Wikimedia Commons Data Roundtripping

Midterm Report

Wikimedia Commons Data Roundtripping

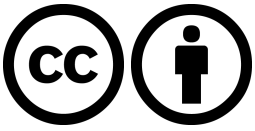
Midterm report 5 March 2019

Authors Albin Larsson, Susanna Ånäs, Maarten Zeinstra, and Paweł Marynowski

Cover image *Unknown actor, Hildur Engström and Julia Caesar in the revue Hertiginnan av Danviken at Kristallsalongen 1906*. Photograph 1906 by Anton Blomberg (1862–1936), Scanned glass negative, Swedish Performing Arts Agency, Public Domain.

With special thanks to the respondents to our survey and the interviewees.

Swedish National Heritage Board Reference Number: RAÄ-2018-3594



Project by
Swedish National Heritage Board

Project team
Albin Larsson, Swedish National Heritage Board, Lead Developer and Product Owner
Susanna Ånäs, Project Administrator (contractor)
Maarten Zeinstra, Researcher (contractor)
Paweł Marynowski, Software Developer (contractor)

Partnering institutions
Swedish Performing Arts Agency / David Jansson, Marianne Seid
The Nordic Museum / Aron Ambrosiani
Nationalmuseum / Karin Glasemann

About the project

Galleries, libraries, archives and museums (GLAMs) share media files from their collections on Wikimedia Commons, the media repository of Wikimedia projects. These media files are used in Wikipedia articles and other Wikimedia projects. Professional contributions from museums account for a substantial part of the content in Wikimedia Commons and help to enrich and bring context to millions of articles on Wikipedia in dozens of languages.

Users of Wikimedia projects are encouraged to add information to the metadata and descriptions of these media files. This helps to contextualise and describe the media. These additions and alterations can take many forms and include:

- New metadata (e.g. links to other works, creators, titles, geolocation, etc.)
- Altered metadata (e.g. different spelling, or fixing errors)
- Translations of metadata into other languages
- Added categorisations and classifications
- Digital alteration of media files (e.g. restoration and crops).

GLAMs don't usually adopt this contributed information about the media records that they provide to Wikimedia Commons. This is a significant opportunity loss for these institutions. The additions made by Wikimedia volunteers can help the institute reach a wider audience, correct mistakes, add details and overall enrich the experience of heritage.

Structured Data on Commons

Wikimedia Commons is publishing a new feature in 2019 that enhances the usability of contributed data and facilitates the extraction of metadata by third parties. Instead of storing media metadata in unstructured wikitext, Wikimedia Commons is adopting the structured data functionality that is known from Wikidata.

Structured Data on Commons makes it easier to access and extract data from Wikimedia Commons computationally. This enables the Data Roundtripping project to prototype and research the reuse of user contributed data in Wikimedia Commons.

Scope

The purpose of this project is to research, design and prototype technical solutions that would make it much easier and less work intensive for GLAM collections managers to review and copy the metadata of the media files they have shared on Wikimedia Commons.

The project itself will not develop any software that is *tailor-made for any specific collection management* system unless such a system is open source, nor develop a tool for viewing or download *statistics* for media files uploaded to Commons by GLAMs. Also, *metadata licensing* is not in the focus of the project.

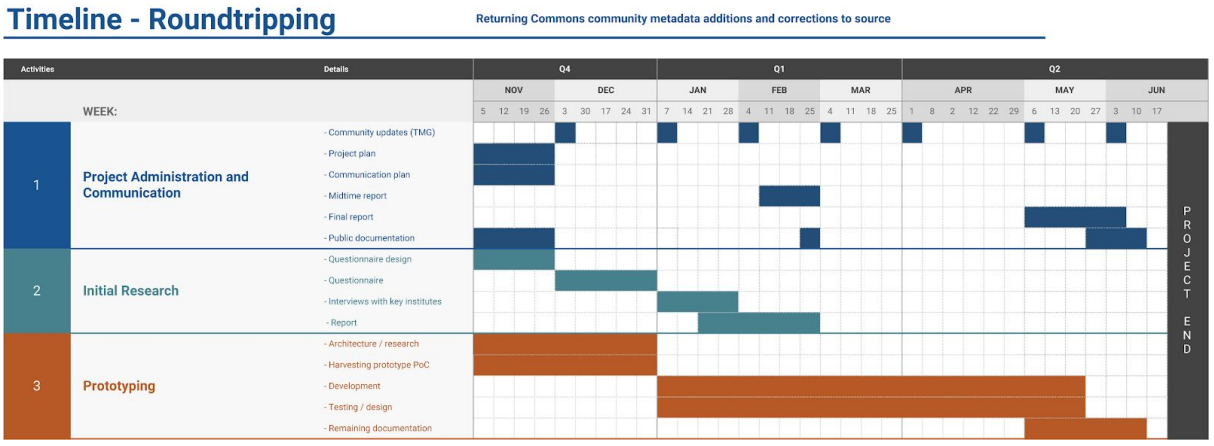
The respondents of the background survey have mainly been from Western countries and specifically from the countries of the project team members. Further interviews are conducted with participating institutions with selected institutions. The pilots are carried out with Swedish GLAMs participating in the project themselves.

