

Structured Data on Commons: A Proposal for the Alfred P. Sloan Foundation

PUBLIC COPY

Budgets and staffing were revised over the course the development of this project and our conversation with the funder. There are items in the proposal that are now out of date, including:

- Page 1 contains an early budget estimate that was later revised to \$2.1 million
- Page 2 contains an outdated mention of the number of staff needed for the project.
- Page 23, Appendix 2 contains a timeline that was later revised and refined. See the document entitled "Structured Data on Commons - Staffing Plan with Tasks - Updated" for the updated timeline presented to the grantor
- Page 26, Appendix 3 contains a budget that exposed individual proposed project staff salaries. That section has been hidden.



Structured Data on Commons: A Proposal for the Alfred P. Sloan Foundation

Prepared by Caitlin Virtue (cvirtue@wikimedia.org) and Wes Moran (wmoran@wikimedia.org)

What We Are Building: The Structured Data on Commons Project will provide a set of infrastructure and tools. These resources will later help our global community of volunteers transform the information throughout Wikimedia Commons' media files from free text into machine-readable data. This switch will enable new uses of Commons' media, from making Wikipedia editors more efficient and articles richer and more dynamic. It will also let new users, like museums, remix the media in their own applications.

Who Will Benefit: The Commons community, [GLAM](#)-style cultural institutions, Wikipedia editors and readers, scientific organizations like Gene wiki, Google, people who believe in free knowledge, the Internet, and CMS platforms like Wordpress and Drupal.

How Long It Will Take: This is a three-year project. In year one, we will build the majority of the back-end infrastructure. In year two, we will complete the back-end infrastructure and build the user-facing (front-end) tools. In parallel we'll also support the efforts of our community of volunteers to use the tools we've built. In year three, we will continue feature evolution and ongoing response to maintenance needs. We will also encourage development of tools with the community and partners that allow for greater engagement and more active cleanup and contribution to metadata. There may be a long tail around continued support of curation for up to 10 years but that is also highly related to community adoption.

How Much It Will Cost: Staff and infrastructure costs over three years would be approximately \$5.5 million.



Resources Required to Make It Happen: Ten staff members across Wikimedia Foundation and Wikimedia Deutschland in Project Management, Engineering, Design and Community Liaison roles.

Outcomes at the End of Year Three: At a minimum, we anticipate 5 million media files will be converted by the end of year three. If bulk editing and/or community adoption is quickly engaged, that number could increase significantly.

Introduction:

In an age where sharing information through social platforms and the Internet as a whole is a standard way of communication, the availability of freely licensed photos, videos, and other media is more important than ever. Wikimedia Commons is the world's largest repository of freely licensed educational media, and our Structured Data on Commons project is a crucial part of Commons' growth and adoption, and the growth of media on the Internet in general. Our project will transform Commons from a useful site to host media to one that integrates that media into the rest of the web by making finding, sharing and reusing this media much more efficient, and much, much easier.

This project will integrate the freely licensed media in Wikimedia Commons with Wikidata, our open knowledge base, and enable our passionate and dedicated community of volunteers to continue to realize the vision of free knowledge for everyone. Over the next three years, [Wikimedia Deutschland](#) (our German chapter) and the WMF will collaborate on the technical work needed to build out the infrastructure to support this project, design and build the tools needed to enable the work, and consult with our community on the best ways to involve their enthusiastic participation in the project. Wikipedia helped usher in the paradigm of "knowledge sharing," and this paradigm is evident on Wikimedia Commons, which grows by about five million media files a year, and will be evident in our Structured Data on Commons project.



The project will be set for active partnership with the community, we believe, by the end of its second year – June 30, 2018. More than half of Commons' media files will be impacted in the short-term, we believe. Magnus Manske, one of the foremost experts in how Commons and Wikidata content is structured, estimates that the Commons Community could instantly migrate between 16-19 million files into a Structured Commons format, as long as they had the right infrastructure (we're providing that) and a fairly standard piece of software (that a volunteer engineer could provide).

Like Wikipedia, Commons has become a "go-to" source on the Internet – used by everyone from casual readers to [major media outlets](#) to higher institutions of learning, and easily discoverable through platforms like Google search. Our project will integrate Commons' media more readily into the rest of the web. Underlying the project is Wikidata, our data repository that debuted in 2012. This proposal explains Wikidata's role in the project, and the way that the Wikimedia Foundation will work closely with Wikimedia Deutschland – just as we did on Wikidata.

From our beginning as a foundation, we have pledged to make the sum of human knowledge freely available to the world. The Structured Data on Commons project continues that vision, continues our work that benefits everyone who has access to the Internet. Thank you for considering funding this project, which we began in 2014. We would be honored to have the support of the Alfred P. Sloan Foundation.

Why the Project Is Needed

All freely licensed photos, audio, and video files on Wikipedia are stored on the Wikimedia Foundation site called Wikimedia Commons. Started in 2004, Commons now has [34 million media files](#) – photos, audio, and video – and it continues to rapidly grow every year: Contributors [added](#) about five million new files last year. Its media files are easily discoverable through platforms like Google search.



Thousands of volunteers integrate these media files from Commons into our Wikimedia projects, like Wikipedia, to illustrate our content and share that media with the public. These files are typically 1) personal photography and media uploaded by individuals; 2) freely licensed media files from locations on the internet like Flickr, YouTube, open access journals, and other repositories; or 3) donations from institutions and organizations with substantial media collections, like UNESCO, NASA, and the British Library.

