

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

Sloan Grant #G-2020-12665. June 1, 2022. Prepared by Amanda Bittaker, Carly Bogen, Jonathan Curiel, and Shari Wakiyama.

Executive Summary

In the second year of Structured Data Across Wikimedia, our vision is becoming more of a reality — with more features delivered to users and more of the foundational infrastructure and services that will carry the project forward in the long term. We have made incremental progress on a broad range of fronts and some breakthrough impact in several initiatives, summarized below. As expected, we’re also refining the project based on a clearer picture of what’s working and — from learnings with editors — what we now know is unrealistic to implement.

Among the developments since our last report:

- As part of our Year 2 plan to add structured metadata to long-form content on one Wikimedia project, we identified the best content to structure on the best project — section topics on Wikipedia — and began planning and implementing. The “Section Topics project” will convey a Wikipedia article’s content by identifying sections within that article and automatically creating topics for those sections. Understanding an article’s section topics is important for building tools that allow contributors to more easily add additional content to those articles. We also implemented essential infrastructure to enable these features that will make development faster and easier for the rest of the program period and years to come.
- Aligning with our Year 2 plan to redesign and improve the search experience on the Wikipedias using structured content, we defined and scoped what we’re calling the “Search Improvements project,” which among many advances will let users easily find what they’re looking for when they don’t get an exact article match, and will help casual readers in emerging language Wikipedias more easily assess the relevance of their search results. We also saw the continued success of MediaSearch on Commons based

on the work we did in Year 1. Searches on Commons have increased approximately 50%, from approximately 95,000 per day before the release of MediaSearch, to approximately 140,000 per day in March 2022.

- Reflecting our new insight into editors’ ability to add images to Wikipedia, we’ve changed the 3-year target of contributing images to 5 million Wikipedia pages. Based on our analysis of editor engagement and articles, we’ve identified key barriers to reaching this goal, and we believe it will be more impactful to add fewer, higher-quality images within the grant timeline. We think that this will allow even more long-term sustainable growth in image addition on articles.

We detail all these developments in the body of this report, which documents how we have again streamlined the project for the better. Even changing the 3-year goal around images allows us to increase access to knowledge while staying true to our contributors’ and readers’ needs, and will still lead to more new editors — including those from emerging markets — working with our projects.

At the end of Year 2, we are starting to see the impact of our first releases and the shape of the rest of our program plan, and starting to achieve our goal of using structured data to attract new contributors and give users everywhere greater access and ability to engage with the knowledge in our projects. That’s why we continue to be grateful to the Sloan Foundation for funding Structured Data Across Wikimedia, and for giving us a concentrated period of time to achieve its short-term aims and, by default, its long-term potential.

Table of Contents

Executive Summary	1-2
Year 2 So Far: The Project’s Progress	3-23
Snapshot of our work	3
Adding structured metadata to long-form content	4-11
Exploration of Search improvements	11-16
Finalizing features for improved suggestion/recommendation capabilities	16-24
Identifying strategies to experiment with and encourage the use of new features	25
Year 3: A Look Ahead	25-26
Other Challenges	26-27
Conclusion	27-28

Year 2 So Far: The Project's Progress

A Snapshot of Our Work from April 1, 2021 to March 31, 2022

What We Envisioned for Year 2	What We've Done So Far
<p>Add structured metadata to long-form content on at least one Wikimedia project; and add new moderation capabilities that allow community members to edit/monitor that metadata if necessary.</p>	<ul style="list-style-type: none"> • Reviewed literature and decided to focus on sections as the focus data structure • Received valuable feedback from the communities on their moderation needs for Section Topics • Completed initial research and design for creating section topics structured data • Added more structure to talk pages to enable more intuitive on-wiki communication for newcomers and experienced contributors alike
<p>Exploration of search improvements.</p>	<ul style="list-style-type: none"> • Designed and defined the Search Improvements project, which will make the search results page more useful and accessible, especially to users in emerging markets. • Scoped a collaboration project with community members to build experimental features that help users in emerging markets find the content they are looking for • Collaborated with Google to determine if the structured data we produce can help to improve the SEO and appearance of Wikipedia results when users are searching on Google
<p>Finalize features for improved related content suggestion/recommendation capabilities for editing.</p>	<ul style="list-style-type: none"> • Created a structured task for newcomers that allows them to easily add suggested images to Wikipedia articles based on their topics of interest • Developed infrastructure to create, store, and use generated datasets about our content, further enabling future natural-language processing work and connections across our content. • Changed the focus of image suggestions to pursue a targeted user campaign strategy, focusing heavily on our editors' and readers' content needs
<p>Identify strategies to experiment with and encourage the use of new features in Year 3 and beyond through a series of community-based pilot projects.</p>	<ul style="list-style-type: none"> • Based on lessons from first events, developed plans to build further newcomer events in Wikimedia Mexico and Wikimedia Argentina, and scoped work with Wikimedia Portugal to do events and a contest around image suggestions with experienced users and GLAMs • Drafted a partnership plan with the Digital Public Library of America to potentially expand a proof-of-concept citation tool that would allow users to use structured data to more effectively caption images on Wikipedia • Continued to partner with interested community members who would like to build tools and gadgets using the structured data infrastructure we've developed

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 2, we made significant progress on adding structured metadata to long-form Wikipedia content. 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. We are already seeing the impact of adding structure to talk pages, with Junior Contributors 1.3 times more likely to publish a new topic. We are excited to see the impact of section topics as it reaches users in Year 3.

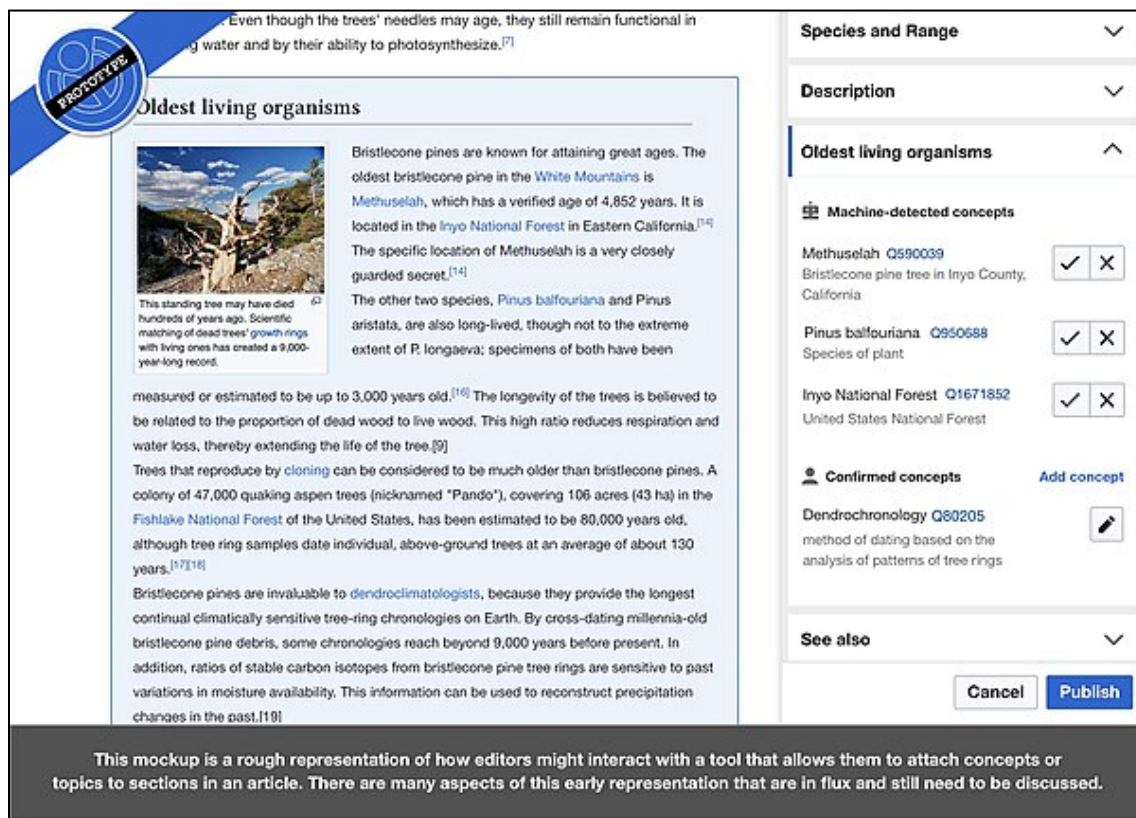
In year 1, we started planning a project to structure text content on the Wikipedias by testing different approaches to suggested topics for sections of Wikipedia articles. We prototyped the required architecture and began developing a plan for a production infrastructure. This system would structure articles into independent sections of wikitext, and automatically detect topics and concepts in those sections, adding those topics and concepts as a structured content element.

