Response to Feedback from Sloan Foundation - Structured Data on Commons

In an attempt to respond to as many questions as possible, we've grouped our response into three major themes: Philosophy, Product, and Program. We have linked to additional information, where appropriate, and also include supporting material in an accompanying Appendix as well as an update to the Staffing Plan.

Philosophy

Crowdsourcing versus Machine Learning

Two reviewers asked why our proposal focuses on crowdsourcing rather than machine learning for generating structured data. It is our belief that crowdsourcing and machine learning are complementary, and we support the utilization of both. This project will help facilitate both types of input by providing a clear and reliable means of associating structured data with media on Commons. We anticipate that the community will utilize machine learning tools (such as the Google Vision API) to help create metadata. The community has already been using machine learning for vandalism detection on Wikimedia projects for more than 10 years. They have also experimented with using image recognition tools on Commons, but the lack of structured data support has prevented such projects from scaling.

Second, we would like to emphasize the importance of data curation by humans, as we believe this is a key reason why our projects are successful and have a high standard of quality. Wikimedia Commons spans an incredibly diverse array of subject matter and media. The complexity of this content will require countless discussions, debates, and judgement calls about how best to organize, label, qualify, disambiguate, and translate associated data. This is work our community has already proven it can do, even at enormous scale, and we believe that this model is well-suited to the next iteration of our projects. By combining the efforts of bots and humans within a structured data framework, we hope to build a media data collection of unprecedented scope and quality.

Product (includes product management, engineering, design)

Community created ontologies

Our openness is a strength. Our community wants to be involved in decision-making and has already-created schemas for different types of media in the current semi-structured model. As we have learned from the development of the Wikidata Community, frequently our community members will turn to and iterate on data structures established by expert communities; however the Wikimedia projects solve problems for more diverse applications than most expert-defined ontologies anticipate, so the flexibility and openness becomes an asset to accommodate for changing needs. For example, our WikiCite Community is both working with and diverging from some of the assumptions created by Bibliographic metadata communities in order to accommodate both citations in Wikimedia projects and the needs of expert communities. As the structured data infrastructure becomes available, our community will help define the ontologies for this project in a flexible, consensus-based, and continually evolving manner. As with the data population, partnering with the community will produce inclusive data models, and encouraging experts to participate in those conversations will make sure that existing data models are accounted for in this work.

Tools and Commons UX and design

There are also some concerns over the user experience across the project and how we plan to address these challenges. Over the past year, the Foundation has been using a process called user-centered design, which uses extensive design research, prototyping and iteration to ensure that our various experiences map to the mental models of our users and address their needs.

We've employed this process successfully with the mobile (iOS and Android) apps, both of which have won awards in the past year. The New Readers project, after identifying a significant gap in our global audience, collaborated with an outside firm to send user researchers to 4 countries to "meet users where they are" and assemble detailed information about how knowledge is consumed in these cultures. This data is now being used to prototype potential experiences which will be tested in these countries. Finally, the copyright violation detection tool and the UX for the cross-wiki watchlists were iteratively prototyped with the community, producing evolutionary designs that exposed new functionality in co-operation with existing users.

We are planning on leveraging elements of the above three programs to work with our existing and future users in an empathetic and iterative way to ensure that the UX for structured data is functional and understood by our users. We have not begun this process, so unfortunately we do not have designs to share as we plan to start with research. Since we are a transparent organization, we will of course share designs and feedback throughout the project.

There is one clear distinction that needs to be made around the use cases of the various interfaces -- we intend that Search (both HTML and API) be the primary interface to the new functionality enabled by structured data. Wikidata Query Service is considered the expert interface, and while it will support queries of the structured data, we do not expect it to have general use -- rather it will improve developer and third party interfaces with Commons. In line with this plan, the basic UX will not change dramatically; we need to make the structured data easy to read, understand and edit, but further changes to Commons are out of scope of this grant, though may be included in future work on the project.

Effects on contribution and discovery

Finally, we wanted to highlight how this project provides value in both discovery and contribution. With structured data, part of the metadata we store can now be assessment data, like "valued image" and "featured image" – the same way that Wikipedia articles are tagged with such designations as "featured article." We can integrate this data into our search algorithms and user interfaces, so structured data will make searching easier on Commons, and make it much easier to find high-quality images, and do complicated searches like “Paintings by women from the 18th century.” Structured data will also make it easier for institutions to add their media and associated metadata to Commons as provenance is an integral part of structured data, thus bringing more high quality content to the project. Both factors could be used by our community to reduce the "Yellow Milkmaid Syndrome" and the weight of other poor quality media vying for user attention in both search results and reuse in favour of high quality digital items created by institutions and better quality contributions by volunteers.

Technical documentation and planning

Some additional clarity was desired around the infrastructure work, particularly on the Commons and Wikidata side. These two projects constitute the bulk of the work in this area:

1. Wikibase Federation: The Wikibase repository for MediaInfo, which runs on Commons, will need to be "federated" with the Wikibase repository of Properties and Items on Wikidata. The MediaInfo record describing a specific file on Commons is stored and maintained on Commons, using Wikibase; The MediaInfo uses Properties and Items (concepts) defined on Wikidata to describe the content, license, material, authorship, etc. of the media file. In order to do this, Wikibase entities (such as Properties, Items, and MediaInfo) need to be able to be referenced and loaded unambiguously across wiki boundaries. This breaks the assumption in Wikibase code that all entities are accessed locally and that there is only a single namespace for each entity type. Also, UI widgets need to become able to talk to multiple repositories, instead of just one.

