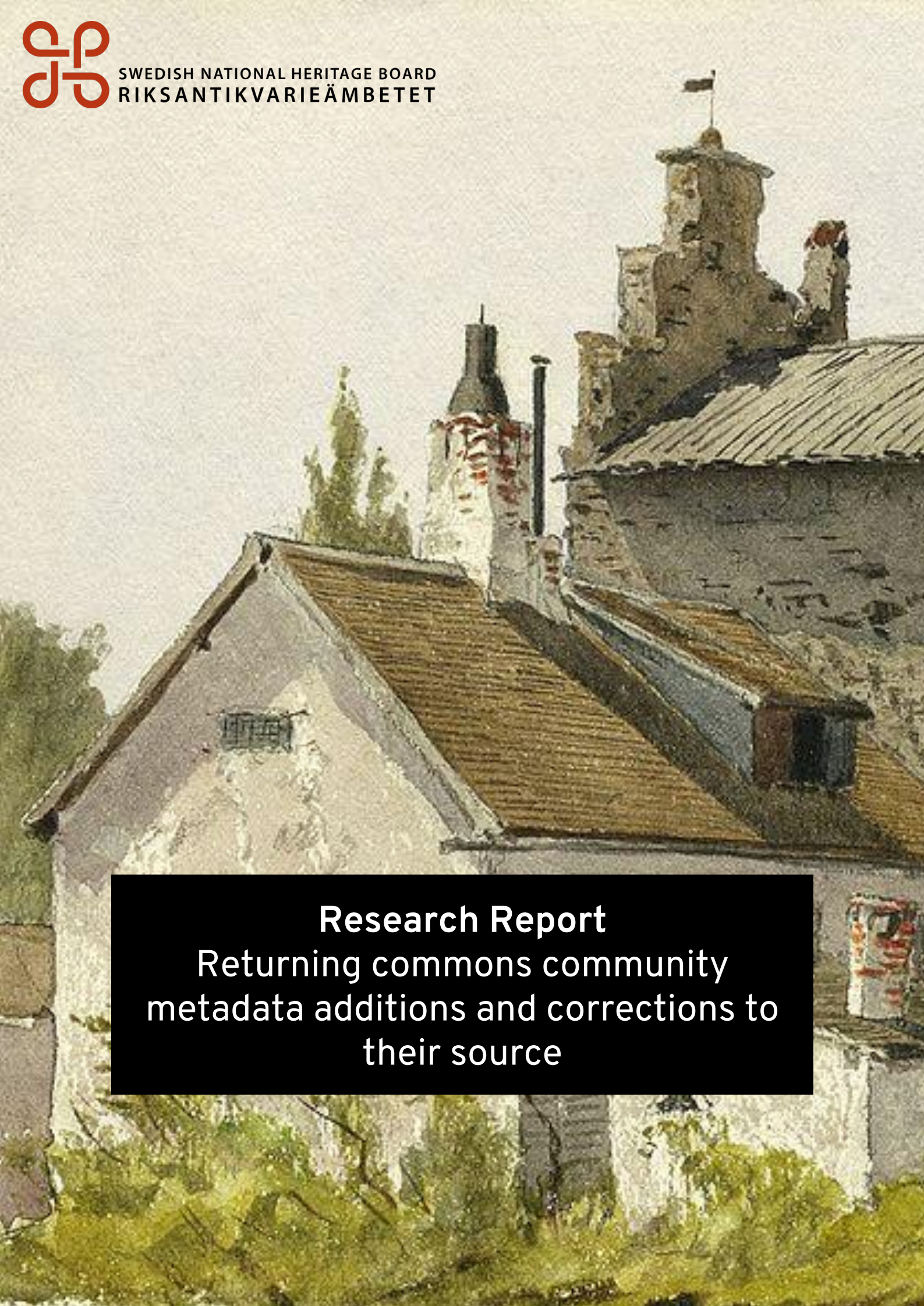




SWEDISH NATIONAL HERITAGE BOARD
RIKSANTIKVARIEÄMBETET



Research Report
Returning commons community
metadata additions and corrections to
their source

Colofon

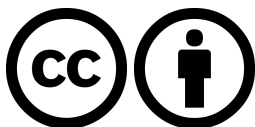
Author: Maarten Zeinstra, [IP Squared](#).