As a result of our work in year 2, we now have a product plan, alignment between our delivery teams, feedback from the community, and a more thorough understanding of what the technical requirements are for the rest of the grant period for adding structure to long-form content. What we're now calling the Section Topics project will create understanding of an article's content by identifying sections in a Wikipedia article and creating topics for those sections. This has several 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 will use the blue-links algorithm and section identification infrastructure above, and be delivered both via the newcomer experience and via notifications for experienced contributors. This will build upon the [prior image suggestions work](#), which we detailed in last year’s report;
- 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 the grant proposal, we also proposed considering new moderation capabilities that allow community members to edit/monitor that metadata if necessary. For this, we explored a UI for users to accept/reject/add/delete section topics. We created mocks and put them in front of users, as shown in the image below:



We received valuable feedback from users indicating that this type of moderation would be a high burden on them. For example, a [user told us](#) that moderation of this kind would undermine their ability to shepherd pending articles onto Wikipedia, prevent misinformation in articles, and do other tasks that otherwise improve the Wikimedia projects:

“There is the issue of the labor to do all this work. Where is that supposed to come from? We are already struggling to keep up with our existing work. (We have) new articles sitting unreviewed in draftspace waiting for possible promotion to mainspace article - (which) is currently **backlogged by over three months with 3,195 pending submissions**. New users

are waiting up to THREE MONTHS to find out whether their article will make it into the encyclopedia. (We also) currently have a backlog of nearly 7000 unreviewed pages in mainspace. . . . That barely scratches the surface of our work. If you make this new system - *if the community were to accept it* - it means we would spend less time creating new articles. It means we would spend less time expanding and improving existing articles. It means we would spend less time cleaning up bias and propaganda and misinformation and other crap that gets into articles. It means we would have less time available to assist new users.”

Based on this feedback and others like it, we decided to start with generating section topics automatically. All features built using section topics will still require human moderation. For example, section-level image suggestions will require a human in the loop to determine whether an image that is automatically suggested for a section is actually added to that section. Overall, it was by working with the community and taking away learnings that we decided to eliminate moderation capabilities for section topics. After the grant timeline, we may still include moderation capabilities if further learnings show that users want or need them.

Our research team reviewed a variety of literature in this area to determine that sections are a useful data structure to focus on. This research included an [article on when editors expand sections into full, separate articles](#); [research from our Design Research team in the context of Section Translation](#); and a paper on [structuring articles via the recommendation of section templates](#). This research and literature review led to an understanding that sections are a logical way to structure and separate the content of the article, and that doing so would support contributors and readers in a variety of future use cases, including:

Use case	When I...	I want to...	So I can...
Section-level image recommendation	...am working on a campaign or newcomer tasks	...see unillustrated sections about my campaign topic or topic of interest and recommendations for matching images	...add images to under-illustrated topics
Organizing a campaign around a topic	...am organizing a campaign	...see articles that don't have sections that discuss a particular topic, <i>e.g.</i> , articles about beaches that don't mention erosion	...add those articles to a list for the campaign to act upon

Improve search	...am searching for information about a concept on Wikipedia	...see sections of articles that are about that concept and related concepts in my search results	...more easily find what I'm looking for and explore related concepts
Structured tasks below article level	...am a newcomer working on structured tasks	...see sections that are about topics I'm interested in editing	...work on areas of my interest and be more successful in my first edits
Section translation & fact translation	...am interested in translating articles from one language to another	...see if sections I am interested in already exist in my target language	...translate sections of interest (in a less-process intensive, more stable way than is currently done)
Reading lists for offline reading	...am reading Wikipedia on my device without access to data	...read a curated list of Wikipedia articles and sections based on a previous search or selection	...access those articles without spending data
Snackable content/visual stories/voice snippets	...don't have time to read lengthy articles or am part of a culture that prefers listening and watching to reading	...see short-form visual or audible stories about the content I'm interested in	...consume content in a way that works for me
Better statistics	...am trying to understand gaps in our content	...see how many articles discuss certain topics, e.g., "biographies of women tend to have more content about marriage"	...make better decisions about development, campaigns focus, etc.

Better ways to catch subtle vandalism	...am patrolling Wikipedia	...easily see when edits result in new concepts being generated that are likely to be off topic to the section of the article they were made in	...flag that vandalism may be occurring and address it
Structuring edit summaries	...am patrolling Wikipedia	...filter my recent changes feed to the topics I'm an expert on	...narrow down the edits I want to patrol
Editor profile	...am an editor of Wikipedia	...have a profile page that shows the types of topics I've contributed to or what the edits I've made are about	...easily show people what type of work I do on Wikipedia

Of the above, we will be focusing on “section-level image recommendation” and “improve search” within the scope of this grant, while opening up opportunities for the other features in the future.

In the grant proposal, we proposed considering exploring the integration of concept metadata with anti-vandalism and quality control systems. We explored this at a high level and it may still be a use case for the future, but won't happen in the timeline for the grant.

In addition to Section Topics, we've also invested in a related initiative to structure talk pages. This investment had two goals: **to make it more intuitive for newcomers to communicate with others so they can grow into productive contributors; and to give experienced contributors more leverage to communicate and coordinate their wiki work with others.** Wikipedia depends on contributors collaborating, and communication is an important part of the collaborative process. By adding structure to talk pages, we aimed to evolve them in a way that gives experienced contributors more leverage to coordinate their work and connect with other editors, while making communicating on-wiki more accessible and intuitive for newer contributors.

Before adding structure to talk pages, contributors with more experience reported being [slowed down](#) by having to do something manual in order for other contributors "to know who posted what, and when, so they can follow the thread of a conversation, post on a user talk page if appropriate, and even just know whether a posting is recent enough to be worth responding to." Newer contributors on the other hand, [found basic tasks](#), like replying to a comment, confusing.

The structure we were able to add to talk pages enabled new features such as a [Reply tool](#) for replying to comments inline, which automatically signs and indents comments and offers a quick way for pinging other users; a [new topic tool](#) for starting new sections inline, which automatically adds users' signatures, and like the Reply Tool, offers a quick way for pinging other people; and [topic subscriptions](#), which allows users to receive a [notification](#) whenever someone posts a new comment in a discussion they are interested in.

[User testing](#) done on the Reply Tool feature in 2021 showed the tool leads a greater percentage of Junior Contributors to publish a comment without a significant increase in disruption. Following these tests, in March 2022, the Reply Tool became available on all Wikis. By early April 2022, the one millionth comment was posted using the tool.

