

# **Structured Data Across Wikimedia: A Proposal for the Alfred P. Sloan Foundation**

## **PUBLIC COPY**

This grant proposal has been accepted and funded by the Alfred P. Sloan Foundation. The following seven items summarize the high level goals of the grant program:

1. Build infrastructure and tools to allow structured metadata from Wikipedia Commons to be added to other content across Wikimedia projects, including Wikipedia itself
2. Redesign and improve the search experience
3. Media added to 5 million content pages
4. Gather community input on the proposed changes throughout the process
5. Encourage the use of new features within the Wikipedia community
6. Increase the number of readers, especially from underserved communities
7. Increase the number of contributors and editors, especially from emerging markets and on mobile

Designs, mockups, and features contained within the proposal are for illustrative purposes only. They are ideas and are not indicative of formal plans or commitments. The Wikimedia Foundation will be in communication with the community for feedback as we further develop plans to fulfill the high level goals stated above.

Note: The time frame stated on the first page of the proposal is incorrect. The correct time frame is July 1, 2020 to June 30, 2023 (3 fiscal years).

# Structured Data Across Wikimedia: A Proposal for the Alfred P. Sloan Foundation

## Introduction

On the internet, any grouping of media with structured data – whether photos, videos, or articles – is infinitely more useful than media without structured data. That’s because machines can read and analyze structured data, which lets people across the internet – from beginning users to the most advanced – more easily search and find that media. The promise of structured data is why, with Sloan’s support, we began our Structured Data on Commons project in 2017. Wikimedia Commons is the world’s largest free-to-use media library, but with structured data it’s becoming something better: A knowledge source with millions of media files that are easier to view, search, edit, organize and reuse on Wikimedia and across the Internet.

**Project scope:** We’re proceeding to a new, key phase of structured data that will connect structured data to other Wikimedia content to improve users’ search on Wikimedia projects, and help our editors add media to articles, while enabling other computer-aided editing features and new ways to access Wikimedia knowledge.

**Time frame:** July 1, 2020 to June 30, 2030 (3 fiscal years)

We’re now evolving our project into a key, second phase that will bring structured data to Wikipedia and other Wikimedia websites – transforming these sites in the same way that we’re transforming Commons.

Our new project has two crucial components:

- We’re connecting structured data to other Wikimedia content, so that structured data goes beyond images on Commons to the core of the Wikimedia projects: Wikipedia itself. With this we’ll be able to bring all the benefits of structured metadata to those projects: newer and easier ways to read, edit, and access the knowledge within the Wikimedia projects.
- We’re giving Wikimedia users a more inviting, more efficient way to search and find content – using the ability of structured data to surface facts on Wikipedia, answer users’ questions about subjects, and provide users other improved ways to find exactly what they want.

Our new project will have a far-reaching impact on Wikipedia and our other knowledge sites. It helps us match content on one Wikimedia knowledge site (like Commons or a different language Wikipedia) to another Wikimedia knowledge site, which will help readers dive deeper into the knowledge ecosystem and help editors more easily connect and enrich the linked data network.

Our new project, we believe, will mirror the impact of our Structured Data on Commons project, which has inspired people around the world to contribute wording that's making Commons media more discoverable in more languages – as happened in November in Nigeria. There, Africans like [Isaac Oloruntimehin](#) helped add over 61,000 pieces of depiction metadata to more than 9,000 photos while testing a new mobile micro-contributions tool called [ISA](#), which makes it easy for anyone – including beginning contributors – to add structured data to Commons images. We write about Isaac's success later in this proposal, and it's especially inspiring because Isaac is a beginning Wikimedian and a student at Nigeria's University of Ilorin. Isaac is the future of Nigeria – and he's also the future of the Wikimedia projects, because editors like him are adding content that's missing from our projects, in languages that need bolstering, and from geographical areas where current Wikimedia participation is low. Structured data is also the future of our projects. That's why our new phase of structured data – from its technical foundation to its social component – is so promising.

Our new phase will expand the scale and reach of our work on structured data, and also solidify its impact across the Wikimedia projects and the wider internet. We will use the new metadata and the structured data on Commons to suggest images for editors to add to Wikipedia articles. We can also create new ways for new and existing readers to access and interact with the content our editors work so hard to create. Our Structured Data on Commons project gave us the foundation to move ahead with confidence, knowing we have the right team in place and also the right vision to make it happen.

## The Importance of Our Future Work: Why We're Initiating the Next Phase of Structured Data on Our Projects

In the last three years, our engineers built the critical new infrastructure that has allowed Commons to incorporate structured data – and their work has led to an important milestone: 11.4 million media files on Wikimedia Commons now have structured data. As we've worked with the Wikimedia community, new user interfaces, databases, backend functionalities, and community tools have also emerged from the three-year project. As importantly, the Wikimedia community has emerged and embraced the project, as we note in reviewing our work in this proposal's first Appendix (see Pages 23-28). But in the course of our three years' work, different teams at the Wikimedia Foundation have come to the same realization: We need more advanced metadata for all content and APIs to provide better search results, which would make the content more accessible, more discoverable, and more useable for translations and other needs.

Here are three issues that highlight our necessity to evolve structured data across Wikimedia:

1. We need new ways to read content for a new segment of readers

We’ve embarked on projects to attract new readers from underserved communities around the world, and our work indicates that readers in these regions and age groups prefer “snackable” content. But as we’ve started to build out systems to provide bite-sized content, it’s clear we don’t have adequate systems in place to provide concise facts from articles. This void is leading editors to “chop up” article content on a case-by-case basis – a level of manual labor that’s impossible to scale.