Progress

The project runs from November 2018 to June 2019. It includes a survey that is published in December 2018, three pilots for tooling in the context of media uploads and metadata retrieval, and the accompanying reports on workshops and co-creation with the team and partnering institutions.

The project team convened in Stockholm 14–15 January 2019. The main goals of the meeting were to apply collaborative design methods with the project team, and to interview the piloting organisations' representatives.

Presentations of the project have currently been accepted in WikiConNL in Utrecht, Netherlands and the Creative Commons Summit in Lisbon, Portugal.



Read complete project documentation in the project page →

The survey

What are the needs and expectations of GLAMs to adopt user contributed information from Wikimedia projects into their collection registration systems?

Maarten Zeinstra / IP Squared has been tasked to research how third-party metadata is adopted in collection management systems of GLAMs in order to determine the scope of possible interventions to adopt third party metadata from Wikimedia Commons.

The research quantifies the underlying reasons for not adopting enriched metadata, such as technical barriers, lack of resources, lack of knowledge, or lack of trust of the source or contributor by the institutes.

The research shows the conclusions of a quantitative and qualitative research. The research provides a series of recommendations for making a minimum workable prototype that highlights the identified opportunities and reduces the identified challenges.

Survey respondents

The 38 respondents were mostly from Finland, Sweden and the Netherlands, or countries with native English speakers or where people are very comfortable with English. Two thirds of the respondents have uploaded media on Wikimedia Commons. Half used mass uploading tools, some upload individual files and a few have had a Wikimedian-in-residence.

Interviews

The detailed survey results above were presented to the project's working group. Members of the working group used this information to gather additional qualitative research results.

An additional five people were interviewed from three institutions.

- Multiple archivists, Swedish Performing Arts Agency
- Head of collection, and a Media producer, The Nordic Museum
- Program manager, Nederlands Instituut voor Beeld en Geluid

Key findings

Direct user contributions

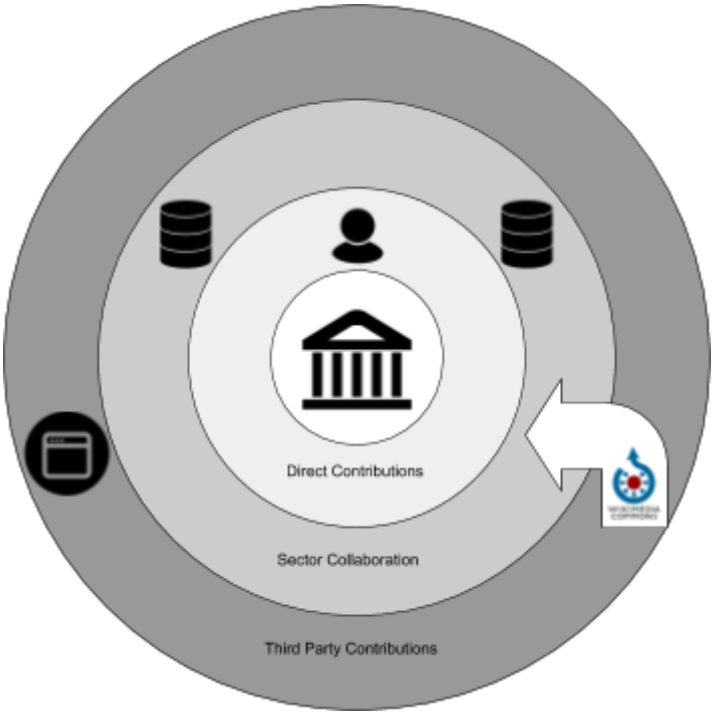
Direct user contributions refer to contact made via emails, contact forms, letters and calls, and direct personal contact. 60% of the respondents have direct user contributions. Reasons for not adopting direct user contributions are lack of verifiability of the source (25%), technical issues (25%), lack of resources (15%), and the lack of trust (15%).