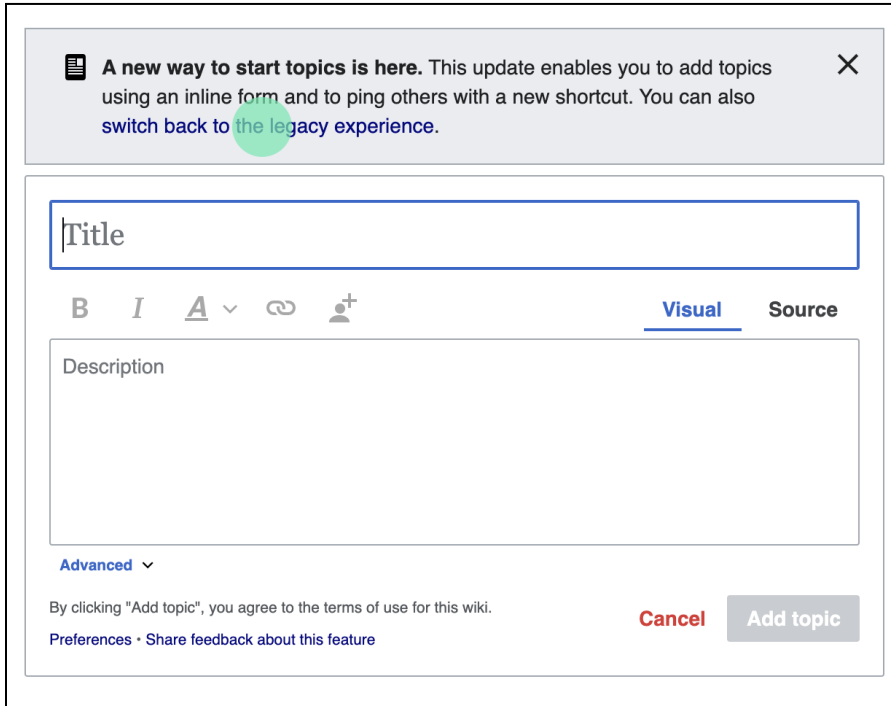
The topic subscriptions tool was released as a beta feature to all Wikis in January 2022. As of March 2022, 99.5% of people who received a new comment notification after manually subscribing to a discussion kept the feature enabled; people read 96% of the new comment notifications they received within two weeks of receiving them; and the average number of notifications sent per day has remained fairly stable with a daily average of about 4 notifications per user per day.

We also ran user tests of the New Topic Tool, ending in March 2022, which showed that the tool increases the likelihood that logged-in Junior Contributors will successfully publish a new topic. 44.2% of Junior Contributors that opened the New Topic Tool were able to successfully publish at least one new topic during the A/B test compared to 37.2% of Junior Contributors using the existing workflow. Junior Contributors using the New Topic Tool are 1.3 times more likely to publish a new topic they start than Junior Contributors using the existing workflow.

Users have universally praised the tool, as indicated by the four comments below:

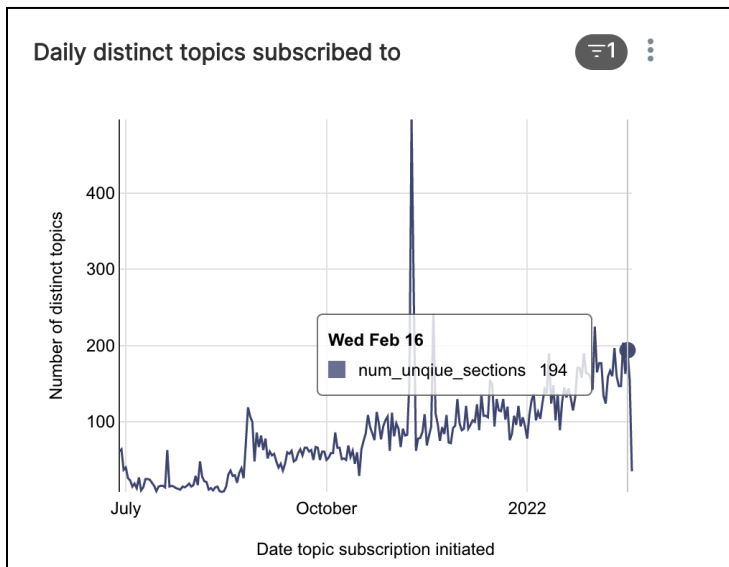
- From User:Czar: "Just wanted to add that this is the best beta feature rollout I've seen on the English Wikipedia in years. It's a very useful tool and improves every aspect of the commenting experience. Please pass my congrats to the dev team!" | [Source](#)
- From User:PMG (Polish Wikipedia): "In my opinion, this is a very good tool and should be enabled for all users. . . . I fully support the introduction of this tool, I believe that it should be permanently enabled for everyone, because even for Wikipedia veterans it makes life easier, and newcomers will not have to undergo a weekly course on 'How to respond to discussions on Wikipedia.'" | [Source](#)
- From User:Soreat (Dutch Wikipedia): "My experience has been extremely positive. I no longer have to manually place a signature, everything is automatic, great. It also makes it easier to work with the Visual Word Processor, because this word processor doesn't really work well in discussion sections." | [Source](#)
- From User:Épine: "I am loving this new tool (which) made discussions a lot easier to have." | [Source](#)

Below: An image that shows the new topic tool at work – and its ease of use.



A screenshot of a web interface for adding a new topic. At the top, a grey notification box contains the text: "A new way to start topics is here. This update enables you to add topics using an inline form and to ping others with a new shortcut. You can also switch back to the legacy experience." A green circle highlights the word "legacy" in the notification. Below the notification is a form with a "Title" input field, a rich text editor with "Description" placeholder text, and a toolbar with icons for bold (B), italic (I), text color (A), link, and user selection. The form has two tabs: "Visual" (selected) and "Source". At the bottom, there is a "Cancel" button and an "Add topic" button. A small disclaimer states: "By clicking 'Add topic', you agree to the terms of use for this wiki." There are also links for "Preferences" and "Share feedback about this feature".

Below: A graph that shows the steady increase in topic subscriptions.



The success of this work to structure talk pages is evidence of the impact we can have by adding structure to Wikipedia content. Editors and readers are empowered by this work, and we look

forward to continuing this impact by expanding structure beyond talk pages to Wikipedia articles themselves.

Exploration of search improvements

Deliverable: Redesign and improve the search experience on Commons and the Wikipedias using structured content.

The need: New users can search and refine searches in an intuitive and familiar interface

The impact: Search is demonstrably better, including updated user interfaces that are easier to use for new visitors

As described in the first-year grant report, in 2021 we released [MediaSearch](#), a new way to find media on Wikimedia Commons. MediaSearch uses categories, structured data, and wikitext from Commons, and Wikidata to find its results. MediaSearch became the default search experience on Commons in May 2021. Since then, searches on Commons have increased approximately 50%, from approximately 95,000 per day before the release of MediaSearch, to approximately 140,000 per day in March 2022. 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. We also received positive feedback from users, such as [this quote from User:DonaldTrung](#):

"I really like this design because [it is] familiar to most . . . people that use the internet. Overall I would say kudos to the team for getting the design right immediately. It looks like . . . a welcome upgrade of the standard search. Keep up the good work everyone."

MediaSearch also allows users to filter their searches by options like “license,” a filter that was used over 40,000 times per month in 2022. After users filter their search by licenses that are acceptable for use on their language Wikipedia, they can then use the newly implemented “copy” feature to directly access code that allows them to insert their desired media into Wikipedia articles, facilitating quality media additions on the Wikipedias.

MediaSearch was built using two reusable components, a front-end extension, and a back-end extension. Both components use modern technologies and are reusable and extendible across all wikis. For example, the front-end extension uses the [vue.js framework](#) that provides many immediate improvements to the front-end development process at the Foundation, and will help to pave the way for further architectural improvements in the future. We are currently exploring the possibility of using both the front- and back-end components developed for MediaSearch to improve the search experience on the Wikipedias.

In Year 2, following the success of MediaSearch, we focused our attention on improving the search experience on the Wikipedias. We designed and defined the Search Improvements project, which will make the search results page more useful and accessible, especially to users in emerging markets. We also scoped a collaboration project with community members to build experimental features that help users in emerging markets find the content they are looking for. Lastly, we collaborated with Google to determine if the structured data we produce can help to improve the SEO and appearance of Wikipedia results when users are searching on Google.