2. We need computer-aided editing for a new segment of contributors

We’ve started investing in micro-contributions as a new method of gaining and retaining new editors. We’ve experimented with adding structured data captions [via](#) our Android app, and evidence shows it’s driving higher retention and numbers of active editors. We’re also venturing into user contributions augmented by AI with our new Computer Aided Tagging (CAT) tool, which uses Google’s Vision API to analyze Commons images and suggest [depicts tags](#). But other Wikimedia sites lack these tools, so we have a new opportunity to extend tools like CAT into our other projects – and a new opportunity to link the new content that emerges across Wikimedia’s projects. With content metadata, we’ll be able to connect related content items regardless of [file type](#), language, or hosted project. Two examples: We could suggest relevant images from Commons when an editor is editing a page on Wikipedia, and we could prompt multilingual users to translate facts missing in other language Wikipedias – as in, “This fact is in the Spanish version of this Wikipedia article. Do you want to add it to the Vietnamese version?”

3. We need a search method that is friendlier and more delightful to use

Although we’ve made significant advancements to Wikipedia’s search function, our search system across the Wikimedia projects lacks the ability to understand a user’s intent. For many queries, this leads to subpar search results. Our current search, for example, doesn’t always surface relevant files or pages because they lack a lexical or textual match, which badly shortchanges our users’ queries – as in a Commons’ search for “big cats,” which only finds files that contain the exact phrase “big cats” and ignores millions of files that contain applicable photos of tigers, lions, leopards, and related animals.

Table 1: A summary of the needs and expected impact of content metadata and understanding query intent

The Need	The Impact
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	An increase in usage on new platforms, along with a measurable increase in new users reached

mobile	
Users in emerging markets and on mobile can contribute using related content suggestions when editing and other computer-aided editing improvements	An increase in editing in emerging markets and on mobile and an increase in effectiveness of editing in those places
New users can search and refine searches in an intuitive and familiar interface	Search is demonstrably better, including updated user interfaces that are easier to use for new visitors

We believe that this new phase of structured data – which improves our ability to serve, edit, search for, and share information from Wikipedia, Commons, Wikidata, and all the Wikimedia projects – aligns with the Sloan Foundation’s goals of Universal Access to Knowledge. Like the other projects you’ve funded under this rubric – including the work of the Berkman Klein Center for Internet & Society, and the Digital Public Library of America – our structured data work is advancing greater access to knowledge for a great number of people, at a time when greater access is more important than ever.

## Our 3-Year Plan: How We'll Bring Structured Data to Wikipedia and Beyond

### What is to be done

This program will add structured metadata to content across Wikimedia projects, making it more machine readable and therefore more accessible, readable, and discoverable by anyone on the Internet. One of the features we will build will use this structured content metadata to help users add media to 5 million content pages. This content metadata will also help us experiment with other computer-aided editing tools that make editing easier and more accessible to more editors around the world, especially those in emerging communities and on mobile.

### Project timeline and key milestones

#### **Year One—Foundations and Research:**

Our work in the first year will generally focus on making immediate impact where we can while laying the groundwork for more advanced use cases and implementations that will come later in the project. Our infrastructure teams will be building the necessary infrastructure for Year 2, while our front-end and search optimization engineers work on immediate impact.

We plan to start with search, and we will cover the following aspects:

- Gathering community input on the proposed changes. Community input will be essential to the success of this program. This will occupy a significant percentage of time and work, and our plans will change in response to community feedback and ideas. We have an incredibly passionate, intelligent, and opinionated community, and working with them will make the project better and more successful. Sometimes, it will also make the project slower and more unpredictable. Community work in the first year will include socializing the problem statement and proposed changes, generating and refining ideas with community stakeholders, and working with the community to craft any necessary technical requests for comment or policies.
- Redesign of the search experience on Commons so that a new user can search and refine her searches in an intuitive interface that feels familiar. This work will include community consultation in the design phase, prototyping one or more approaches, and consultation with the Search Platform Team.
- Design and prototyping of improved related content suggestion/recommendation capabilities for editing (first deployed on Commons, but perhaps later used on other projects). Currently finding and adding media, references, links, and other connections between content involves complex digital literacy skills, such as search techniques and navigating wiki content structures.

Current Commons search results for “Hudson River School” only shows files where the phrase is included, due to lack of understanding of the query intent.

The screenshot shows the Wikimedia Commons search interface. At the top left is the Wikimedia Commons logo. The search bar contains the text "hudson river school" and a "Search" button. Below the search bar, it indicates "Results: 20 of 3,150". There are advanced search options for sorting by relevance and filtering by search type (Gallery, File, Help, Category, Creator, Institution). A sidebar on the left lists various site navigation options like "Main page", "Upload file", and "Tools". The main content area displays a list of search results, each with a thumbnail image and a text description. The results include entries for Thomas Moran, Asher Brown Durand, Alfred Thompson Bricher, and a book titled "DINNER (held by HUDSON RIVER SCHOOL MASTERS' CLUB)".

Not logged in | Talk | Contributions | Create account | Log in

Special page

Search Wikimedia

## Search results

hudson river school

Search

Results: 20 of 3,150

Advanced search: Sort by relevance X

Search in: (Gallery) X (File) X (Help) X (Category) X (Creator) X (Institution) X

Search categories · View other tools

Create the page "**Hudson river school**" on this wiki!

