) of subject mappings for Wikidata, which the synchronization script utilizes to add subjects to Wikimedia Commons files.

This approach means that a publicly editable mapping document is continually checked as data is regularly updated, and these new subjects were added rapidly as passes were made for the bot to update all other metadata for the media files as well. Changes or additions made to the mapping will be automatically reflected in future synchronizations, as well. This document is not only open on GitHub, but was primarily developed by Evan Robb, a librarian who does not work for DPLA, but works for one of its main providers. This also shows how aggregation gives everyone in the network a reason to spend time on efforts that benefit the collective, and not just their own institution — and that Wikimedia projects can foster that as well.

Overall, we are excited about how this work benefits the broader ecosystem of linked open cultural heritage data. For example, for 10 years, DPLA's aggregation collected institution names and subject terms without any entities. This data is often created on the member provider's end with the use of authority files and stored with URIs. But DPLA never previously invested in implementing entities in its data model, and aggregated simply by ingesting all of these types of terms as string values. As a result of this project, the linked data work invested in by member institutions is now reflected in the aggregated DPLA data, making searching and discovering the information and images within the combined libraries much easier for users.

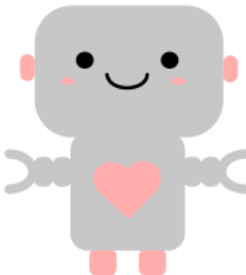
2. To address visibility for records that have irreconcilable descriptive metadata and cannot be easily updated by the above process, DPLA will develop a tool for suggesting depicts statements based on metadata subjects and evangelize that the tool be used by DPLA's community to improve search on Commons.

DPLA developed [DepictAssist](#), a tool that uses subjects to suggest potential depicts statements for images on Commons that were contributed by member institutions, built on top of the Suggested Tags feature developed as part of the Structured Data on Commons grant. The data added using this tool then becomes part of DPLA's aggregated metadata. The tool, in various iterations, was shared in many forums, including a meeting of the LD4 Wikidata Affinity Group, the DPLA Wikimedia Working Group, and an in-person meeting of 8 IUPUI University librarians. This feedback led to improvements and documenting other design needs which will be addressed in the future. DPLA now considers DepictAssist as part of its services, and are continuing to maintain and develop it as they anticipate pushing it out to broader audiences.

The DepictAssist interface, showing the drop down of available institutions

DepictAssist

by DPLA



Choose from the dropdown of all available institutions:

[Find images](#)

3. To leverage DPLA's detailed modeling of sources in structured data records, DPLA will work with the community to create a Wikipedia citation gadget. The gadget will follow the data modeling used by DPLA for their contributing institutions, but will be documented sufficiently to allow for customisation for use by other institutions should they desire to undertake similar work.

DPLA discussed and gathered feedback on the idea for a citation gadget at various events, including Wikimania, WikiConference North America, LD4, the DPLA Wikimedia Working Group and other network meetings, and elsewhere. Gathering community consensus on a citation template for use across the Wikipedias proved challenging, and DPLA is still discussing the final designs with the community. DPLA continues to work on automated ways to add the citation template from Structured Data on Commons to the Wikipedias, and plans to complete an initial batch of adding citations to images in the near future.

4. DPLA will conduct outreach to other regional and national aggregators such as the California Digital Library, Digital NZ, TROVE, and Europeana to share successes and learnings, encourage them to engage in similar programs, and advocate for more contributions to Commons globally.

Over the course of the project, DPLA conducted outreach in many places, including conferences such as Wikimania, the Wikimedia & Libraries conference, and the American Library Association, as well as with many meetings with key peer institutions, such as the US National Archives, the Smithsonian Institution, and the Biodiversity Heritage Library. They also formed the DPLA Wikimedia Working Group, which brings together key players from the DPLA network to help direct efforts. The working group seeks to support and further the work started as part of this project, improving the capacity and sustainability of the project through such initiatives as improving documentation, supporting project participants, and driving new cross-network collaborations. The working group aims to shape and help further the collaborative work of the DPLA Member Network, to share expertise with and learn from hub and DPLA colleagues, and to help address common needs and challenges. The group includes 10 members from across DPLA's member network and the library profession. It is DPLA's largest working group, representing a wide array of interests and experience.