Wikimedia Commons operates on MediaWiki, the same software that powers Wikipedia. MediaWiki was developed primarily for hosting text like Wikipedia, so Commons' millions of media files don't have machine-readable metadata. Instead, each media file is accompanied by free-form, non-structured descriptions that are not consistently machine-readable. This makes it difficult to link the media there to other educational resources on the Wikimedia sites, like Wikipedia or Wikisource. It also makes it difficult for other websites to take advantage of the media – to link to it and connect the media to their own content. Moreover, unstructured data makes it more difficult for Internet users to search for this media – to find the photos, video, and audio that would be ideal to reuse but are essentially unseen, hidden because the media's details, the words that describe the media, are incomplete and disconnected from the wider Internet.

The Commons volunteer community has long asked for features that require a database structure that would let them describe the media more fully. This includes multilingual categories that would allow more non-English speaking volunteers to effectively tag and find images. (See [here](#) for a 2004 discussion, [here](#) for a 2008 discussion, and [here](#) for more recent discussions.) Structured data is the solution. Files on Wikimedia Commons already have a significant amount of metadata, from museum records to copyright information, but it is not easily accessible in a robust structured and linked format – a format that allows software to both understand what the fields mean (structured), but also connect them to concepts defined by the structured data (linked).



Structured data – and Wikidata – are changing the way people across the Internet can access information from repositories like Wikimedia Commons. One way to think of structured data: It's a kind of DNA that explains information in a much more integral way. Structured data provides meaning for a media file from multiple angles. It offers multiple ways to search for that media, and multiple ways to understand that media. Unstructured data only tells part of that media's history. Unstructured data diminishes the value of information contextualizing the media – and makes it much harder to find the image, audio, or video. For much of Commons' content, it's as if the media files don't exist at all, since they're so difficult to find on the Internet, and difficult to find on Commons. We're changing that. (For another explanation of structured data, see Appendix 5 of this proposal.)

The Structured Data on Commons project will provide the means to transform the information throughout Wikimedia Commons' media files from free text into machine readable data, so that Commons' millions of media files are much easier to view, translate, find, edit, curate, use, and reuse.

By implementing our Structured Data on Commons project, and expanding Wikidata's technical capabilities to Commons, we will:

- Make searching for media files much more effective. Searching is currently hard because the information on every Commons media file is unavailable in a machine-readable and machine-understandable form.
- Allow automated license-compliant reuse of media files.
- Give developers the ability to write flexible software for accessing the information and use it in new ways. This would grow access and use for the content by those people who already use Commons – and a wider community of re-users that need a more consistent and accessible dataset.



- Provide infrastructure that encourages cultural and scientific institutions to release their media alongside metadata under a free license on Wikimedia Commons rather than corporate repositories like Flickr, especially in contexts unsupported by aggregators like [Europeana](#) and the [Digital Public Library of America](#) (DPLA).
- Let users edit file information much more easily, without having to both know [MediaWiki markup](#) and the nuances of its implementation on Commons. This would create a more “form-like” user experience, similar to other popular media-sharing platforms.
- Allow users to view the information about the media file in any language. As it is, most information – including reuse restrictions and other licensing information – is only available in English.

With these changes, we expect Wikimedia Commons to become more interoperable with other media-sharing platforms, and a better source of media for content management systems like [Wordpress](#), social media like [History Pin](#), and digital educational tools more broadly. With this increased capacity for reuse, we expect Wikimedia Commons to become a more broadly appealing platform for sharing media files, especially among educational, cultural, and heritage organizations that share our mission for educational content and with whom our volunteer communities already have relationships.

Wikidata, our structured data repository that debuted in 2012, makes this all possible and is a foundation of the project. Wikidata revolves around “items” that are the equivalent of Wikipedia articles, and each item contains “statements” about everything from names to book titles and periods of history. Wikidata makes these items and statements much easier to centralize, link, and share across all Wikimedia sites. Wikidata already has 23 million items and 100 million statements, and Wikidata’s growth is central to our free knowledge ecosystem’s long-term growth – and is a key foundation of the Internet’s overall expansion of linked and structured data. Because the data on Wikidata is in structured form, Wikime-



dia projects and third parties can easily reuse the data, and computers can easily process and “understand” it. Wikidata is increasingly becoming a powerful tool [not only in STEM fields](#), where structured data has a long tradition of use, but also in other fields like cultural heritage and the humanities where fewer tools are available for linking data across the Internet. Data entered in any language on Wikidata can immediately be made available in all other languages on Wikidata, instantly making Commons’ content accessible to everybody, regardless of language.

We see our Structured Data on Commons project as an evolution of Wikidata – a natural expansion of a project that we funded with help from the Allen Institute for Artificial Intelligence, the Gordon and Betty Moore Foundation, and Google, Inc. These three institutions saw the need for Wikidata, and its long-term benefits, and that vision of a more connected Internet is now coming closer to fruition in Wikidata's fourth year. Extending its impact to Wikimedia Commons is a vital part of that process.

We began the Structured Data project in August 2014, and we continued discovery and testing through 2015 to understand the needs and complexities involved. We [involved community input](#) from the beginning. Our plan is to continue working with the active community of Commons' contributors and editors, the Wikidata community, and the broader community of re-users of Commons files, to build new tools and develop practices for structured data. The community will then, over time, gradually migrate unstructured data into a machine-readable format. As with Wikipedia, volunteer contributors are essential to the growth of Wikidata and Wikimedia Commons, and all implementation of these efforts will also include community outreach and integration with existing community workflows and tools.