In the grant proposal, we said we would explore a variety of ways to use structured data to improve the search experience, including:

- tags for enhancing search quality and navigation via faceting and filtering
- deployment of machine-learning models to learn relevance and automatically improve results based on user feedback loops
- usage of query intent classification and other query-parsing approaches like semantic analysis and natural language understanding to define which models and query methods are most effective in delivering user results
- a series of experiments focused on proven ways of improving search result relevance for users

After exploring these options through user research, community consultation, and evaluations of technical feasibility, we are focusing on the last as the most promising avenue and the most needed by our users. To determine this, we conducted research in the Arabic and Spanish language communities on the Wikipedia Search experience, which resulted in findings such as:

- Emerging language Wikipedias have less content than larger Wikipedias (such as English Wikipedia). As such, there is a higher chance for readers to end up on the Special:Search page without relevant content, or it is hard and inefficient to examine results to find what they are looking for. It is also possible that linguistic factors in emerging languages make it more difficult to hit an article exactly from the Go bar since, as one example, languages like Swahili with robust morphology — especially using prefixes — might reduce the effectiveness of autosuggest.
- Most testers reported that their usual route to a sought Wikipedia article begins in Google, and typically via the URL/search bar. Absence of a Wikipedia article at the top of the Google results is interpreted as an absence of coverage on Wikipedia. Searches often end at this point.
- Images are important to testers.
- Testers are not familiar with sister projects such as Commons or Wikibooks. An updated display of results from sister projects sees an improvement in tester reactions and a greater general tester uptake of presented information.

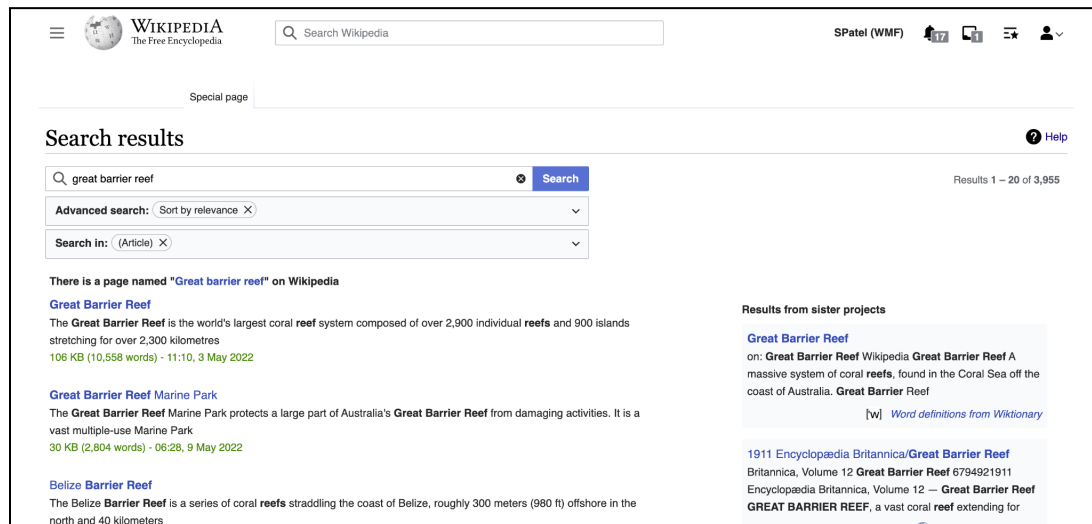
- Testers respond positively to surfacing sections in search results.
- Testers generally react positively to better-contextualized text snippets.

The Search Improvements project is using structured content to give users a more inviting, more efficient way to search and find content on the Wikipedias. In Year 2 we identified and defined incremental improvements to the Wikipedia search results page that use structured content to help users in emerging markets find the content they are looking for. Specifically, we aim to use structured content to:

- 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 Special: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;
- use the section topics described above to show relevant section information about articles.

Below is an initial design exploration for the Special:Search improvements, as compared to the existing interface. You can see a clearer, more consistent user interface, and related information surfaced in the right panel to enhance discoverability.

Existing interface of Special:Search:



New concept for Special:Search:

The screenshot shows the Wikipedia search interface. At the top, there's the Wikipedia logo and a search bar containing 'Cambio climático en Nicaragua'. Below the search bar, the results are displayed under the heading 'Resultados de la búsqueda'. A message states: 'La página "Cambio climático en Nicaragua" no existe. Puede solicitar que se cree, pero considere verificar los resultados de búsqueda a continuación para ver si el tema ya está cubierto.' Below this, there are several search results, each with a thumbnail image and a brief description. The first result is 'Cambio climático (sección en Caribe (región))', which is highlighted with a blue border. Other results include 'Clima (sección en Geografía de Nicaragua)', 'Cambio climático en América Latina', 'Cambio climático en El Salvador', 'Nicaragua', and 'Relaciones Cuba-Nicaragua'. On the right side, there is a 'Vista rápida' (Quick view) panel for 'Caribe (región)', which provides a summary of the region and includes a section for 'Caribe (región) Imágenes' with several small images. At the bottom of the quick view panel, there is a 'Resultados de Wikiviajes' section for 'Centroamérica y el Caribe'.

We are also working to partner with interested community members who would like to build tools and gadgets using the structured data infrastructure we've developed to improve the search experience on one or more Wikipedias. We have released a request for proposals to active community members for a project that is search or discovery related, uses structured metadata tags, and helps users search or discover Wikipedia content.

In addition, we are following up on an emergent opportunity to explore improving search results on Google using our structured data. We are working with Google to see if the structured data we produce can help to improve the SEO and appearance of Wikipedia results when users are searching on Google. The goal is make it easier for users to find the most relevant section of a page for their search. Google has been experimenting with surfacing subsection links and snippets for structured pages.

Examples of proposed tabbed snippet design:

https://en.wikipedia.org/wiki/Quantization_of_the_e...

Quantization of the electromagnetic field - Wikipedia

The **quantization** of the **electromagnetic field**, means that an **electromagnetic field** consists of discrete energy parcels, photons.

Second Quantization Electromagnetic Field And V... Quantization Of EM Fie >

Second quantization starts with an expansion of a scalar or vector field (or wave functions) in a basis consisting of a complete set of functions. These expansion functions depend on the...

[Continue on en.wikipedia.org »](#)

https://en.wikipedia.org/wiki/Stephen_Colbert

Stephen Colbert - Wikipedia

Stephen Tyrone Colbert (/koolˈbɛər/ kohl-BAYR; born May 13, 1964) is an American comedian, writer, producer, political commentator, actor, and television ...

Alma mater: [Northwestern University \(BA\)](#) Parent(s): [James William Colbert Jr.](#)
Years active: 1984–present Relative(s): [Elizabeth Colbert Busch](#) (sister)

Early Life Early Career In Comedy Television Career Politics

Colbert was born in Washington, D.C., the youngest of eleven children in a Catholic family. He spent his early years in Bethesda, Maryland. He grew up in the Charleston suburb of James...

[Continue on en.wikipedia.org »](#)

In Google’s experimentation so far, they have identified that tabbed snippets improve the traffic to the pages the snippets are presented for, accounting for an increase in clickthrough rate of 1% - 8% based on the position of the result.

Additionally, we may be able to use Section Topics data to the schema.org metadata of our pages, further improving SEO in Google Search results. 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 is complete and we are producing those topics, we plan to explore this further.

We are excited about the progress we’ve made in Year 2 towards a demonstrably better search experience on the Wikipedias that will allow new users to search and refine searches in an

intuitive and familiar interface. We're confident that the learnings from the successes of MediaSearch will lead us to demonstrate similar success on the Wikipedias in Year 3.

Finalize features for improved related content suggestion/recommendation capabilities for editing

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 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.

What is changing?

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. So we aim to increase content numbers more slowly over a longer time period, using the infrastructure and tools built during the grant period.