The inaugural membership includes:

- Dominic Byrd-McDevitt (Chair), Data Fellow, DPLA
- Meredith Doviak, Community Manager, National Archives Catalog, National Archives and Records Administration
- Eben English, Digital Repository Services Manager, Boston Public Library & Technical Lead, Digital Commonwealth
- Christine Fernsebner Eslao, Metadata Technologies Program Manager, Harvard Library
- Jamie Flood, Senior Wikipedian and Outreach Specialist, National Agricultural Library
- Giovanna Fontenelle, Program Officer, Culture & Heritage, Wikimedia Foundation
- Rachel Meibos Helps, Wikipedian-in-Residence, Harold B. Lee Library, Brigham Young University (Mountain West Digital Library)
- Evan Robb, Digital Repository Librarian, Washington State Library (Northwest Digital Heritage)

- Angela Stanley, Assistant State Librarian for Innovation & Collaboration, Georgia Public Library Service (Digital Library of Georgia)
- Greta Suiter, Manuscripts Archivist, Ohio University Libraries (Ohio Digital Network)

Lastly, full documentation was created for all of the services created as part of this project. DPLA's digital asset pipelines are fully documented on the DPLA GitHub account, including the [DPLA ingestion repo](#), documenting how Wikimedia markup is generated from item records, and [ingest-wikimedia](#), which documents the upload and metadata synchronization.

After the grant period, DPLA plans to continue and deepen their relationship with the Wikimedia community. As a direct result of the work of this project and previous Wikimedia Foundation grants, DPLA was able to [secure further funding from the Sloan Foundation](#) for more Wikimedia programs over the next three years. They see that work as a continuation of the work started as part of the Structured Data Across Wikimedia funding, but with a broader scope than structured data. As such, they plan to maintain and continue to develop DepictAssist, continue to reconcile and add subjects to DPLA uploads, work on the SDC-powered citation concept, and continue to be active in global outreach with peer institutions.

Experiment with additional features to suggest relevant content to readers and editors during the article edit process

Deliverable: Newer and easier ways to read, edit, and access the knowledge within the Wikimedia projects.

The need: Emerging and underrepresented communities often prefer to interact with visual content, and Wikipedia content is largely represented in long-form text.

The impact: Bridge the knowledge gap in emerging markets and grow visual content for newer internet users.

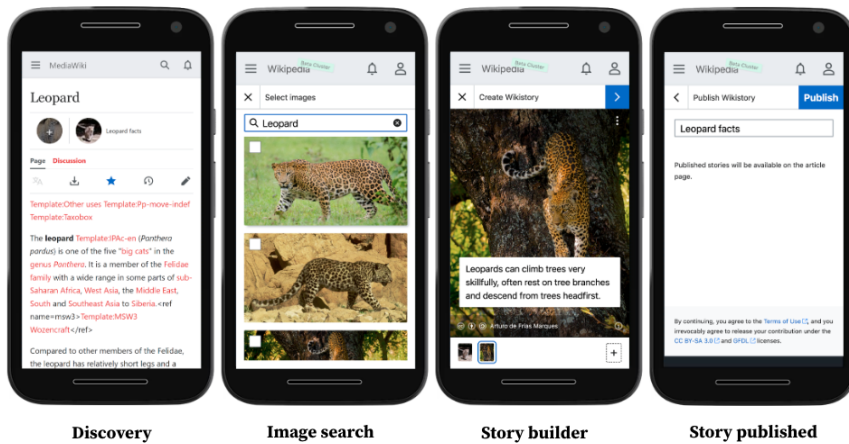
Lastly, in a related project, we launched Wikistories, a mobile web-based tool that empowers editors to create a new content format: short, visual and reliable knowledge for quick consumption and easy sharing. Wikistories is a story creation and consumption tool in our projects for editors and readers who want to engage with visual and reliable knowledge in a quick way using mobile devices. Wikistories is also a visual format that captures and distributes encyclopedic knowledge that is less suitable for long form articles. Wikistories takes encyclopedic content from Wikipedia, connects it with images from Commons, and allows users to create visually engaging content using that combination. To find and suggest images for stories, Wikistories uses the same Structured Data backend, developed as part of this grant project, that powers the MediaSearch Commons search interface.

We consider this work strategically important as we explore and experiment with new content formats, as it can help us to bridge the knowledge gap in emerging markets and grow visual content for newer internet users. The goal is in an increase in the diversity of content added, especially for smaller or emerging communities of contributors, whose languages and topics of socio-cultural interest have remained vastly underrepresented on the Wikimedia platform. We aim to grow visual content and lead to more contributor & reader engagement with encyclopedic content in emerging digital communities, and help underserved users to create, curate, contribute and engage with knowledge through visually-driven experiences.