Cover image: Drawing by A.T. Gellerstedt, 1873, (14th century stone house, Stora Hästnäs, Visby, Gotland, Sweden), via [Wikimedia Commons](#), public domain.

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

Made possible by the [Swedish National Heritage Board](#).

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



License: Unless otherwise indicated this document is licensed under a Creative Commons Attribution 4.0 license. You are free to distribute, share and build upon this work as long as you credit the author of this document. You can find the complete text of this license here <https://creativecommons.org/licenses/by/4.0/legalcode>

Credit 'Research report – Returning Commons community metadata additions and corrections to source' by Maarten Zeinstra, IP Squared (2019) / CC BY 4.0.

Table of contents

Introduction	4
Research results	7
Recommendations	9
1. Lower technical barriers for adoption by creating simple export functionality	
2. Focus on altered metadata, contextual metadata translations, and authority references	
3. Generate trust by showing user information	
4. Present structured data on Wikimedia Commons as an authority file	
5. Integrate unique identifiers	
6. Integrate other authority files	
Detailed Survey Results	11
Respondents	11
Direct user contributions	14
Sector collaborations	15
Third party contributions	16
Technical capabilities	18
Tracking changes on Wikimedia Commons	20
Detailed Qualitative Research results	22
Wishes from the interviewees	23
Campaigns	24
Wikimedia as a an authority	24
Next steps	25
Prototype tool	25
Addendum 1. Questionnaire	26
Addendum 2 Qualitative research questions	33

Introduction

Galleries, libraries, archives and museums (GLAMs) share their 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 museum to Wikimedia Commons account for a substantial part of the commons and help to enrich and bring context to millions of articles on Wikipedia in dozens of languages.

Users of Wikimedia Commons 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. descriptions about the content, creators, geolocation, etc.),
- Altered metadata (e.g. different spelling, or fixing errors),
- Translations of metadata,
- Added categorisations and classifications of media files,
- 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.

Wikimedia Commons is adding a [new feature](#) to its platform in 2019 that enlarge the usability of contributed data and ease the extraction of this metadata by third parties. Instead of storing data in unstructured wikitext, Wikimedia Commons is adopting the structured data functionality that is known from Wikidata.

This research document is part of a larger project called '[Wikimedia Commons Data Roundtripping](#)' by the [Swedish National Heritage Board](#). The purpose of this project is to research, design and prototype technical solutions that would make it easier and less work intensive for GLAM collections managers to review, copy and add sources to the metadata within their collection management systems of media files they have contributed on Wikimedia Commons.

The project aims to:

- Research the desirability and requirements of GLAM-collections managers in regards to retrieving metadata added to their files post-upload on Wikimedia Commons,
- Develop and test a prototype tool that supports GLAM-collections managers in identifying, reviewing and retrieving added or changed metadata to media files,
- Report on lessons learned and recommend future actions.

[IP Squared](#), a strategic information advice agency, is tasked to research whether third-party metadata is adopted in the collection management systems of GLAMs in order to determine the scope of possible interventions to help to increase the adoption of this third party metadata.

The research will show that reasons for not adopting this enriched metadata can be found in many areas. For example, due to technical reasons, based on a lack of resources, a lack of knowledge, or a lack of trust of the source or contributor by the institutes. This research reports tries to find and quantify these reasons and provides recommendations to lower these barriers, given the new functionality of structured data on Wikimedia Commons. It does so using a quantitative and qualitative approach.