The 5 million number is not achievable in the grant timeline for several reasons:

- Most wikis will not allow bots or other bulk edits to add images, so we can't use the same bot strategy as we did with Structured Data on Commons to quickly add vast numbers of images. Changing local Wikipedia policies requires a long timeline and involves social work, not technical work.
- Implementation has shown that there are a limited number of image suggestions available per wiki, especially for smaller Wikipedias in emerging communities. For example, across Cebuano and Arabic Wikipedias, where we worked with local communities to add images, , there were 473,051 unillustrated articles with potential images for illustration, which led to 68,653 images added.
- Cultural and logistic issues emerged through the user-feedback process. For example, some wikis discourage over-illustration of pages; not all pages actually need illustration; and some topics have far more related contributions on Commons than others.

In exploring ways to add images, we partnered with bot writers from the Cebuano and Arabic Wikipedias, which are two wikis that allow bots and are two of the only communities where mass addition was possible. While our partnership led to an addition of 68,653 images with a lower-than-average reversion rate (this is a good thing!), that is not replicable on most other wikis.

Below: A screenshot of a [Cebuano Wikipedia article](#) about a cabbage plant species, newly illustrated with images from a bot

The screenshot shows a Wikipedia article for *Iberis bernardiana*. The article text includes a summary and a list of references. Two botanical illustrations of the plant are displayed. The right sidebar contains taxonomic information: **Iberis bernardiana**, **Siyentipikinhong Pagklasipikar**, **Kaginharian: Plantae**, **Kabahig: Tracheophyta**, **Kahutong: Magnoliopsida**, **Kahanay: Brassicales**, **Kabanay: Brassicaceae**, **Kahenera: Iberis**, **Espesye: Iberis bernardiana**, **Siyentipikinhong Ngalan**, **Iberis bernardiana** Godr. & Gren., **Laing Ngalan**, *Iberis bubanii* Deville, *Iberis benthamiana* Boiss. & Reut., and *Iberis aniensis* Foucaud & Rouy. The bottom of the page shows a category list: **Mga kategoriya: Tanom nga ripolyo paghimo ni bot | Paghimo ni bot 2022-03 | Tanom nga ripolyo | Iberis**.

A better plan for high-quality image additions

As stated in the grant proposal, we are trying to increase access to editing in emerging markets and on mobile, and increase effectiveness of editing in those places. We think a better strategy to achieve that impact is pursuing a targeted user campaign strategy, focusing heavily on our editors' and readers' content needs. Our plan has two target user groups with separate strategies and features:

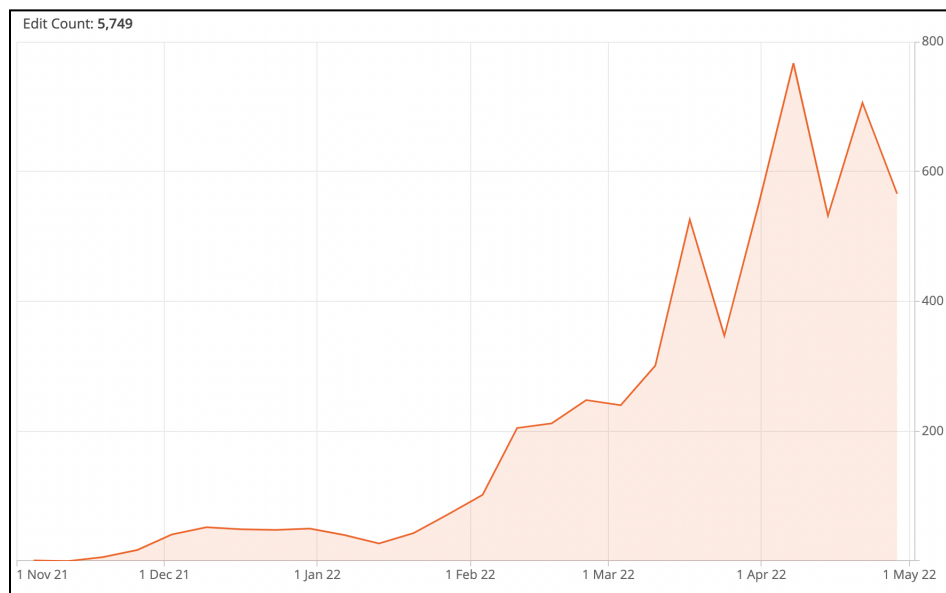
- We will increase contributions from newcomers (especially newcomers with topical expertise) via a structured task that walks newcomers through the process, with a goal to increase the number of users who add imagery to articles.

- 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.

Pursuing this targeted user campaign will give us high-quality contributions and will allow even more long-term sustainable growth in image addition on articles. Through user events and campaigns organized around our work, we have seen that our outreach efforts to attract expert contributors need more precise image matching. For example, a GLAM user group in Argentina was gathered in an event to test the image-suggestions tools. They wanted to focus the event on adding images that were related to Argentina. Solving that problem for the users was not easy, because there was no way to narrow down the suggestions to articles and images about Argentina specifically. Our existing article topic system ([ORES](#)) did not identify many articles related to Argentina. To adapt, we will use the section topic service to identify sections of articles that are related to Argentina and suggest images for those, giving a user group from Argentina many more images and articles to choose from for their event. This work to give users more precise information about sections is coming in Year 3 of the grant, and will also increase the number of available image suggestions. Thus, the section topic work will reinforce the image-suggestion work and enable more users to add more images to articles.

In Year 1, we initiated a successful proof of concept to automatically suggest media from Commons for editors to add to articles on Wikipedia as described in the first-year grant report. In Year 2, the POC has been further developed into a structured task for newcomers that allows them to easily add suggested images to Wikipedia articles based on their topics of interest. This newcomer task was launched in November 2021 and is now the default suggested edit for 40% of newcomers on both desktop and mobile on Spanish, Arabic, Czech, Bengali, Persian, French, Turkish and Portuguese Wikipedias. More Wikipedias will get the feature in the coming weeks. The feature has resulted in 2,582 unreverted image contributions as of March 31, 2022.

Below: A chart showing the growth of image contributions using the add-an-image tool. The y-axis represents total edits per week across the wikis with the add-an-image tool enabled.

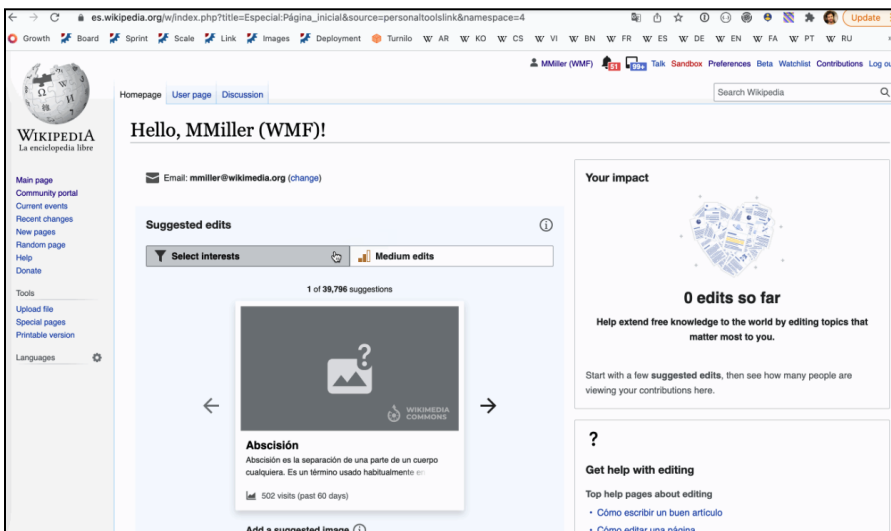


One of our most engaged newcomers so far is [User:Mikael1910](#), working on Czech Wikipedia. They've added over 153 images, with no reverts, since the tool has been available. We asked this user to tell us how they got involved, and they told us: "I attend(ed) the Wikipedia course for senior citizens and I registered prior to the first lesson . . . the add-an-image feature was shown to me, and I decided to try it out. I like the tool." Here is an example of an image that this user added to an article:

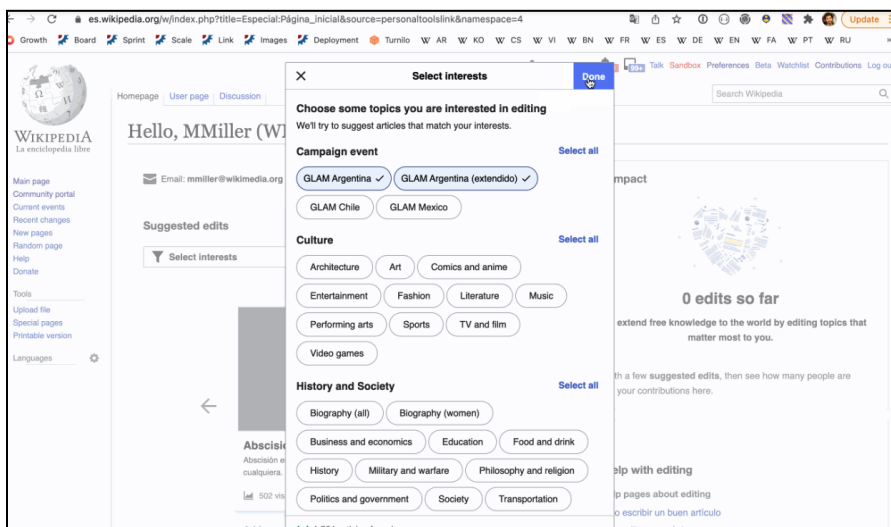
The screenshot shows the article page for "The C Programming Language" on the Czech Wikipedia. The main content is in Czech, describing the book and its authors. A thumbnail image of the book cover is displayed on the right side of the article. The page layout includes a top navigation bar with user options, a search box, and a sidebar with various navigation links. A banner at the top of the article area promotes a writing challenge.

We've also run several events to encourage GLAM experts to explore this feature in Mexico and Argentina, in collaboration with their local Wikimedia affiliates. For these events, we wanted to know if we could help people who are experts in their field but new to Wikipedia successfully contribute knowledge. At the event in collaboration with Wikimedia Argentina, 20 users added 49 images to articles. The event with Wikimedia Mexico had over 50 engaged participants, including some representation from GLAMs like the National Library of Mexico. [Here](#) is the Wikimedia Mexico blog post about the event. For these events and future ones like them, we developed the ability for users to filter via a combination of topics, like articles in Argentina about transportation. This lowers the barrier to participation for users like these, so they can focus on their areas of expertise, where they have the most impact in sharing knowledge. The series of eight images below demonstrates the process that newcomers are walked through to determine whether to add an image at one of these events.

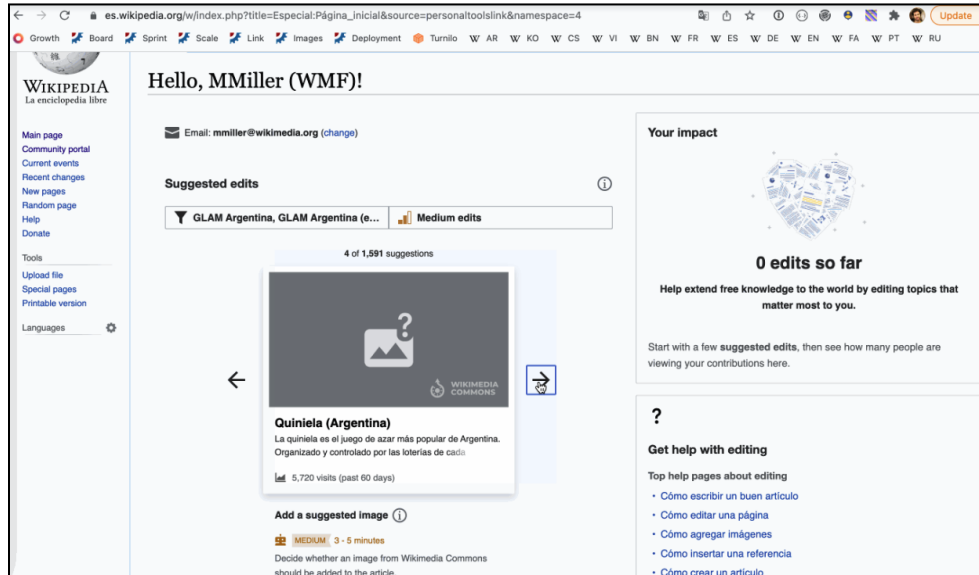
Step 1: The user is presented with a “Suggested edits” tab (middle left in image below), where they’re asked to click the “Select interests” drop-down menu.



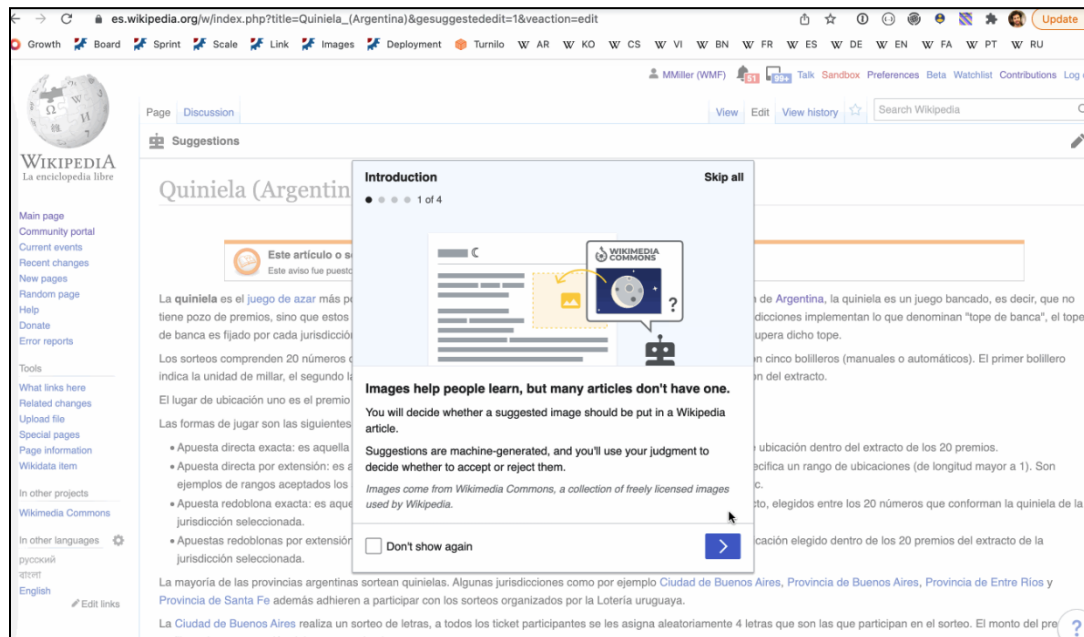
Step 2: The user then selects the “Campaign event” that they’re participating in – which in the case below means the “GLAM Argentina” options, and then the “Done” button.



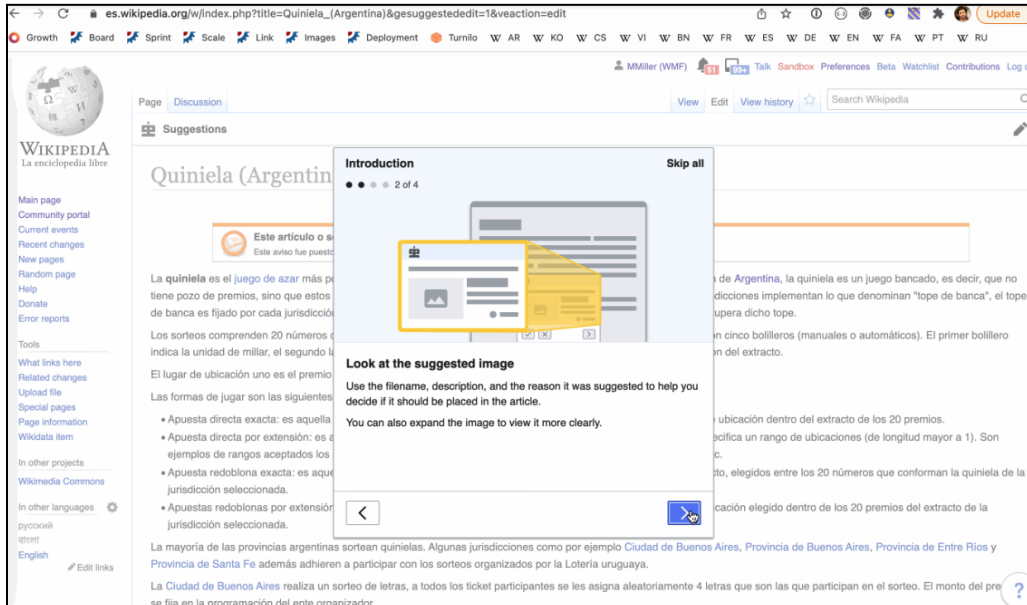
Step 3: The user then sees a choice of Argentina-related articles that they can add images to — letting them hit the right arrow button, as seen below, to see other potential articles. The image below shows that the user has stopped on the [Spanish-language article on Quinela \(Argentina\)](#), which is about an Argentine game of chance.