**Thomas Moran**  
Thomas Moran was an artist of the **Hudson River School**.  
413 bytes (10 words) · 13:58, 22 November 2014

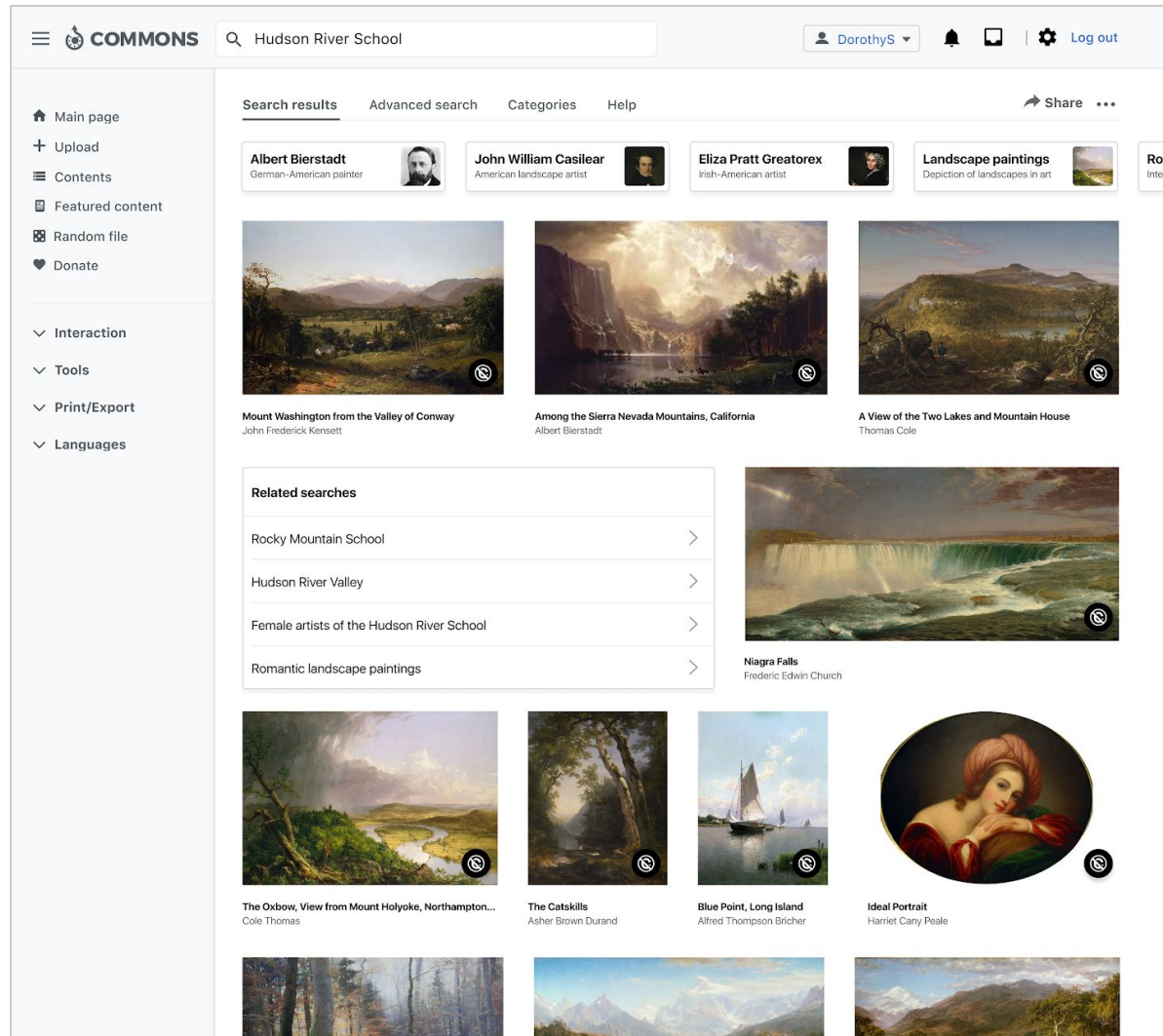
**Asher (Brown) Durand**  
[August 21, 1796 - September 17, 1886] was a founding painter of the **Hudson River School of Art**. Before he became a painter, Asher Brown Durand worked as  
4 KB (237 words) · 16:11, 1 March 2017

**File:Alfred Thompson Bricher.jpg**  
English: Alfred Thompson Bricher (1837–1906), american painter and member of the **Hudson River School**.  
(200 × 254 (7 KB)) · 06:08, 17 January 2015

**File:Hudson River School Painters.jpg**  
DescriptionHudson River School Painters.jpg English: Thomas Cole, George Innes, Frederic Edwin Church, Albert Bierstadt. Date 11 January 2013, 23:35:48  
(800 × 600 (298 KB)) · 20:09, 25 September 2017

**File:DINNER (held by HUDSON RIVER SCHOOL MASTERS' CLUB (at) "THE TEN EYCK,ALBANY, NY" (HOTEL.) (NYPL Hades-276056-471120).jpg**  
by **HUDSON RIVER SCHOOL MASTERS' CLUB (at) "THE TEN EYCK,ALBANY, NY" (HOTEL.) (NYPL Hades-276056-471120).jpg** English: DINNER (held by **HUDSON RIVER SCHOOL**)  
(1,805 × 2,377 (2.63 MB)) · 14:41, 31 December 2017

**FUTURE Search concept** for Commons – better results and recommendations that are topically related, along with usability improvements in a modern interface.