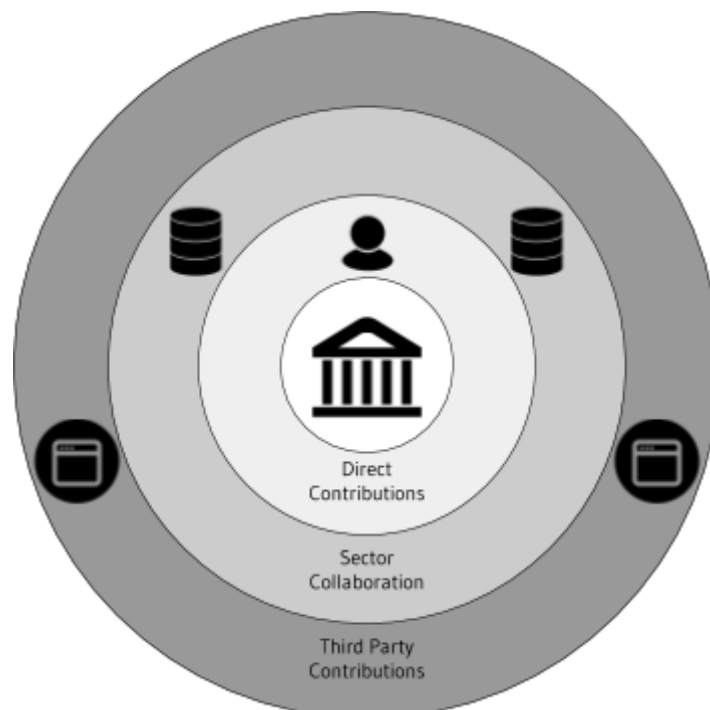
The research presented in this document is structured around the following research question:

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

First, a questionnaire was developed and communicated to GLAMs across the world. The outcomes of the questionnaire are used to provide a quantitative perspective of the needs and expectations of GLAMs to adopt user contributed information from Wikimedia projects into their collection registration systems. The design of the survey can be found in the addendum one.

The survey did not limit itself to user contributed information on Wikimedia Commons. The quantitative survey makes an additional distinction between other types of data contributions to give an overview of the needs and expectations of the GLAMs. The survey splits user contributions into three categories:

- Direct user contributions (e.g emails and phone calls),
- Sector collaborations (e.g. authority files, and thesauri), and
- General third party contributions (e.g. crowdsourcing and Wikimedia Commons).



This approach provides research data to describe indications what might block general adoption of third party information. These indications helps to determine if the found barriers are distinct to Wikimedia Commons or are general sectoral issues of the GLAMs.

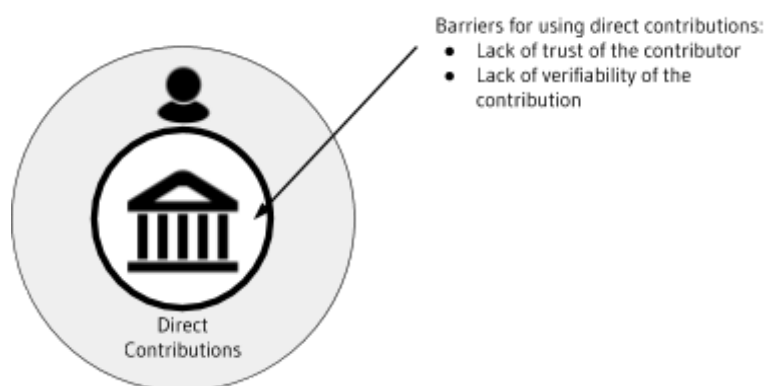
The survey also indexes technical capabilities of GLAMs to adopt metadata from third parties. It gathers data on the technological readiness for adopting third party contributions to the GLAMs collection management systems. These include bulk import of metadata and methods for disambiguation of data.

In a second phase an interview script was developed based on the outcomes of the questionnaire to support local interviews with the aim to get at the challenges and opportunities of selected institutions for further collaboration. The focus of these interviews were to provide indications of the current practices of maintaining collection metadata and verify the outcomes of the quantitative research. The design of the interview can be found in the addendum three.