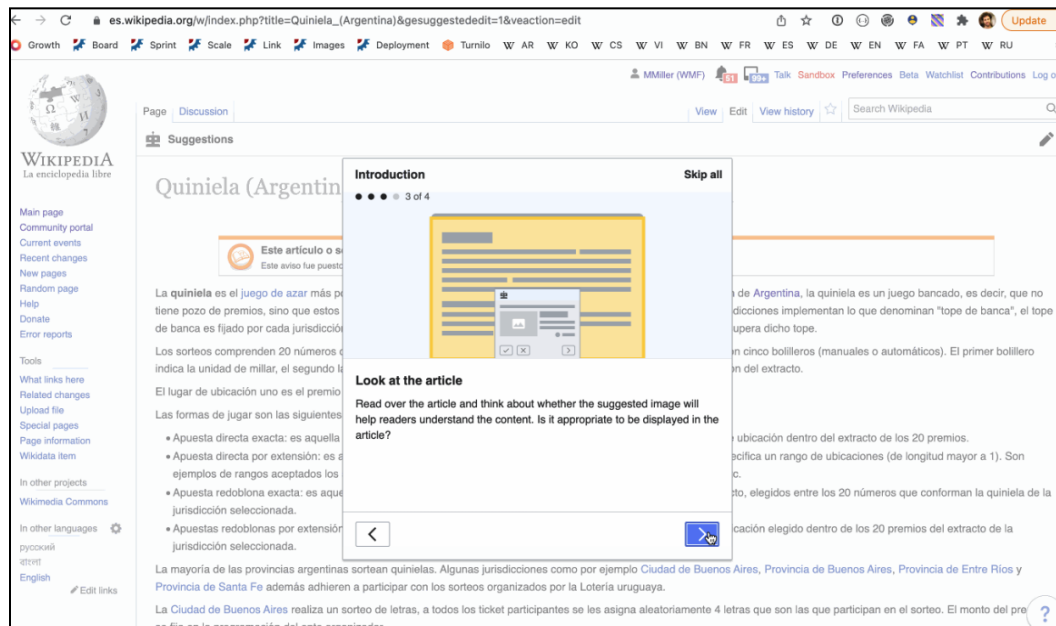
Step 4: By clicking on that article choice, the user then sees a brief guideline atop the article, as seen below. The guideline's first slide explains that the user will decide whether an image should be part of the article, and that the choices are machine-generated and come from Wikimedia Commons. The user advances the slides by clicking the blue-boxed arrow.



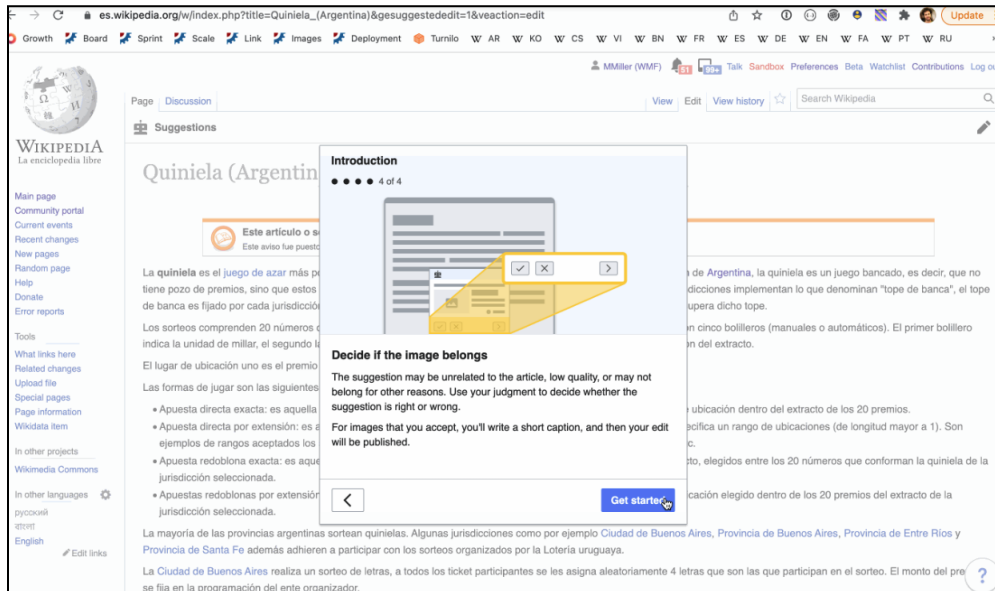
Step 5: The next slide then gives them clear instructions to help them decide on using the suggested image.



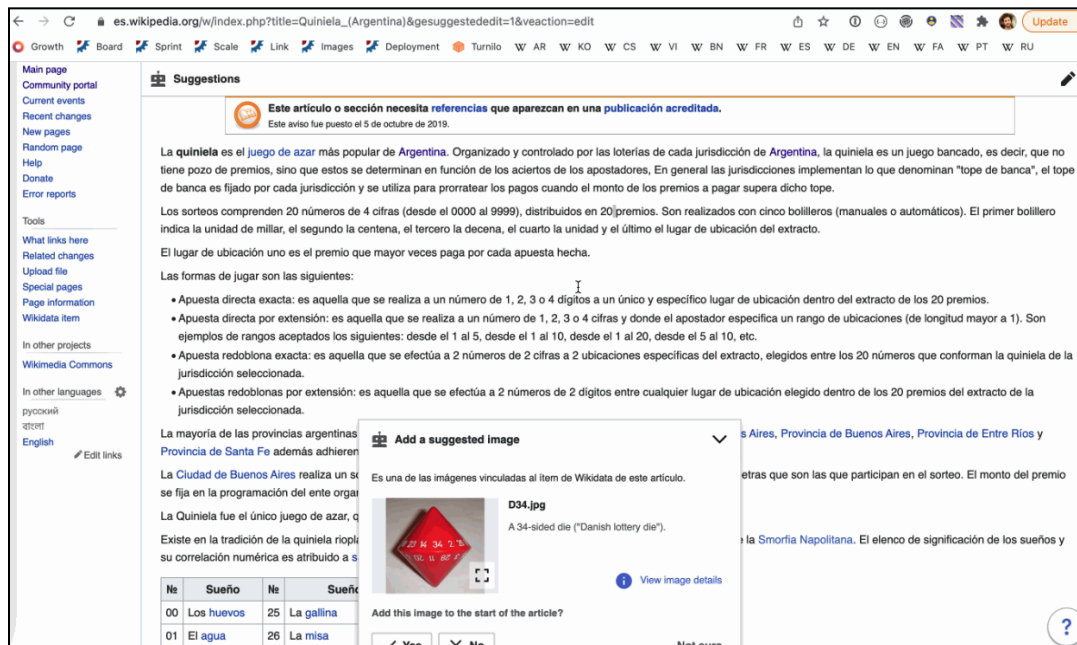
Step 6: The next slide, as seen below, then instructs the user to read the article that's being considered for an image — and whether the image under consideration “will help readers understand the content.”