Sector collaboration

Two thirds adopt information from other sources within the GLAM-sector. The most used authority files, thesauri, ontologies, etc. are Getty authority files, OCLC authority files, Wikidata, U.S. Library of Congress Authority Files and Kulturnav. The most cited reason for not working in sector collaboration is the lack of (technical) resources.

Third party collaborations

Third party collaborations help to index, digitise, transcribe, etc. collection information. Only about 37% of the respondents use general third party contributions in their digitisation and quality assurance projects. The tools are instead usually custom-made for an institution or a sector. Even when institutions adopt third party contributions, only a low percentage update their own collection information on a regular basis. Constraints in resources and technical constraints are the most cited barriers to adopt this information.



Desired Wikimedia Commons interactions

The interviewees brought up several needs and desires that Structured Data on Commons could leverage. These are described in more detail in the report.

Needs that were mentioned often included for example contributing to a rich set of name variants in different spellings, aliases and pseudonyms to make searching easier. Types of definitions for content include tags, translations for descriptions, and well-defined sources for statements. The idea of campaigns to enrich metadata content on Wikimedia Commons emerged through the interviews. Two of the three pilots will be carried out as a campaign.[1]

In addition to taking advantage of changes made on Wikimedia Commons, the interviewees would like to be able to push changes in their databases to Commons.

The barriers to making use of the user contributions according to the interviewees are related to lack of trust and verifiability of the source. The report presents several recommendations that could tackle this challenge.[2]

Recommendations

The research brings forward a set of recommendations for lowering the technical constraints of adopting third party contributions from Wikimedia Commons. These will inform the development of a tool created in the Data Roundtripping project, but they can also be used by GLAM organisations and the Wikimedia community to enhance interactions with the contributed content.

1. Focus on altered metadata, contextual metadata translations, and authority references
2. Generate trust by showing user information
3. Present structured data on Wikimedia Commons as an authority file
4. Integrate unique identifiers
5. Integrate other authority files

Read the report *Research Report – Returning commons community metadata additions and corrections to source →*

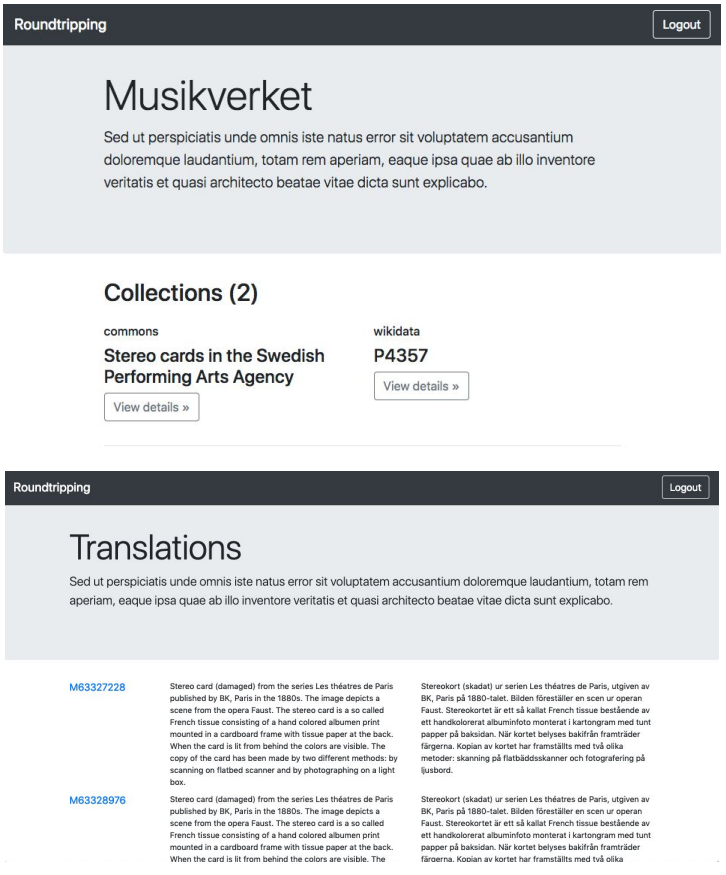
[1] Research report, p. 25.
[2] Research report, p. 8

The team meeting

The project team collaboratively investigated the affordances offered by current and newly developed technologies in Wikimedia Commons by creating a wide set of scenarios in a meeting in Stockholm 14–15 January 2019.

On the second day of the meeting, museum experts from Swedish Performing Arts Agency and The Nordic Museum were interviewed. An interview script had been developed based on the outcomes of the questionnaire and informed by the collaboratively created scenarios.