The Wikimedia Foundation is collaborating on this project with Wikimedia Deutschland, which oversaw Wikidata's initial operations, and continues to manage Wikidata's broad



technical and engineering functions. Our Structured Data project will leverage the experiences of staff that have worked well together for many years, and worked to create Wikidata in 2012.

Commons' Growth as a Project

In the past five years, Commons has been growing as fast as Wikipedia has – faster, even. People around the world are uploading images – on their own volition and through contests like [Wiki Loves Monuments](#), the world's largest photo competition. Museums around the world, as well as institutions like NASA, are donating large troves of historic photographs to Commons.

The Wikimedia community's GLAM-Wiki outreach program is building relationships with cultural heritage organizations, which is helping them share many different types of cultural heritage media on Commons. Digital infrastructure built by cultural heritage organizations prioritizes structured metadata for media files, and this feature allows organizations like DPLA, [Trove](#) and Europeana to aggregate both the media files and the associated metadata into central discovery tools.

Wikimedia could be a neutral and openly-licensed shared platform in this ecosystem, providing a platform for organizations outside the scope or support systems of DPLA and Europeana. Connecting Commons with Wikidata through structured data would allow media files to be connected to Wikidata concepts represented in Wikipedia and to search tools, thus making the Commons repository more accessible to Wikipedia's half billion readers. Without this infrastructure, we cannot build an API that returns consistent and meaningful data to partners who want to use Commons content in their own contexts.

While the Wikimedia volunteer community manually enriches the textual data about the files on Commons, it is nearly impossible to extract that information to use it for search and



discovery tools or to provide this enriched metadata and multi-lingual translations back to the institutional platforms.

The Wikimedia volunteer community has started integrating metadata from GLAM institutions into Wikidata (like the [Sum of All Paintings project](#), and the Flemish Art Museum pilot). But the media files representing the actual objects on Commons cannot be fully integrated with that data format.

Structured data is a requirement for this next stage of development and will accelerate Commons' value to educational, cultural and scientific institutions. Additionally, it will enhance tools for Readers and Editors that the Wikimedia Foundation introduced in earlier years. A good example is MediaViewer, which we first introduced in 2013 and brought out for wider use in 2014 (https://www.mediawiki.org/wiki/Multimedia/Media_Viewler). The tool improved the way users could display images, and the way that users could share the images, and see the images' metadata – but it could only do so much without having proper structured data, and communities did not widely embrace the tool. Another example: Our Visual Editor has difficulty surfacing high-quality images because of the templates, categories and free text that Commons employs. With structured data, Visual Editor would be able to more easily provide editors the best media for illustrating our projects. With that data, users of Visual Editor could do a simple query to get high-quality images on a certain topic with a certain license. Computers can't deal very well with free text. Structured data on Commons unlocks the ability to do many other projects.

What We'll Accomplish After Three Years

After three years, the infrastructure will be in place to migrate Commons' millions of media files to structured data. A sizable amount of the images are ready for migration. As noted in the introduction, Magnus Manske, an expert in how Commons and Wikidata content is structured, estimates that the Commons community could instantly migrate between



16-19 million files into a Structured Commons format if they had the right infrastructure and a fairly standard piece of software.

How soon after this initial migration would the other files be converted? Most challenges for the tools to do this work, Magnus estimates, would be related to solvable template/data features of existing Commons content. The “X-factor” is the capacity of volunteers and bot tools to migrate the material. Whenever you migrate a large set of material from one data structure (semi-structured, historical Commons data stored in wikitext) to another (structured data on Commons), there is a long tail of data-cleaning where, for a sizable window of time, a large portion of the content stored within the file will have a mix of structured data that’s easy to clean and integrate, and unstructured (or “uncleaned”) data that requires clean-up by a combination of semi-automated tools and human evaluation to be accurately read by computers.

It may take five or even 10 years for the majority of Commons’ media files to be moved over into a Structured Data format. We can’t offer an exact timeframe since it depends on the input of the Commons’ community. But we are optimistic. Every month, more than 7,000 Commons contributors make five or more edits on Commons (see <https://stats.wikimedia.org/wikispecial/EN/TablesWikipediaCOMMONS.htm>), and every month more than 1,300 Commons contributors make 100 or more edits on Commons. Both Commons and Wikidata [have](#) about 7,000 active editors. Whether on Commons, Wikipedia, or another Wikimedia project, our sites are populated with passionate contributors. A key is designing software, tools, community support, and incentives for contributors to participate in the Structured Data on Commons project. That’s what we’re doing.

Once the infrastructure is in place, we anticipate that the Commons community will have capacity for migrating over for certain data types, including [“Categories.”](#) Almost every image on Commons is placed in content-based categories, which are as close as Commons



comes to tags. (See image at right.) This information is ripe for synchronization with Wikidata identifiers.

More than [1.5 million \(of the 4.2 million\) of these categories](#) have already been synchronized with Wikidata items, which make them prime candidates for this migration. Many of these categories also have subcategories, which could be examined systematically with the right tools. Additional data items on almost every file could more readily be migrated:

- Copyright statements
- The majority of “creators of works” fields (some of which is dependent on major technical infrastructure planned for this grant)
- The majority of upload date fields
- The majority of creation date fields
- The identities of uploaders of files (which requires major technical infrastructure changes planned for in this grant)

Subsections of Commons content that, with higher confidence, can be migrated, in large part, to structured data, include:

- More than 31 million media files using a version of the [Information Template](#) that provides some basic information that can be moved into structured data, some of which will need cleanup for consistency to create uniform data. Included among these templates, are several types of data ripe for migration:





- Of these, more than [24 million language-identified file descriptions](#) (captions), which are encoded correctly for correct integration into structured data.
- Of these images, 6.2 million files have location geo-coordinates that could accurately describe the picture's location and be integrated into structured data

Other subsections of Commons that have been curated with more robust description metadata could be migrated more quickly by communities. And these subsections would be the most useful parts of the initial migration. They include¹:

- [905,000 pieces of Artwork](#)
- [637,353 books](#)
- [754,324 archival quality descriptions of photos](#)
- [87,883](#) Airplane images
- [18,326 maps](#)

An important consideration: The Commons community has long demanded multilingual categories that allow more non-English speaking volunteers to effectively tag images. Structured data on commons allows this to happen.