Step 7: The next slide, as seen below, then reviews other reasons that the image may be inappropriate for the article — for example, the image may be of low quality. The slide’s blue button then gives the user a “Get Started” option.



Step 8: The user then sees the image atop the Wikipedia article, giving them a chance to click either “yes” or “no” on adding the image.



Meanwhile, we are also progressing in the second fork of our image-addition strategy: helping experienced editors contribute more effectively. In development is a task that notifies experienced editors of image suggestions for articles on their [watchlists](#), and covers a broader array of images. We expect to launch the first version of this in June 2022.

To enable these experiences, we've also developed infrastructure to create, store, and use generated datasets about our content. This enables the management of the dataset for image suggestions. It will also further enable future natural language processing work and enable connections across our content. Beyond the grant period, these centralized data infrastructure capabilities will allow teams across the Foundation to create new data products and relate more content, making it easier to build products that enable access.

We built a data pipeline orchestration pathway using an open source tool called Apache. This allows data producers to schedule processes to extract, transform, and load data ([ETL](#)), which is a common data integration process that makes data usable. An SDAW example: For the image suggestions product, the ETL takes relevant content from Commons, Wikidata, and the Wikipedias, and generates a new dataset that suggests Commons images for unillustrated Wikipedia articles. After the ETL, this Image Suggestions dataset is ready to be used through an API. Editor tools and bots can consume the Image Suggestions dataset to help editors find an image that will enrich articles. In addition, the teams are building reusable components on the platform that allow common patterns to be implemented more simply by data producers, as with loading data to our storage solutions. For data access, we are also building a generic method to read the data that is designed to scale across all of our generated datasets.

Long term, these data-platform capabilities will enable teams across the Foundation to build data products more easily and quickly — think building three new features a year instead of one. It will remove the need to build bespoke infrastructure and reduce overall technical debt. Our work thus far on SDAW has illuminated the need for shared infrastructure that can be used across multiple teams and features, and has helped us make this shared infrastructure a priority across the Foundation. A prime example of fulfilling that need is the data platform, which has allowed two teams to develop two image-suggestion tools that allow images from Commons to be added to the Wikipedias while optimizing for different audiences.

Additionally, we plan to start a project that delivers an event platform to enable data products to be created based on data events (currently our data products rely on large data dumps and batch schedules). This platform would allow data producers to create data products, for example, when a new wiki page gets created or an existing page is edited.

Identify strategies to experiment with and encourage the use of new features in Year 3 and beyond through a series of community-based pilot projects.

Our Year 2 work leaves us in a good place to begin Year 3, and we have worked on a few forward-looking strategies to help us launch strongly. In Year 3, we will wrap up building the infrastructure and features and begin to focus more on getting people to use them. We will continue to build upon the newcomer events in Wikimedia Mexico and Wikimedia Argentina, as those have proven successful so far. We also have plans to work with Wikimedia Portugal to do events and a contest around image suggestions with experienced users and GLAMs.

We are also exploring a partnership with the Digital Public Library of America (DPLA) to potentially expand a [proof of concept citation tool](#) that would allow users to use structured data to more effectively caption images on Wikipedia; hold events with DPLA's partners to increase the usage of Structured Data on Commons to describe GLAM images, thereby making them available for the image suggestions pipeline; and to teach other cultural heritage institutions how to more effectively contribute their images and structured data to Commons. This will help us make more images available to share on Wikipedia, further empowering our image suggestion tools. This work will build on our [existing partnership with DPLA](#).

We will also continue to partner with interested community members who would like to build tools and gadgets using the structured data infrastructure we've developed. We hope to sponsor a variety of community-led experiments in this space in Year 3 and will be releasing a request for proposals to help encourage this type of participation. An example of this type of tool is [WikiCrowd](#), which helps Wikimedia editors more easily connect and enrich the linked data network.

We will also continue to communicate with our communities, including the 40+ community members who are subscribed to our [Structured Data Across Wikimedia newsletter](#).

Year 3: A Look Ahead

In Year 3, we plan to focus on releasing the in-progress features to our communities, refining them, and gaining community adoption.

In the grant proposal, we said that milestones from this period of development could include:

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

- We aim to continue refining the image suggestions features, both for newcomers and for experienced users. We also aim to take 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
 - We are in the process of building improvements to the Search experience on Wikipedia, and will partner with interested community members who would like to explore building tools or gadgets in this space that take advantage of our Structured Data Across Wikimedia infrastructure.
- Experiment with connecting topically related content across languages as part of the contribution process
 - We are already doing this as part of the image suggestions work, and plan to continue to explore how structured data might help our translation features.
- Exploration of integration of concept metadata with anti-vandalism and quality control systems
 - Once Section Topics is available to users, we hope to hear more feedback from the community on this.
- Design and launch 2–5 community-based pilot projects focused on experimenting with and encouraging adoption of new features. The goal of the pilots will be to increase adoption of reading, editing, or reuse features in emerging or diverse user segments.
 - The plan for this is described in this report’s “Identify strategies to experiment” section on Page 25.

Other Challenges

We are confident that we’ll deliver the grant in scope and on schedule, but we want to describe some “behind the scenes” challenges that we have faced — and overcome — as we’ve continued our work on the project:

- Chief Product Officer Toby Negrin left the Foundation in November 2021, after which Carol Dunn and Margeigh Novotny were named interim Chief Product Officers. Combined with a series of other staff transitions, along with a recognition that technologies and our capabilities have evolved over the past 1½ years, we spent time re-evaluating the Product department’s overall plan — while at the same time recommitting to Structured Data Across Wikimedia.
- As happened across the United States during the pandemic, when much higher percentages of people left their jobs, the Structured Data Across Wikimedia team experienced an unusual amount of turnover. As a result, we’ve had fewer staff covering multiple priorities, especially in product management — which has led to a delay in identifying and scoping products to be delivered for this grant.

- The pandemic also significantly reduced the real-time availability of team members during the grant period.

In the past six months, we've spent much time on hiring replacements for the team members who left. In addition, all new staff require onboarding. All these duties require time and commitment beyond the day-to-day time and commitment required for work on Structured Data Across Wikimedia. Overcoming these challenges is another reason we're proud of the results we've achieved since our last report.

Conclusion

We have delivered, iterated, and learned a lot in the past 12 months of the Structured Data Across Wikimedia program. We've shipped a new feature in a new form factor and delivered two major pieces of infrastructure. We've seen major uptake and impact from another feature, the first feature we shipped in Year 1. We've partnered with our communities and movement-aligned organizations to introduce these new features to our existing and new audiences and planned to ramp up engagement even more in Year 3. We've worked with bot editors to add 69,000 images to Wikipedia articles and learned from that experience how to sustain continued impact in image enrichment through targeting high-quality human-led additions.

We also continue to grow as a team and organization as we've overcome the aforementioned challenges to deliver and have impact. We are more efficient, managing several work streams to achieve our goals within the grant timeline, and to deliver quality technical solutions that have impact now and position us for the future.

The heart of the program is managing our content to enable new knowledge experiences for contributors and readers. We achieve this goal by building solutions that create structure around our content. In creating that structure, we create metadata for our content, and this metadata is then used to make the connections we seek between and among knowledge entities.

This requires building the platforms to support all the structured content requests and needs from our community, including:

- **Machine Learning/Artificial Intelligence Platform:** using technology to create and make connections.
- **Data Platform:** storing content insights and connections, and making them available.

- API Platform: making content insights and connections more accessible to the world.

We've worked on all of these in Years 1 and 2 of the program, and will continue to build them out and use them to make knowledge easier to access and participate in Year 3.

These platforms are the foundation for us to achieve our goal to become the essential infrastructure of the ecosystem of free knowledge by 2030. That's what Structured Data Across Wikimedia is establishing: A future when the Wikimedia projects are more essential than ever for readers, for editors, and for anyone searching for knowledge and discovering how integral the Wikimedia projects are to the wider internet.