## Year Two—Implementation and Experimentation:

In this phase, we'll build upon the previous work in metadata and search. We will add structured metadata to long form content on at least one Wikimedia project.

Milestones from this period of development would include:

- New moderation capabilities that allow community members to edit/monitor that metadata if necessary
- Exploration of the use of metadata 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
- Finalize features for improved related content suggestion/recommendation capabilities for editing (first deployed on Commons, but perhaps later used on other projects).
- 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. (See pilot strategy in Year 3.)

**Current Wikipedia results** for “who are the female nobel laureates?” returns a number of different list articles as the top results.

The screenshot shows the Wikipedia search interface. The search bar contains the query "who are the female nobel laureates?". Below the search bar, there are options for "Advanced search" (set to "Sort by relevance") and "Search in" (set to "Article"). The search results section displays a message: "The page 'Who are the female nobel laureates?' 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." Below this message, there are two list articles: "List of female Nobel laureates" and "List of Nobel laureates by country". The "List of female Nobel laureates" article is highlighted, showing its size (29 KB, 871 words) and last edit date (14 October 2019). On the right side, there is a "Results from sister projects" section with a link to "Maired Maguire".

**FUTURE Search result concept on Wikipedia that immediately provides topically relevant matches in an updated interface.**

The screenshot shows a Wikipedia search results page for the query "who are the female nobel laureates?". The page features a search bar at the top with the query entered, and a user profile for "DorothyS". The search results are displayed in a list format, showing the year of the award, a small portrait of the laureate, their name, a brief description, and the field of study. The list includes Marie Skłodowska Curie (1903), Bertha von Suttner (1905), Selma Lagerlöf (1909), Marie Skłodowska Curie (1911), Grazia Deledda (1926), Sigrid Undset (1928), Jane Addams (1931), Irène Joliot-Curie (1935), Pearl S. Buck (1938), Gabriela Mistral (1945), and Emily Greene Balch (1946). To the right of the list, there are sections for "Related searches" (e.g., "how many female scientists have won nobel"), "Results from other Wikimedia projects" (including a Wikiquote by Doris Lessing and Commons images of Nadia Murad and Gertrude Eilion), and a "Share" button.

**FUTURE computer aided editing concept** for Wikimedia projects in which relevant images can be suggested as part of the editing process.

WIKIPEDIA  
The Free Encyclopedia

DorothyS
🔔
📱
⚙️
Log out

↶
↷
A
🔗
🗨️
Cite
⋮

Publish...

## Transiting Exoplanet Survey Satellite

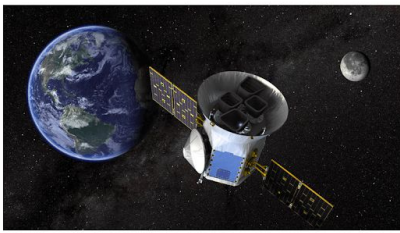
From Wikipedia, the free encyclopedia

The **Transiting Exoplanet Survey Satellite** (TESS) is a [space telescope](#) for NASA's [Explorers program](#), designed to search for [exoplanets](#) using the [transit method](#) in an area 400 times larger than the area of the sky surveyed by Kepler during its mission.<sup>[6]</sup> It was launched on April 18, 2018, during its two-year primary mission, it will search for 20,000 transiting exoplanets,<sup>[10]</sup> compared to the 3,500 that Kepler found when it launched;<sup>[11]</sup> however, as of December 2018, it has discovered more than 1400 candidate exoplanets, of which 105 have been confirmed.<sup>[12]</sup> The *first light* image from TESS was released publicly on September 17, 2018.

### History

The genesis of TESS was as early as 2005, when it was proposed to be funded from private funding by individuals. In 2008, MIT proposed that TESS become part of the [Small Explorer program](#) at Goddard Space Flight Center, but it was not selected. It was resubmitted in 2010 as an [Explorers program](#) mission, and was approved in 2013 as a Medium Explorer mission. TESS passed its [critical design review](#) (CDR) in 2015, allowing production of the satellite to begin. While Kepler had cost US\$640 million at launch, TESS cost only US\$200 million (plus US\$87 million for launch).<sup>[18][19]</sup>

Suggested contribution



File:Tess virtual map.jpg

NASA

Add image to article
Cancel

**Space observatory**<sup>[1][2]</sup>

NASA / MIT

2018-038A@

43435

[tess.gsfc.nasa.gov](mailto:tess.gsfc.nasa.gov)

[tess.mit.edu](mailto:tess.mit.edu)

**Planned:** 2 years

**Elapsed:** 1 year, 7 months, 17 days

**Spacecraft properties**

<b>Bus</b>	LEOSTar-2/750 <sup>[3]</sup>
<b>Manufacturer</b>	Orbital ATK
<b>Launch mass</b>	362 kg (798 lb) <sup>[4]</sup>

**FUTURE metadata concept** on Wikipedia that uses machine analysis to provide topical tags that can be confirmed or edited by users if the suggestions by the automated system are not correct. The confirmed data would be a starting point for connecting conceptually related content across Wikimedia projects and improving the ability to provide concise facts.