How We'll Ensure That New Commons Files Have Structured Data

The Wikimedia Foundation's [UploadWizard](#), [cross-wiki upload](#), core upload, and related upload campaign tools maintained by the Foundation, and mass upload tools being developed by the volunteer GLAM-Wiki Community (principally [Patty Pan](#) and [GLAMPipe](#)) will need adjustments, to be compatible with structured data of Commons. These tools are all built around the assumptions that media should have structured data and descriptions, so

¹ These numbers don't indicate all files of that genre just the ones curated with data in the best standard



it's a matter of adjusting the data pathways through our API/framework and expanding the fields available in the upload forms.

Impact on the Wikimedia Movement

By bringing structured data to Commons, we will benefit the Wikimedia movement in five ways:

1. Categories and metadata would be more easily multilingual. This allows for greater interoperability and collaboration between volunteers from multiple language communities. It also allows for multilingual search and discovery. This has been a long-term ask from the Commons Community, becoming one of the most supported items in a Community Generated Wishlist of technical features. See https://meta.wikimedia.org/wiki/2015_Community_Wishlist_Survey/Commons#Allow_categories_in_Commons_in_all_languages
2. Developers will expand the project even more. Structured data on Commons provides the fundamental infrastructure required for consistent use of Commons data through APIs and other machine readable endpoints, so that developers both within and from outside the Wikimedia Community can create consistent, reusable, and reliable software that edits, helps with reuse, and allows analysis of Commons media and its associated data. Current tools that try to do these rely heavily on short-term software solutions that consistently break and/or produce bad end-user data when the MediaWiki core changes or the volunteer community changes wikitext.
3. With better Commons search capability, contributors can more effectively illustrate Wikimedia content, especially for abstract or unconventional topics. The current process depends on a knowledge of English, a knowledge of how to use the Commons Category system, or a knowledge of a very specific terms used by uploaders of the media file.



4. A better Commons search allows for better user experience for the broader reader community. With structured data, they'll be able to find the right media, when they want to – which will encourage greater sharing outside Wikimedia projects.

5. It will allow for easier and simpler partnership with content providers, especially knowledge collecting/sharing organizations. Until recently, with the grantee development of Patten, mass upload of well-described content to Commons was very technically complicated. Uploading without that contextual data, or oversimplifying well-developed metadata, created a loss in quality of the image descriptions. Patten reduced that complication in half. Structured data on Commons further reduces the intellectual and technical hurdles needed to upload en-mass – which will prompt more free content into our platform for sharing.

The Immediate Known Benefit to Other Organizations

DPLA, Europeana, and a number of community partner organizations want to integrate clearly open-license material and its metadata into Wikimedia Commons, and by extension Wikipedia. Because of these organizations' high profile, it's very likely that other GLAM partners would follow soon after. Volunteer developers will also be able to more simply upload already openly licensed content from both GLAM and other kinds of open knowledge platforms (Open Access journal articles, images from organizations like NASA or New York Public Library with open licenses, openly licensed video from YouTube, etc).

GLAM and other partners who have donated media will be more readily able to review and integrate improvements and translations to file metadata created by Commons volunteers to their own catalogues and digital libraries. This back-and-forth improvement – made by both Commons contributors and the cultural institutions themselves – improves the metadata, and the possibility for discovering these images, not only on Commons but on the Internet in general.



Improved search and API functionality for Open Content allies like Creative Commons and Internet Archive. Furthermore, there is demand and need for greater access to Commons by reusers; WMF has, for instance, talked to the developers of USC's Scalar platform about embedding Wikimedia content in academic websites, but the lack of strong machine readable metadata prevents a full utilization of our content.

Increase In Sharing via Commons

The benefits to other organizations detailed above have the potential to create a virtuous circle of sharing, use and re-use around Wikimedia Commons. We believe that our project will result in a significant increase in the number of cultural and scientific groups/institutions sharing their images via Commons, and extend the vision that we – and independent experts – have for our sites. Knowledge-access advocates like [Benjamin Good](#), an Assistant Professor of the Department of Molecular and Experimental Medicine at the Scripps Research Institute, are [pushing](#) academic communities to integrate their data onto Wikidata – just like more cultural institutions like the British Museum have recognized how important it is to integrate their media into Wikimedia Commons. Instead of being siloed away behind paywalls and subscriptions, their work becomes part of our freely available global source of knowledge.

One of the big reasons why many institutions are still reluctant to spend significant amounts of time understanding how to bulk upload content is that it needs to be integrated into Commons' particular form of storing information. Structured data can integrate the parts of digital collections that institutions invest money and expertise on when provided alongside clear tools for uploading; such capabilities allow far more community members to be liaisons for these organizations, and for organizations to participate more readily on their own.