   As an example, within Commons, a Wikibase claim can be created, such as "depiction of: horse". "Horse" will be linked to the Wikidata item for the concept of a horse, thus allowing users to potentially find it through queries like "Photographs of mammals from 1910". It also automatically gives us translations of "horse" since Wikidata includes labels in multiple languages. This requires that Commons be able to simultaneously utilize local structured data and Wikidata data. The work for this effort is documented here: https://phabricator.wikimedia.org/T76007 and here: https://phabricator.wikimedia.org/T149580

1. Storage level integration (MCR): This allows multiple kinds of content to be stored as parts of the same wiki page, sharing a history, and being treated as a unit with regards to protection, watching, moving, deletion, etc. MCR (mutli-content revisions) breaks the fundamental assumption of MediaWiki, that per revision there is only one type of content.

We feel confident that we’ve discussed these efforts at length both internally and with the community, understand the design challenges and have the necessary expertise in Wikidata and MediaWiki to execute them successfully.

Networking & hardware estimates

One reviewer has asked for a more specific breakdown of capex which is reasonable considering a project of this magnitude. There are several factors that make doing this in a conventional format more difficult but also hopefully provide some reassurance that capacity issues are understood and well-managed. In general, we expect the $100,000 of capex to be distributed across Wikibase, Search and Wikidata Query system proportionally each year of the project.

1. Currently, Wikidata is stored in the same format as all other project content and in the same systems, which makes it both difficult to isolate particular usage and, more importantly, not particularly useful to do so from an operational perspective. This is because English Wikipedia content and traffic are so great that other projects are essentially free riders and inconsequential in the aggregate. This is also true for search.
2. A critical (but not well-communicated) attribute of the project is that it is primarily a conversion of unstructured data in the form of wikitext to a structured format in wikibase. The amount of data that is added is minimal. Therefore, considering #1, we do not foresee a step function increase in the storage requirements of any of the systems involved.
3. Technical operations at the Foundation are performed by a relatively small team of highly skilled personnel. While there are some specialized roles such as DBAs, people generally move around as needed which is highly efficient, but makes direct allocation of costs difficult. The team is also highly efficient in terms of capex, with very lean capital outlays over the life of the project.

Program

Institutional Support

We would definitely like to draw attention to the letters of support from major Institutional partners in this effort in the appendix. Digital Public Library of America, the National Archives, and Europeana have all voiced their support (see appendix). Our volunteer communities and affiliates have created amazing allies in the GLAM sector because we are mission aligned, and dealing with the same kinds of problems, both technical and social, that these older institutions aspire to work on.

Volunteer work structure

Community liaisons and GLAM staff will work directly with the community, socializing the work of the project team. Primary tasks include:

1. Keeping the community, and especially experienced volunteers, updated about the development of the tools via help pages, workshops, and talks.
2. Taking feedback from the community to the development team in order to enable them to be a part of the development process.
3. Supporting the community in decision processes around how to make use of structured data possibilities.

4. Helping identify edge cases, such as complex templates, that may require more support.

When structured data-aware tools are ready for use (including VisualEditor, Upload Wizard, and Media Viewer), a mix of Wikidata-familiar volunteers and early adopters from the Commons community will develop initial best practices and start communicating that out into collaborators in their own circles of influence. The community itself has built a number of tools that may augment WMF-built tools for adding structured data. For example, Magnus Manske, an expert in how Commons and Wikidata content is structured, has built a prototype tool called CommonsEdge that is both an API and an interface to that API. He expects the tool to add metadata to 16-19 million Commons files. We welcome these volunteer-led innovations.

Use will grow from there. We expect greater collaboration and increased multilingual contribution of structured data through community events like Wiki Loves Monuments, Wiki Loves Earth, and others. These kinds of volunteer-led projects will iteratively improve on the tools and teaching materials available for supporting new contributors.

We have a successful recent precedent for this type of work with volunteers in Wikidata. We have let the community build the data model and they've done an excellent job. We expect the same will happen on Commons.

Diversity issues

Structured data means better searchability. Better searchability means having the ability to do large-scale analysis and find gaps and biases programmatically. With this analysis we can much more specifically show people where our biases are and help the community address them with tools like Wikidata Human Gender Indicators. The Wikidata community is already using this approach (leveraging our Wikidata Query Service) to identify coverage and bias issues and we can use the same approach for media.

Update to Staffing Plan

We have made a few small changes to the staffing plan we sent you previously. The changes are bolded in the attached document. A couple of allocations were not transferred over from the budget breakdown and the task allocation spreadsheet. The distributed nature of the work across multiple teams and a centralized new Structured Data team were balanced in the most recent review. The changes do not affect the budget as this information was captured in the original budget. The changes now reflect the full headcount and distributions across years.