To validate the Wikistories concept, we conducted qualitative user research in Kenya, Nigeria, South Africa and Indonesia. The research in the African countries focused on potential Wikistories contributors that are content creators and consumers in different spheres, not necessarily in the Wikimedia Movement. The collaboration in Indonesia focused on participants that are content creators in Indonesian Wikipedia, Commons and other Wikimedia projects.

Wikistories is now available in the Indonesian language Wikipedia, and was launched in collaboration with Wikimedia Indonesia, who published a [blog post in Indonesian](#) describing the effort. The Indonesian community was selected for their experience curating image-related projects, a high mobile contribution in an underserved market, and active GLAM (Galleries, Libraries, and Museums) willing to collaborate. After release, we collaborated with Wikimedia Indonesia to hold a workshop and a contest where the tool was introduced to newcomers and contributors in different Wikimedia Indonesian communities. The team also demonstrated Wikistories at WikiNuSantara 2022 — an event organized by Wikimedia Indonesia in Padang, and received encouraging feedback.

The Wikistories mobile interface, including structured data based image search using the MediaSearch backend



Wikistories on Beta Indonesian and English Wikipedias, showing image attribution from Commons metadata











Wikistories available for viewing on the Beta English Wikipedia “Cat” article

Wikipedia Beta Cluster 🔔

Cat

Small domesticated carnivorous mammal

  Edu's cats  Testing story with duplica...  Just cat  Long text story  Third cat story  Male cat 

Article Discussion

🌐 Language ☆ Watch 🕒 History ✎ Edit ⋮ More

This article is about the species commonly kept as a pet. For the cat family, see *Felidae*. For other uses, see *Cat (disambiguation)* and *Cats (disambiguation)*.


Script error: No such module "Autotaxobox".

The **cat** (*Felis catus*) is a **domestic species** of small **carnivorous mammal**.^{[3][1]} It is the only domesticated species in the family *Felidae* and is commonly referred to as the **domestic cat** or **house cat** to distinguish it from the wild members of the family.^[4] Cats are commonly kept as house pets but can also be **farm cats** or **feral cats**; the feral cat ranges freely and avoids human contact.^[5] Domestic cats are valued by humans for companionship and their ability to kill small **rodents**. About 60 **cat breeds** are recognized by various **cat registries**.^[6]

The cat is similar in **anatomy** to the other felid species: it has a strong flexible body, quick **reflexes**, sharp teeth, and **retractable claws** adapted to killing small prey like mice and rats. Its **night vision** and sense of smell are well developed. **Cat communication** includes **vocalizations** like **meowing**, **purring**, trilling, hissing, **growling**, and grunting as well as **cat-specific body language**. Although the cat is a **social species**, it is a solitary hunter. As a **predator**, it is **crepuscular**, i.e. most active at dawn and dusk. It can hear sounds too faint or too high in **frequency** for human ears, such as those made by **mice** and other small mammals.^[7] It also secretes and perceives **pheromones**.^[8] Female domestic cats can have kittens from spring to late autumn, with litter sizes often ranging from two to five kittens.^[9] Domestic Screenshot and shown at events as registered

Cat

Temporal range: 9,500 years ago – present



Various types of cats

Following the deployment, there were several [community](#) and GLAM engagement activities that acted as channels of feedback on user experiences including early adopters [survey](#) & [moderated research](#). In December 2022, the Wikimedia Foundation collaborated with Indian media company [The Paperclip](#), during which time The Paperclip created four stories relating to football and history, sourced entirely from knowledge available on Wikimedia projects. These stories were published on Paperclip’s [Twitter](#) ([1](#), [2](#), [3](#),[4](#)) and on their website during the 2022 FIFA Men’s Football World Cup tournament for relevance. The main objective of this collaboration was to see how the current long-form content of Wikipedia articles could be truncated into smaller bite-size, easy-to-consume formats that follow a storytelling approach and narrative. We observed the creation process of the Paperclip team, and the engagement of their readers, and noted some observations and takeaways.

From Srinwantu Dey, Co-Founder, The Paperclip:

We tried to curate stories of different flavors with a strong Indian connect and a unique perspective that delight or surprise our readers. The biggest challenge was how to transform a slice of encyclopedic fact to a well narrated story that will attract and retain readers...Our finale story (about Fred Pugsley) was of how a refugee from Rangoon lit up Calcutta football scene – it was a story of human condition and resistance and sparked a lot of relevant emotion among the readers including a few of them who shared personal anecdotes of their parents and grandparents who also suffered similar war experience and traveled to India from Rangoon on foot.