Right now, it's hard to find good media on Commons and reuse them, further discouraging institutions from uploading their files to Commons. Once we make it easier to find high-quality content and to reuse it, institutions have a much greater incentive to add media. Now, the main incentive for institutions to upload to Commons is the volume of Wikipedia pageviews from pages that contain their media files. By improving Commons itself, and expanding the way people can search for images and reuse these images, we greatly expand the usefulness of Commons – and the incentive for partners to upload images there. It starts with structured data.

The Longer-Term Benefit to Other Organizations

A number of organizations do not have support from DPLA or Europeana, or other similar aggregators of educational media, especially in regions like South and Southeast Asia, Latin America and Africa. Moreover, many organizations in these contexts do not have the technical capacity for hosting their own digitized collections. With structured data improvements, Wikimedia Commons could provide a better platform for digital records, and thus we could become a reliable venue for sharing cultural heritage content under free formats and free licenses in these under supported contexts.

Structured data on Commons allows greater work on embedding and dynamically reusing Wikimedia Commons content with proper attribution. This would allow more flexible embed tools for a wider range of content management systems and platforms, would make it easier for adoption of Wikimedia content to usage, while also complying with our licensing.

Our vocabulary for describing media files will be concepts described in Wikidata and synchronized with major identifiers. Creating such a large controlled, yet still dynamically evolving, vocabulary for tagging could advance Wikidata's usefulness for tagging multimedia in almost every context, whether or not the media content enters our ecosystem. This would



allow for cross-internet discovery of relationships between media files, again creating those foundations for a Semantic web that Wikidata is so important for.

Content brought into Wikimedia Commons will be able to be more faithfully and more consistently archived by Internet Archive and other digital archiving services, allowing for better discoverability of that content in digital records, even if Wikimedia projects disappear.

Metadata: How Our Project Compares to Flickr

To put our Commons project in context, it's helpful to compare Commons to another media-file repository that has widespread impact on the Internet: Flickr. While Flickr is a commercial site that is owned by Yahoo, it adopted the use of Creative Commons licenses in 2004 – and more liberal CC licenses [in 2015](#) – and Flickr is now the Internet's largest repository of Creative Commons-licensed photos. Flickr has 10 billion overall images.

Flickr's upload tool allows for tagging, licensing, access rights, people identification, and grouping with a form that includes auto-fill and drop-down menus. Though Flickr has a limited number of metadata fields, its tools' flexibility and easy user interface does allow for consistent, emergent, and widespread tagging across Flickr. However, Flickr's data still doesn't have the level of specificity that structured data allows. Flickr allows tagging with words like "painting," "dog," and "book." This could mean the image is a painting of a dog and a book, or even a painting of a book by a dog. Flickr's metadata can't distinguish between the potential meaning of these tags. We will be able to do that. And Flickr can't distinguish the tag "gift" that means "a present" in English and the tag "gift" that is the German word for "poison." We will be able to do that. Structured data can distinguish between different concepts with the same spelling. Structured data can distinguish the meaning of words, even across language barriers. Structured data is much, much smarter than unstructured data.



Flickr’s method of people identification, copyright licensing, and access rights are similar to what Commons would be if it had structured data. In the digital library/catalogue community, similar form upload tools – with more robust, semantic metadata structures – are available for structured data, usually designed around specific applications for that institution and utilizing some variation of Dublin Core Metadata (<http://dublincore.org/metadata-basics/>) or other metadata and identifier standards. Our current upload wizard (https://commons.wikimedia.org/wiki/Commons:Upload_Wizard) has some of these features, but we expect our contributors to fill in many more fields than is typical of community-generated sites. And these fields – including creator and medium – are not readily “suggestable” to the uploader. Already, the Wikibase/Wikidata software provides this kind of suggested fill-in, with the databases’ index of labels.

Others That Are Trying to Solve the Metadata Problem

Other institutions are trying to expand their crowdsourced metadata capabilities, but not at the scale that we are. Because adding metadata and image description is so person-intensive, a number of institutions with large troves of media are trying to defer that work to “crowds” using structured tools. Two examples: Zooniverse (www.zooniverse.org/) and Biodiversity Heritage Library (www.biodiversitylibrary.org/). Both organizations are asking people to do simple tasks that provide metadata tagging and improve descriptions. Another example: the National Archive of the United States is experimenting with its Citizen Archivist program to enrich their collections:<https://www.archives.gov/citizen-archivist/tag/>. The Archives encourages volunteers with this language on its “citizen archivist” home page: “Tagging is a fun and easy way for you to help make National Archives records found more easily online. By adding keywords, terms, and labels to a record, you can do your part to help the next person discover that record.”

For the most part, these platforms are designed around certain subsets of data being crowd-curated – typically tags or descriptions, or transcription of the media content itself.



(Most of the descriptions are not as extensive or extendable the way they would be on Commons after our project is implemented.) These platforms also err on the side of caution by building systems that ask for and produce a relatively small number of volunteer-driven metadata types.

A number of smaller organizations (researchers, research libraries, etc.) are trying to do crowdsourcing for metadata and other record enrichment work (transcription, caption writing, etc.), but they rarely achieve much activity because they require a critical mass of interested participants, the right topics for engagement, the right model for engagement, or a long-term committed editorial group.

Wikimedia's massively open review structure, and the demonstrated community strategies for monitoring contributions developed by both our Commons Community and Wikidata communities, allows for a radically more open editorial space – where almost everyone can contribute to all layers of most metadata while still maintaining quality. And unlike these other collaborative crowd-sourcing projects, we are the only community that currently aspires to do this at-scale while representing the metadata in hundreds of potential languages.