The screenshot shows the Wikipedia article for Ada Lovelace. The main text includes a lead sentence: "Augusta Ada King, Countess of Lovelace (née Byron; 10 December 1815 – 27 November 1852) was an English mathematician and writer, chiefly known for her work on Charles Babbage's proposed mechanical general-purpose computer, the Analytical Engine. She was the first to recognise that the machine had applications beyond pure calculation, and published the first algorithm intended to be carried out by such a machine. As a result, she is sometimes regarded as the first to recognise the full potential of a "computing machine" and one of the first computer programmers.<sup>[2][3][4]</sup>

The structured data sidebar on the right lists the following information:

- Paragraph 1:** 90 words, 567 characters
- MACHINE-DETECTED FOR VERIFICATION**
- Person:** 3
  - Augusta Ada King ✓
  - Lord Byron ✓
  - Charles Babbage ✓
- Organization:** 1
  - Countess of Lovelace ✓
- Date:** 2
  - 10 December 1815 ✓
  - 27 November 1852 ✓
- Nationality:** 1
  - English ✓
- Paragraph 2:** (collapsed)

Additional biographical data from the sidebar includes:

- Born:** The Hor... 10 Dece... London,
- Died:** 27 Nove... Maryleb
- Resting place:** Church... Hucknal
- Known for:** Mathem
- Spouse(s):** William I

### Year Three—Refinement, Polishing, and Reporting:

The conclusion of the three-year plan would involve combining the results from the previous two years to build new product features/capabilities that can serve multiple Wikimedia projects in many ways.

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
- Redesign of the search experience for new users on projects beyond Commons
- Experiment with connecting topically related content across languages as part of the contribution process
- Exploration of integration of concept metadata with anti-vandalism and quality control systems

- 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

## Project oversight and management

Work for this project will build on top of software, database structures, and team structures developed during the Structured Data on Commons project. We expect to reuse and extend existing features developed for Wikibase Federation and Multi-Content Revisions.

New software developed for the Structured Data Across Wikimedia project will be handled by existing teams and staff that have experience with our structured data and search technology stacks.

Program responsibilities	Team	Team management	Oversight
Overall management and delivery of program, including scheduling, risk management, change management, and communication	Program Management	Sr Program Manager: Amanda Bittaker	Chief Product Officer: Toby Negrin
User interface research and design Delivery of front end features and some platform tasks	Structured Data	Product Manager: Ramsey Isler	
Technical implementation of metadata and search API	Search Platform	Engineering Manager: Guillaume Lederrey	Chief Technology Officer: Grant Ingersoll
Assistance/consultation with backend work and integration of metadata into core user workflows	Core Platform	Director of Engineering: Corey Floyd	
Investigate query intent	Research	Principal Research Scientist, Head of Research: Leila Zia	
Financial management	Finance	Sr Financial Analyst: Alan Lau	Chief Finance Officer: Jaime Villagomez

## A brief discussion of the relevant technical issues

### Current state of content metadata and search

The current search infrastructure of Wikimedia is powerful, but it is difficult for new users to discover and use advanced search features, and difficult for other development teams to use the search infrastructure for new products. It is based on Elasticsearch and a standard full-text search approach. It supports ~900 wikis in ~300 languages. It has > 1 billion documents indexed, representing > 4 TB of primary storage. It serves ~1000 queries per second for the completion suggester, ~500 queries per second for full-text search, and ~150 queries per second for article recommendations.

The keyword-based search has language-specific analyzers for a few dozen languages, while other languages use default analyzers. Pageviews, incoming links, popularity and other signals are used on top of the textual content to improve ranking. The top ~20 wikis use machine learning-based ranking. No faceting or other ways to refine user intent are present.

While multiple advanced features are available for power users to refine searches (see Help:CirrusSearch for details), they are difficult for new users to discover and use.

Each wiki has its own search, with support for cross-wiki searching. File search from any wiki is directed to Commons.

Structured media metadata is accessible via search and captions are indexed for keyword search. Statement data about media is not used as a search signal on Commons. There are not yet UI elements for refining media search that leverage the structured media metadata.

There is a robust A/B testing infrastructure in place. Any major change to full-text ranking goes through an A/B test.

All updates are real-time, which does not allow for significant post-processing.

### Where do we want to be with search?

We want to modernize the search experience for Commons users by deploying industry best practices for search to create a more compelling user experience focused on satisfying user intents at a deeper level. These best practices are built by leveraging existing signal data as well as other advanced techniques. Our end goal is to deploy a search tech stack across Commons and other wikis that understands user intent, learns from aggregate user behavior and looks and feels like a modern search experience. Technically, we will support the main use cases of this project by bringing forward many of the features we use in other projects and built into our search engine (Elasticsearch). This will enable us to experiment on the next round of features we think can power search across all other sites.

### What are our options for getting there?

Building on our successful strategy of improving search in other areas like Wikipedia, we propose to take an experiment-driven approach to achieving our goals of an improved user experience for search users on Commons as well as downstream users of Commons data in other projects. We will start in year one by bringing over and leveraging many of the techniques we've used on other wikis to improve the relevance of Commons search. We will begin with simpler approaches like better indexing and weighting of content and richer query parsing (e.g., phrase identification, tuning of the importances of specific document features like titles and captions) based on doing an in-depth analysis of query behavior in Commons. This analysis will guide our intuition in choosing later approaches for understanding user intent. We also think providing users with better type-ahead results, as well as better faceting and filtering capabilities, will enable users to more proactively affect the quality of results without having to guess at how to rephrase their query. During this phase we will gain critical experience in understanding the behaviors of users and how they react to changes by leveraging the A/B testing framework mentioned above in the current infrastructure section. A/B testing is a critical part of success in search due to the often ambiguous intents of user queries. Finally, we will use the time in this phase to identify key user intents for use in the next phase described below.