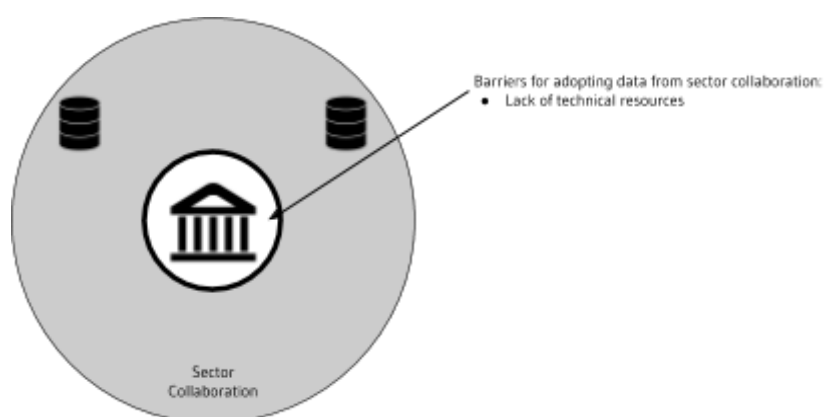
Research results

The survey results makes it clear that there is an interest in extracting enriched metadata from Wikimedia Commons. It is also clear that most organisations will struggle to ingest this data. Automatic processes to ingest data and sector collaboration using authority data are not common practice among the institutions that responded to the survey.

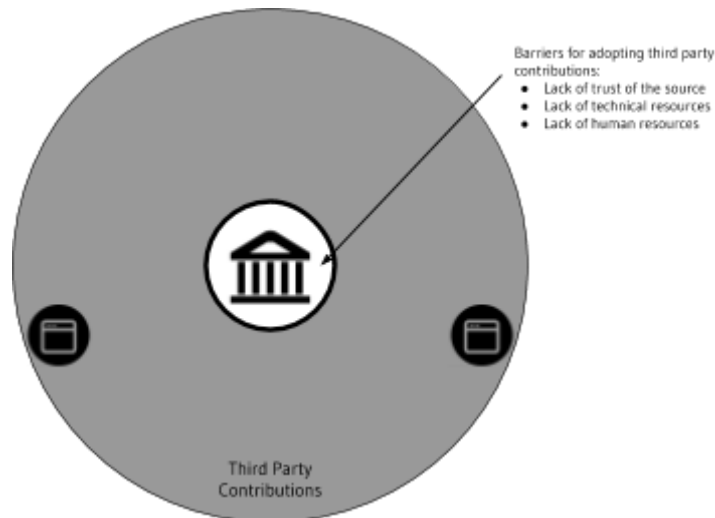
This is evident from the barriers that are indicated by the respondents. The barriers to ingest direct user contributions are mostly based on the actual person contributing information and the type of information contributed. Here GLAMs mostly cite a lack of trust and verifiability of the source. The barrier of institutions to adopt new data from a person, which might not be an expert, is higher or lower depending on the content of the contributed metadata. A simple typographical error is easier accepted than adding substantive information about records.



When looking at data from other institutions and from authority files there is usually little question about the quality of the information, instead technical resources are mentioned as a barrier to adopt this type of metadata. Instead an authority file is usually adopted by linking, instead of duplicating metadata.



When asked about third party contributions our respondents cite a general lack of resources as the highest barrier to adopt third party metadata. This includes technical resources as well as human resources. Additionally some indicate that the trustworthiness of this information also comes into play.



The barriers to adopt third party information are stacked. With direct third party contributions there are very little constraint in (technical) resources, constraints in human resources (e.g. time and expertise) might still occur. However the central of all contributed information is verifiability and trust of the source. As long as that barrier is not lowered data adoption by GLAMs will proof difficult.

This lack of trust and verifiability is less problematic for sector collaborations (e.g. thesauri and authority files), as these are developed by other trusted parties in the heritage field, usually other GLAMs.

This is supported by interviews with stakeholders. During these interviews the barrier of trust is highlighted. Interviewees indicated that they can spend up to one hour per change to validate the source and suggestion before changing their records, but likewise would always link to a trusted authority file despite not having different metadata for one resource.