Based on the response of the interviewees, three scenarios were further developed into pilots to be investigated in the Data Roundtripping project. Based on the interviewees' input, the team concluded that there is large added value when Wikimedia Commons contributors **add translations of existing metadata**, **add descriptions about the subject matter of contributed content**, and **link to other sources that verify metadata of a media file**.[3]



User interface mockups for the Roundtripping tool

[3] Research report, p. 9.

Pilots

The Data Roundtripping project will investigate the use of Structured Data on Commons alongside currently available means for exchanging metadata between the GLAM and Wikimedia projects. According to the current project plans and the current schedule of the Structured Data on Commons project, one pilot will deal with data stored in Wikitext, one will target data in Wikidata, and one will likely use Structured Data on Commons. These diverse pilots will give insight to differences in the environments and allow for future recommendations.

For the purpose of the project, the **Data Roundtripping** web application will be developed. It will consist of two parts: **The backend** retrieves and processes the required data, while the **end-user tool** provides the actual user interface to users at GLAM institutions. The first part is to be in place by late February, and the latter will be improved during the entire spring. The tool is aimed to support the pilot projects as well as the research.

1. Swedish Performing Arts Agency

Low barrier pilot for engaging audiences in a translation campaign

Swedish Performing Arts Agency's pilot is created around an upload of 1200 glassplate photographs. In previous uploads they have themselves translated all descriptions into English before upload. In the pilot the aim is to allow them to upload the images with Swedish information only, arrange a translation campaign with the images and ingest the translated descriptions back to their own system.

Workflow

The images are uploaded to Wikimedia Commons using traditional upload methods. This will be followed by a crowdsourcing campaign, asking the audience to translate the descriptions from Swedish to English. There will be a **Translation tool** created for an easy setup and participation in the translation campaign.

During the campaign, Swedish Performing Arts Agency will be able to see the progress and statistics of the campaign in the **Roundtripping tool**. After the campaign, they will be able to review differences between current data and their original contribution and export the new data.

Swedish Performing Arts Agency can import the exported data into their system in the manner they choose. The process will be documented in the project documentation.

Takeaways

This is a low barrier pilot, since verification of the translations can be done without additional research. It tackles contributions added to the current technological framework in Wikimedia Commons, which will continue to coexist with Structured Data on Commons.

2. Nationalmuseum

Making use of third party contributions via authority data in Wikidata

The purpose of this pilot is to take advantage of authority IDs that have been added to the institution's contributions in Wikidata by other users. The IDs to be collected are Wikidata IDs themselves and ULAN IDs found in the Wikidata entries.

Workflow

The **Roundtripping tool** features a tool to query Wikidata for the institution’s contributions. Nationalmuseum will choose the properties which it wants to retrieve, and download the data as a csv file (Comma Separated Values). They will import the data into their system in the manner they choose, and it will be documented in the project documentation.

Takeaways

The pilot gives an opportunity to compare structured data approaches to non-structured data ones. While observing the museum’s choices of data ingestion, the project gains valuable information for recommendations to guide future work.

3. Nordic Museum

Depicting image content with the Structured Data on Commons capabilities

The topic of Nordic Museum’s pilot is to record metadata of the imagery from their exhibition about British fashion.

In this pilot the project team will work in collaboration with the curators of the museum to identify what kind of data should be recorded related to their images. Campaigns for collecting crowdsourced data will be prepared based on that.

Workflow

The work is based on identifying the most suitable data to be stored in Wikimedia Commons in collaboration with the museum staff. Most likely the work will focus on the use of the **depicts** property, and/or recording the **author(s).**

The process starts with a traditional media mass upload to Wikimedia Commons, and the structured metadata will be added programmatically separately. A campaign will be arranged to ask for user contributions to enrich the metadata.

Next, the user contributions are retrieved to the Nordic Museum using the **Roundtripping tool**. Progress can be tracked during the campaign, and the data can be exported and downloaded after the campaign ends.

Like with the previous pilots, the institution will import the data into their system in the manner they choose, and it will be documented in the project documentation.

Takeaways

In the Nordic Museum pilot it will be possible to compare structured data approaches to non-structured data ones with metrics developed for that purpose. The institutions will document the ways in which they bring the data back to their system, and the documentation will be made available as part of the reporting.

Albert Nycop as Bror Pettersson in Jon Blund at Östermalmsteatern, 1908, Atelier Jaeger - Swedish Performing Arts Agency. Scanned glass negative. Public Domain.