As we move out of the foundational phase of the process, we will look to implement a number of industry best practices for search, always with an eye towards experimentation. Given the current volume of queries on Commons, we think we have enough data to begin to deploy modern machine learning and Natural Language Processing (NLP) techniques to better understand users' queries and thereby produce better search results. Our current infrastructure supports leveraging user behavior data to build more advanced relevance models like Learning To Rank (LTR) and other click-based models. We will experiment with deploying one or more LTR or click-based models during this phase of implementation. After we have the baseline relevance models in place we will also experiment with building and using one or more query intent classification models. Query intent classification is an approach that takes in a user query along with other features of the request and injects into the query processing pipeline one or more labels (think of them as hints) indicating what this user is most likely looking for. For instance, it might indicate that a user is interested in a specific category of things (e.g., "cats") or that they are looking for a very specific thing (e.g., "calico cat perched on a ledge"). From these intents, we can then do further semantic analysis on the query to identify parts of the query that enable us to better craft a query to our backend search system. For instance, in the "calico cat" example, we might identify "calico cat" as a phrase and weight its importance higher. We might also focus the use of the word "ledge" on a particular feature of the documents that identify physical features in the image. In the category case ("cats"), we might simply apply a pre-filter to the search that automatically boosts or restricts the query to the matched category.



For the tagging of content in other Wikimedia projects, we will explore a combination of techniques ranging from simpler string matching (e.g. regular expressions) approaches through to machine learned models (e.g. Named Entity Resolution) built on top of Wikidata. In the early stages of matching, we will focus on using the highest quality entities and focus them on matching against high quality content, both of which can be determined using ORES and other systems we have in place. We will also use the popularity of pages and data as a focusing factor for where we spend our time, as the task at hand is computationally quite expensive.

### How will we choose the options we take?

We will start by choosing options that are known winners in most other aspects of search within Foundation projects as well as those proven in other sites and shared in popular open source search communities that we participate in. Frankly, there is much low-hanging fruit on Commons search that we expect many of the standard relevance tuning practices (e.g., filtering and faceting, better type-ahead, better phrase search, better weighting of terms) outlined for the first phase of work to make noticeable differences. Many of these are also quick to implement, so they should give us some early wins.

For the second phase of infrastructure and modeling work, we will rely much more heavily on a data science and experimentation approach driven by A/B tests. All of the modeling and NLP work can be tested against historical data and reviewed by Wikimedia Foundation staff and Commons community members before being deployed into a broader end-user focused A/B test. Throughout all the phases, we will have members of our research and data science team advising and providing feedback on the impact of our choices, thereby enabling us to make data-driven decisions about what options we take throughout.

## The Road to Structured Data Across Wikimedia: How Structured Data Is Enabling New Uses and Tools for 57 Million Media Files

At its foundation, our new project will mirror the community-wide effort that has been a hallmark of our Structured Data on Commons project. We saw how well that collaboration can work on a recent Saturday in Nigeria’s capital, Abuja. There, a Nigerian university student named [Isaac Oloruntimilehin](#) (photo at right) [added structured data](#) to [a photo of an Abuja park](#), which was taken by a longtime Wikimedia editor and photographer, [Fawaz Tairou](#), who also lives in Nigeria. By adding “green space” and “garden” to the image, Oloruntimilehin helped ensure that Tairou's photo can be better searched on Wikimedia Commons – not just today but years from now. This makes images of Africa, enriched by Africans, discoverable both on Wikimedia projects and for those who





reuse our content. Because the tags use Wikidata, they are language agnostic – Isaac may have added the tag in his native languages, say English, and others can read it in their own preferred language, say Igbo or Hindi. This new dimensionality multiplies the impact of each metadata edit across all 280 languages that Wikimedia supports. And this new dimensionality fits in with our [2030 Strategy](#) to evolve our technology, products, and platform and to reach hundreds of millions of new knowledge seekers – including those in the poorest parts of the world where access to information is scarce.

Oloruntimehin made the additions during a [challenge](#) to test a new mobile micro-contributions tool called [ISA](#) that makes it easy for anyone, including beginning contributors, to add structured data to Commons images. With ISA, organizers can choose a predefined set of images on Commons and then ask contributors to tag these with multilingual structured metadata. During a challenge, points are counted for each contribution, so it's possible to organize competitions around tagging or other micro-contributions.

The Wikimedia Foundation facilitated the tool's development through [Wiki In Africa](#) (a South African nonprofit), the Histropedia team (longtime MediaWiki developers), and an African developer. The Abuja challenge was held during [Wiki Indaba](#), a conference that [brought together](#) Africa's Wikimedia community and Wikimedia Foundation staff members to brainstorm about current activities and also look to Wikimedia's future. We think ISA will be a popular fixture in that future because it makes adding structured data so much easier and so enjoyable. The [initial campaign](#) for Wiki Loves Africa showed deep engagement, with 37 participants adding 61,000 pieces of depiction metadata to over 9,000 images.