The scale of our community – its reach, its passion, its impact, its potential – is evident on Wikidata and Commons. Contributors make [10 million edits a month](#) on Wikidata and [two million edits a month](#) on Commons. As noted on Page 11, Wikidata and Commons each have about 7,000 active editors – people who make five or more edits a month. More than 1,000 people make more than 1,000 edits every month on Wikidata and Commons.

The Structured Data on Commons project is a game-changer. Through this project, our content will also be shared more widely outside Wikimedia. The more that Wikimedia's knowledge is shared around the Internet, the better we fulfill our mission to reach the most people possible. Fifteen years ago, Wikipedia shepherded in a novel approach: We can all



contribute to the understanding of the world – together, in a community that connects people around the world, and in a way that benefits everyone who has access to the Internet. Our Structured Data on Commons project is a continuation of this approach. We believe the time is ideal for this project. We believe the time is now.



Appendix 1: Risks of the Project

Our plan is not without risks – risks that we believe are surmountable, but risks that we have to anticipate nevertheless:

Risk 1 (High): The community discussion for adoption of structured data on Commons could identify large unanticipated technical changes that require extended development time.

Risk 2 (Medium): Community adoption of features into existing workflows becomes obstructed by community politics of some sort. The Wikidata team has demonstrated effective adoption of Wikidata features both in Wikidata and the broader use of Wikidata in other projects, including Commons and dozens of language Wikipedias. Applying lessons learned of transparent planning and community management in the Wikidata context to Structured Data on Commons will be a necessary part of its success. Additionally, previous community polls and discussions have shown widespread enthusiasm in the features possible from these changes to Commons.

Risk 3 (Medium): The backlog of tools and software features that break during the infrastructure change impede effective integration of all of the new features planned for structured data on Commons. It's crucial to have support from and prior planning across Wikimedia development teams, including Discovery, Multimedia, Reading, Technical Operations, and Architecture among others.

Risk 4 (Low): Relationships between WMF and/or WMDE and the Commons community change in some fundamental way that makes community members skeptical of the initiative. With the current community dynamics, this is unlikely. The positive rollout experience of Wikidata, the change in WMF Executive Director, and the upcoming community strategy process all have been well received and indicate a period of warming relations. The Wiki-



data team has deep experience managing open, transparent, and community inclusive processes for major technical rollouts despite a complicated context.

Risk 5 (Low): Complexities in feature development and roadmaps clearly defining work between Wikidata as currently funded by WMF and future work.

Risk 6 (High): We are going to make significant changes to pretty much everything in Commons. This will be very disruptive to re-users of our content and builders of community and third-party tools. To some degree we need to mitigate this by building migration paths or compatibility layers, but there will be a painful breakage, which will include a long period of engaging with and supporting change in software, tools and processes supported by our Communities and third parties .



Appendix 2: Timeline

The bulk of the infrastructure work to make this happen would occur in the first year as defined below. And then the cleanup and additional supporting efforts for curation, planned and unplanned features, including front-end work to support structured data input, occur in the second year. There may be a long tail around continued support of curation for many years but that is also highly related to community adoption.

Year 1 Infrastructure	Year 2 Integration	Year 3 Engagement
People [1]: Build Structured Data team		
Tools [2]: Exploration & Testing	Tools [2]: Deployment	Tools [2]: Iteration and maintenance
	Media Viewer Licensing	
Search & Query: Explore features	Search & Query: Deployment	Search & Query: Iteration and maintenance
Community: Review concepts & data models	Community: Feature launch feedback	Community: Feedback and curation support
	Community: Support migration of tools, especially other community media import tools, like PattyPan and	Community: Encourage tool development for engagement, cleanup and contribution
Partners [3]: Review data models	Partners [3]: Promote features and seek new potential partners	Partners [3]: Encourage tool development for engagement, cleanup and contribution

[1] Product Managers, Project Manager, Dev/Engineers, UX Design/Research, Community Liaison

[2] Current tools like MediaViewer, Upload Wizard, Visual Editor

[3] Partners include groups like Europeana, Internet Archive, DPLA, other GLAM, and Open Access content from Journals, NASA, New York Public Library, YouTube, etc.

Year 1 - FY 16-17

- Groundwork for Structured Data Project
 - Hire and build team and collaboration channels.
 - The end result will be a good technical foundation to store metadata related to multimedia files in a structured and machine-readable way. (This will allow many future innovations like multilingual search and display of file



- information, easier license-compliant re-use of our content, and a richer multimedia experience on Wikipedia).
- Allow Wikidata concepts to be used as a vocabulary for describing media
- Groundwork for integration into existing tools (e.g., Search, MediaViewer, Upload Wizard, Visual Editor)
- User Interface Concept & Design (for multimedia use cases)
 - Integration of existing structure of the file description page
- Community Engagement
 - Demonstrate features of structured data on Commons to Commons community, seek consensus and initial community adoption of structured data for some data types
 - Engage known partners and allies, such as Europeana, Internet Archive and DPLA, in initial evaluation and engagement with data models, APIs and other features of change
 - Continuous close communication with the Commons and Wikidata community on the progress of the project to involve them along the way.