Interpretation of the content is part of the storytelling process and Wikipedia content policies provide guardrails to avoid the insertion of bias. Attributions and references are important acknowledgments of the authenticity of the content and upholds the integrity of the topic being narrated. At the same time, stories about difficult topics need careful handling, including the use of appropriate language for these to be inclusive to the wider audience. During the collaboration, the teams learned from each other about sensitive storytelling and balancing encyclopedic authenticity. The Paperclip team commented that they particularly learned a lot about citation methodologies for the creative commons license.

The story about Pugsley eventually also made it to the Indian mainstream news with an article in the Indian Express: [The greatest refugee story Indian football ever had: Fred Pugsley, who travelled 500 km on foot from Burma, wins hearts online](#). The stories had over a combined 1 million views, were liked by nearly 7,000 unique visitors, and were retweeted by more than 1,700 users. This was an encouraging outcome of the exercise that demonstrated potential opportunities for the presentation of encyclopedic content across media for a variety of audiences.

Since release on the beta cluster on Indonesian Wikipedia in July, 2022, 244 editors have created 946 Wikistories. Most of the stories are about culture, cities in Indonesia, monuments, animals, and personalities. In the next fiscal year, we plan to focus on organic adoption and retention of the tool by editors through various re-engagement activities with the Indonesian Wikipedia community. We will also focus on the definition and development of [Wikistories editorial moderation features](#) as proposed by the community.

Conclusion

As we reflect on the past three years, here is one prism through which to see the project: Statistics that indicate what has changed. By this prism, *much* has changed for the better:

- **Total media requests have boomed.** This measures the total amount of images, video, and audio files that users have clicked on, and when we compare February 2020 to July 2023, total media requests have increased by 165,741,708 per month. This is likely due in part to how much easier it is to search for and discover media on Commons after our work, and due to the increased number of images on Wikipedia articles – which is also due to our work.
- **Total number of Commons’ media searches has skyrocketed.** The total number of media searches on Commons has increased by 60% since the start of the project.
- **More than half a million images have been added to Wikipedia articles.** Since January 2022 (which is the earliest we have a measurement for), users have added more than 587,000 images to articles on the Wikipedias.

- **More than 50,000 images have been added by newcomers in emerging languages.** Over the course of the project, over 56,200 images have been added to articles on the Wikipedias by newcomers in emerging languages.

This last statistic is a key bellwether of the project's success since our Structured Data Across Wikimedia plan from three years ago highlighted the project's potential for users in emerging communities — especially mobile users. That's why we are so excited by Wikistories, the mobile web-based tool that we detailed in this report's previous section, and that uses the same Structured Data backend developed as part of this grant project. It's an example of the momentum that this project has built over three years that will carry our work forward in the years to come.

That work, we are sure, will help bridge the knowledge gap in emerging markets and growing visual content for newer internet users. And as we move forward, it will rely on an approach that we spotlighted throughout this report: Community involvement. Structured Data Across Wikimedia worked because of the active participation of scores of users from around the world. The last three years have confirmed how crucial it is to iterate our work with different language communities, exemplified by Search Preview. As mentioned previously, the Search Preview interface is now live on Portuguese, Indonesian, Russian, Norwegian, Hungarian, Catalan, Dutch, and Ukrainian Wikipedias, with plans to extend it to more Wikipedias later this year. Search Preview has been an unabashed hit with every language community we've brought it to. And like other aspects of Structured Data Across Wikimedia, it has established a momentum of change and improvement that will carry us into the foreseeable future.

That future is bright, we can say with conviction. Your support has helped elevate structured data where it belongs on the Wikimedia projects: As a means to make Wikipedia and our other projects more useful, more impactful, and more beneficial to everyone who seeks knowledge from them. Structured data can help every action that users take on our projects — from a search query to creating new content. We're now seeing those actions happening in real time. And we're witnessing the real-time benefits that are occurring. The story of Fred Pugsley, which has generated so much attention across India, started with [an 11-paragraph Wikipedia article about him](#). By collaborating with the Indian media company The Paperclip, the Wikimedia Foundation helped turn that article into something much bigger. Structured data is transforming knowledge acquisition not just across our projects *but beyond*. That's something we had hoped for when we started Structured Data Across Wikimedia. We're proud that it has become a reality. And we're thankful that we can celebrate that fact with the community and with the Sloan Foundation.