ISA was initially built for [Wiki Loves Africa](#), the annual Africa-centered contest, to provide better multilingual and structured descriptions for images uploaded to Commons. But ISA is also designed to be a host for any small competition. The tool helps create campaigns to improve one or several specific categories of images on Commons. In real time, the tool



provides stats that allow organizers to identify their challenges' best contributors. People are having fun with it. At Wiki Indaba, for example, the ISA challenge winner – with [412 contributions](#) – was Nigerian photographer Fawaz Tairou, who (as seen in the photo left) [was all smiles](#) at receiving the challenge's prize: A mug with an image of the ISA tool.

The Alfred P. Sloan Foundation funded Structured Data on

Commons, and your generous support has been critical in advancing our goals this far. Several

key technical developments from Structured Data on Commons will continue to be critical for our future work, and we are very thankful for your support.

## **Key long-term benefits realized from Structured Data on Commons:**

The new ability to store, edit, and read structured metadata on Commons is a major breakthrough added to the site throughout 2019. It is enabled and extended by multiple factors.

### Wikibase federation

Wikibase is the software behind Wikidata, and now Commons has its own installation of Wikibase as well. Thanks to a new feature known as federation, the Wikibase systems on Wikidata and Commons can now talk to each other and share resources.

We expect federation to be a crucial piece of functionality for the foreseeable future as we implement structured data and Wikibase instances on other Wikimedia projects.

### Increased Wikibase/Wikidata proficiency within WMF

Previously, Wikibase expertise was mostly limited to developers at Wikimedia Deutschland. Since SDC started, we've built an internal team of developers who have gained valuable expertise in Wikidata, Wikibase, and the techniques needed to integrate it all with other Wikimedia projects. This institutional knowledge will prove valuable for years to come.

### Multi-Content Revisions

Another crucial piece of new infrastructure built as part of SDC, Multi-Content Revisions (MCR), allows multiple types of data (wikitext, JSON) to be managed as a separate document (slot) on the same page, without losing the shared history and page-level functionality associated with the main content.

MCR is now used on Commons to store structured data from the WikibaseMediaInfo extension, and it will be important infrastructure in the future as we expand on our metadata capabilities across projects.

### A substantial amount of new structured metadata on Commons

After nearly three years of work on the SDC project, there are currently three distinct types of structured data on Commons:

- Data on file pages, displayed via templates powered by Lua scripts that pull from Wikidata or Wikibase on Commons

- Data on Category pages, displayed via templates powered by Lua scripts that pull from Wikidata
- Data stored per file in MCR slots, accessed and edited via the new WikibaseMediaInfo extension and/or APIs

Although adding structured data to templates via Lua was possible in a limited way before SDC, our program’s commitment to structured data support has accelerated adoption, encouraged volunteer developers to do more integration, and given those same developers new tools and data to work with.

Additionally, the entirely new functionality of the WikibaseMediaInfo extension has added new UI elements to Commons as well as new API capabilities.

Just recently, in November 2019, community member Jarek Tuszyński ([Jarekt](#) on the Wikimedia projects) completed SDC-based Lua updates to the Information template, which is the primary template for approximately 51 million files on Commons.

Jarekt was able to complete this work due to new software support that the SDC team built. The new functionality allows templates powered by Lua scripts to access the new Commons “M ID” data just like Wikidata Q IDs.

Table 2: Number of files or pages by type of structured data as of Dec. 1, 2019:

Type of Structured Data	Number of files or pages (as of Jan 14, 2020)
Number of file pages with structured data via Lua template scripts	<b>8.7 million</b>
Files with structured data via MediaInfo data item (captions or statements)	<b>3.3 million</b>
Total number of pages with structured data	<b>11.4 million</b>

## Tools and APIs for third-party tool builders

SDC has inspired a number of new tools and new APIs, all of which will continue to be important in the years ahead. As noted earlier in this proposal, [ISA](#) is letting contributors add structured data to Commons images.

The Swedish National Heritage Board experimented with tools to enrich metadata about collections and then retrieve that data from Commons for its own collections database.

[Lucas Werkmeister](#) built two key tools using structured data on Commons – one for [batch editing](#), and one for [image annotations](#). The batch editing tool is helping enrich thousands of files with structured descriptive data. [Hay Kranen](#) created a [new version](#) of his Viz Query tool, to work with the Commons SPARQL endpoint. It allows users to build complex queries in a easy-to-understand, visual way.

There's more to do – which we're thrilled about since it means we can integrate structured data even more into the Wikimedia projects. We planned for the future when we began Structured Data on Commons in 2017, and our new project – which we'll undertake from 2020 to 2023 – will take us even closer to our projects' true potential. Thank you again for all your support, and for considering this new proposal about the future of the Wikimedia projects.

## How the Proposer is Qualified to Do This Work

The Wikimedia Foundation has stewarded Wikipedia and the other Wikimedia projects since 2003. Founder Jimmy Wales entrusted the Foundation to lead the projects forward and to find the best staff possible for our global mission. And that's what we've done: Hired key people to every position, including those who've been leading our structured data project. [Amanda Bittaker](#), Program Manager for the Structured Data on Commons program, has years of experience in nonprofit program management, evaluation design, and project implementation. [Ramsey Isler](#), Product Manager for the project, has years of experience as a developer, designer, writer, and technical director. Both Amanda and Ramsey have track records of success that mirror the Wikimedia Foundation's track record of the past 16 years, when we've guided Wikipedia and our other projects into the largest source of shared knowledge in human history.