Year 2 - FY 17-18

- Integration into existing tools (e.g., MediaViewer, Upload Wizard, Visual Editor)
 - Easy language independent tagging so that contributors can better engage with content in languages they are comfortable with
 - Automatic import of license information and other data (e.g., from sites like Flickr, YouTube, Europeana or DPLA)
 - More accurate license information in the MediaViewer
- Search & Query Services
 - Mediafiles can be queried and searched by various criteria independent of language (e.g. topic, person shown, rights holder, license type, resolution etc.)
- User Interface Implementation (for Multimedia Use Cases)
 - Integration of existing structure of the file description page
- Community engagement



- Grow depth of Commons community-led integration and use of structured data features as they become available, and responsively support community feedback and concerns
- Actively seek external allies and partners to utilize new features available through structured data on Commons in both the cultural heritage and among other communities actively using structured data
- Support migration of community built tools, like PattyPan and other media import tools, from current Commons structure to Structured data

Year 3 - FY 18-19

- Support for existing tools (e.g. MediaViewer, Upload Wizard, Visual Editor)
 - Continued feature evolution and ongoing response to maintenance needs
- Search & Query Services
 - Continued feature evolution and ongoing response to maintenance needs
- Community engagement
 - Continued support for feedback
 - Encourage development of tools that allow for greater engagement and more active cleanup and contribution to metadata.



Appendix 3: Budget

[Removed]

Appendix 4: How Other Groups Would Use Our Technology

With this project, other groups will be able to use our technology to add structured data to their images in their own repositories. Consider the core technology for this project: Wikibase, which is the software that drives Wikidata, will be used to extend Mediawiki to create a robust media repository. This action makes Mediawiki an attractive Content Management System for media repositories, especially those wishing to invite open collaboration alongside robust contextual metadata. Current open source software in this space tends to have been developed for archival/library/authoritative use cases (Omeka, or DSpace) which favour robust metadata, but also are designed around hierarchical knowl-



edge production (mimicking the expectations of academic and cultural professions); or falls into the radically technically open spaces using Project Hydra (academic focused), Django or Wordpress for example), which require expertise in designing strong metadata storage and collaborative contribution. None of this second class of frameworks wants to do structured data for media out of the box, so deployers of the software need a lot of developer time and design expertise. Potentially, Mediawiki with the features for Structured data on Commons could fill this gap between robust structure and flexibility.

In terms of social technology: We already have some use of the Wikidata descriptors, to tag archival collections (see <http://wikimedia.fi/2016/04/15/yle-3-wikidata/>), and other organizations are adopting Wikidata for other uses. So bringing structured data to Commons is about ensuring our own ecosystem is keeping up with the rest of the digital knowledge and cultural heritage community. In 5, 10 or maybe 15 years, if we do the migration of Commons structure well and the technology works well, other communities would adopt either our strategy for metadata (relying on a hybrid of Wikidata identifiers and our other Commons metadata fields), or very similar schemas/technologies for storing the data that can be interoperable. We are such a big player in this space, and have so much flexibility, that our communities are able to create dynamically usable strategies that can become standards for other communities. Commons' copyright/licensing schema alone is going to be very valuable once it's provided as structured data. We have one of the more globally comprehensive libraries of open-license options, with several thousand distinct license statements.

Appendix 5: A Birthday Photo Explains Structured Data

Here's one way to see the difference between structured data and unstructured data: When people upload content to a social media site or blog, they share their information by using wording that's essentially contextual data in an unstructured format – things like image descriptions and tagging people in photos. A host of a birthday party might take pho-

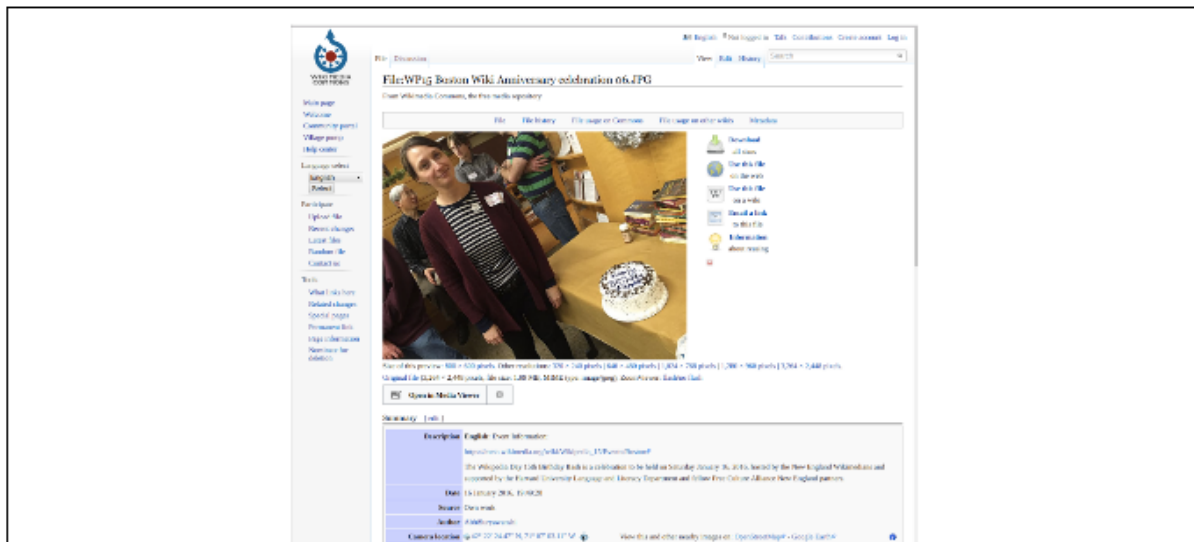


tos of people and birthday cake, upload the images to a website, but only describe the people in the photos and not the cake. Since “cake” is missing from the images’ description, anyone searching for “cake” won’t find the photos. Other people can rediscover the images, but their discovery is entirely dependent on understanding the context of the media beforehand.

The vast majority of people who search for cultural heritage and educational media haven’t seen the photos they want. They haven’t seen the cake, as it were. They don’t even know it exists. So it’s crucial to make sure that – from the beginning – the researcher can find the item from a variety of different directions. Not only do we need to know that the birthday cake is in the image, but we need to distinguish between works simply taken by someone, depicting a specific person, or depicting people at an event hosted by that person. A computer can’t distinguish between the “Joe Smith” in the Facebook-description-like string that reads “Alice and Sue eating cake at Joe Smith’s birthday party” from within a descriptions that reads “Joe Smith blowing out candles on his cake.” Google Image search almost exclusively relies on the first kind of data even though it could support some degree of structured data, because it’s the most common way that people describe media on the web. This leads to a lot of “false positive” results and frequently misses important material for media searches, because people fail to describe all of the elements in their media files. This is one reason that Google and other companies are spending so much money developing machine learning for object/content presence in images.

Structured data allows us to label the meaning of each part of a description. In structured data, we could write the previous example, “Alice and Sue eating cake at Joe Smith’s birthday party,” as “Image1 Subject = Alice;Sue. Image1 Event = Joe Smith’s birthday party. Image1 Depicts = eating; birthday cake; two people.” Wording for the second image, “Joe Smith blowing out candles on his cake,” could be written as “Image2 Subject = Joe Smith. Image2 Event = Joe Smith’s birthday party. Image2 Depicts = birthday candle;

birthday cake; one person.” Now, with this structured data, a computer can help us narrow down searches, to get, say, only pictures of Joe vs. pictures of his cake. And we can

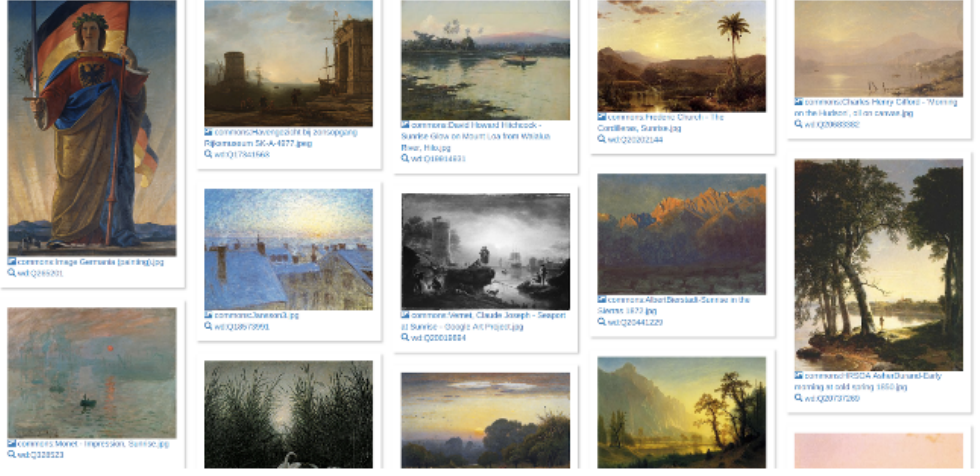


A picture of a Wikipedia 15 birthday cake at a gathering in Boston for the 15th anniversary in Boston. Though the textual description of the image contains lots of information about the event, the lack of structure to this data makes it of limited use for search and discovery.

also do things like sort the images by number of subjects or by the event.

This becomes extremely important when we start thinking about educational media and making it relevant towards a broader public. For example, with structured data you could find all oil paintings in Wikimedia Commons that depict both people and birthday cakes, and those images could be used to generate DIY cultural objects, like online birthday cards without fear of violating copyright. Or, an educator, researcher or student could ask “What photographs and oil paintings do we have representing the town of Rheims published during the 1890s?” Moreover, the results of this intersection of content could be embedded in a Wikipedia article or external website that dynamically updates, as Commons content in that topic grows and volunteers improve the other metadata. Already, we can create a sim-

ilar search result as an outcome of the volunteer-led [Sum of All Paintings project on Wikidata](#). For example, we can ask “Which paintings depict sunrises and have images on Wikimedia Commons?”: <http://tinyurl.com/j34fkcz> . However, Wikidata only describes distinct concepts (such as a painting or sculpture) and does not have a scope that allows describing all of the digital media files on Commons.



The results of a query of Wikidata asking “Which paintings depict sunrises and have images on Wikimedia Commons?” Wikidata provides metadata for distinct items like paintings, including a single image that represents that item; however, without structured data on Commons we cannot perform queries like “Which photographs on Commons include sunsets?” Describing each photograph is outside of the scope defined by the Wikidata community for its project.

Structured data emancipates media files. But structured data is also more work. It takes more time to write more than a basic description of what’s depicted. And the more detailed the data, the more work that’s needed. Most institutions that care for educational material spend a lot of money on creating metadata and ensuring that it is fairly accurate, so that experts, researchers, and the public can find the right bit when they need to.



The main work for educational websites like Europeana (which is the European Union's site for storing and accessing digital works of cultural heritage), [Trove](#) (a project of the National Library of Australia), and DPLA is to aggregate metadata and expose search layers on top of that metadata. Individual institutions have been doing the initial work of developing structured, detailed contextual metadata, but they haven't had the tools to show their collection in relationship to another organization's collection. So, all that money spent digitizing the images is of limited use until aggregators help make sure that it's available for audiences that need the images. Google's search algorithm doesn't incorporate all of the metadata that libraries and other institutions use, so developing Commons as a platform that can both store strong metadata and expose described media content to Google makes us a more valuable ally for that sector. The Structured Data on Commons project is the kind of project that libraries, museums, and other institutions have been waiting for.