Users come to our sites [20 billion times a month](#), making Wikipedia the most widely used educational resource on the Internet. Earlier this year, the *Washington Post* [called](#) Wikipedia “the best part of the Internet” and “the Internet's good grown-up,” highlighting its “moral maturity and repeated contributions toward the common good.” Wikimedia Commons is now the world's largest free-to-use library of photos, videos, illustrations, drawings, and music – a site that universities, libraries, museums, and other institutions increasingly rely on to hold their collections and to reach a global audience of hundreds of millions of people. Our projects work because we collaborate with volunteers and institutions around the world, and because

we're staffed by people who have the knowledge and experience necessary to see our projects through from start to finish.

Amanda and Ramsey worked with other Wikimedia staff members and members of the Wikimedia community to transform Structured Data on Commons from a good idea into a practical reality. Also overseeing the project: Top Wikimedia executives who are committed to its continued evolution and expansion. [Toby Negrin](#), the Wikimedia Foundation's Chief Product Officer, brings nearly 20 years of experience with data integration, research, and design. [Grant Ingersoll](#), the Wikimedia Foundation's Chief Technology Officer, has two decades of experience in open source software development and natural language processing engineering, most recently as CTO and co-founder of Lucidworks, which specializes in AI-powered search solutions. [Katherine Maher](#), the Wikimedia Foundation's CEO and Executive Director, greenlighted Structured Data on Commons as one of her first acts as the Foundation's newly appointed head in 2016. The Wikimedia Foundation comes to the Sloan Foundation for support because we have past, successful experience with the project, and because we have a detailed plan to move it forward – just as we've done for the past 16 years with Wikipedia and our other knowledge sources, which have become essential sources of knowledge for people around the world.

## Conclusion: Visualizing Structured Data's Greatest Potential on the Wikimedia Projects

Finding new ways for Wikimedia's community to contribute knowledge and finding new ways for people to *access* that knowledge is the future of Wikimedia. When it comes to search, we need Commons to be more like the rest of the internet: Inviting, easy, and rewarding to use. The reward for Wikimedia users won't be "likes" or "new followers" but a much more precise account of available Commons' media, and a much more precise understanding of how Wikimedia can help them with their searches. Structured Data on Commons has brought our users much closer to that understanding. But we can get closer still – closer to realizing users' expectations of what they should get when they search for something. And closer to our own expectations of what we can do – what we *should* do – for roughly one billion users.

One of the many facets of our work across the Wikimedia projects is constant change. Wikipedia adds thousands of articles a day. Commons adds thousands of new images a day. Every new article and every new image adds to the volume of knowledge that's waiting to be found by Wikimedia users and those across the Internet. We want the wait to be over. This new phase of our structured data project will give users a much greater ability – faster, more efficient, more enjoyable – to access the knowledge they're after. It enables users to contribute to that knowledge more easily. And this new phase will continue the momentum we built in the first phase. That phase led us to this one, where we can see the future of structured data across the Wikimedia projects. We love what we see. That's why we want to make it happen. And with the continued help of the Sloan Foundation, we can do just that. Thank you again for your consideration.

## Appendices:

### Information Products

Structured Data Across Wikimedia will build upon software and databases developed during the Structured Data on Commons project. By building on what already exists, we can quickly develop new and more robust information products that serve multiple use cases that have become apparent during our current work.

#### **New MediaWiki extensions**

The core new software product from the Structured Data on Commons project was the WikibaseMediaInfo extension, which enables the addition of Wikibase structured data to multimedia files stored and displayed within a Media Wiki instance. For Structured Data Across Wikimedia, we will expand on this code to build a new extension that performs a similar function, but for wikitext articles with long-form text content.

The new extension, like most Wikimedia Foundation software, will be open source. Other new extensions may be developed as well.

#### **Topical structured data for articles on Wikimedia Projects**

As Commons begins to receive more structured data, there has been an increased interest in using that data outside of Wikimedia. Researchers, GLAMs, and third-party tech organizations have all begun to look at how this data may assist them.

As we've worked on Machine Vision with Google via the Computer Aided Tagging feature on Commons, staff at Google have mentioned their interest in monitoring how the addition of structured data in the form of "depicts" statements would affect Google search results for those images. We've also recently had preliminary discussions with engineering teams at Reddit as they attempt to use Wikidata as a foundational piece of their in-progress topic modeling systems.

Structured Data Across Wikimedia will add structured conceptual/semantic data to even more types of content and provide useful tools for these interested parties. Additionally, this new structured data can create new opportunities for community-developed software and research that amplifies and extends our work.

## Statistics and data dashboards

As we develop new systems for semantic/conceptual search and analysis of text, we'll create large amounts of data that can be used to develop statistics. Some topics we could develop statistical products for include but are not limited to:

- Changes to search effectiveness via the new work undertaken in this project
- The number of concepts referenced in an average paragraph on a given Wikipedia article
- The number of Wikipedia articles that reference a given topic/concept

## New APIs

Existing SDC APIs allow third parties, community volunteers, and other teams at the Wikimedia Foundation to build tools on top of what we've done. We expect to continue this work with new APIs that expose the new metadata as well.

## Presentations and documentation

As we've developed the Structured Data on Commons project, various Wikimedia staff members have presented our work and findings in public. This has included [podcasts](#) and presentations at Wikimedia events including WikidataCon, [Wikimania](#), [WikiConference North America](#), and Hackathons.

We expect these information sharing and educational efforts to continue with Structured Data Across Wikimedia as we expand beyond Commons and engage other projects and communities.

## Technical Discussion

Architecture Diagrams for future proposed work

# Hypothetical tech stack