Interviewees indicated that Wikimedia Commons have a large added value when Wikimedia Commons contributors:

1. Add translations of existing metadata
2. Add descriptions about the subject matter of contributed content
3. Link to other sources that verify metadata of a media file.

Additionally institutions indicated that if Wikimedia Commons would become more similar to an authority file in use and operation then it is more likely that they will adopt this structured information.

Recommendations

It is therefore recommended that this project tries to lower the constraints in technical resources and other technical issues by developing a tool that works lowers the identified barriers for adoption.

1. Lower technical barriers for adoption by creating simple export functionality

It is not necessary to adopt API standards as most respondents and interviewees indicated that they do not use APIs for ingesting data from other parties, except for authority files as linked data.

Practically this means that the minimum viable product of this project should not include complicated data export functionality. Being able to download information as a Comma Separated Values (CSV) would be sufficient.

2. Focus on altered metadata, contextual metadata translations, and authority references

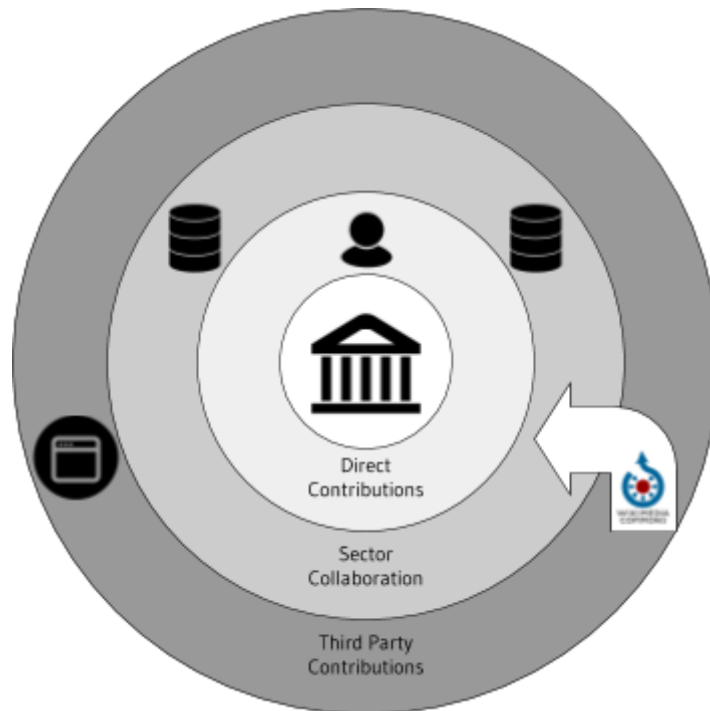
The survey and interviews have shown that altered metadata, contextual metadata, translation and references to authority files are most valued by the GLAMs. Contextual metadata included structured data of objects, persons or entities that are depicted by the contributed media files.

3. Generate trust by showing user information

A barrier for adopting information from a direct contributor relies on a level of trust of that contributor. This also applies to data added by Wikimedians. Showing that the edits were made by people who generally do not make edits that are reversed helps build trust in the added data.

4. Present structured data on Wikimedia Commons as an authority file

This project has an opportunity to promote the structured data of Wikimedia Commons as an authority file itself. Therefore moving the perceived barriers from 'third party collaborations' to 'sector collaborations'.



The research showed that sector collaborations do not suffer from high barriers of lack of trust, thus aligning Wikimedia Commons with these authority files lowers that barrier for adoption.

A secondary recommendation related to this is to highlight the linked data functionality of structured data of Wikimedia Commons. GLAMs should be able to link to contributed media on Wikimedia Commons using a URI. This allows further adoption of Wikimedia Commons as an authority file for media files.

5. Integrate unique identifiers

A large percentage of respondents indicated that they have public unique identifiers for objects in their collections. A good step to promote the new capabilities of structured data on Wikimedia Commons is to add these identifiers to contributed media on Wikimedia Commons.

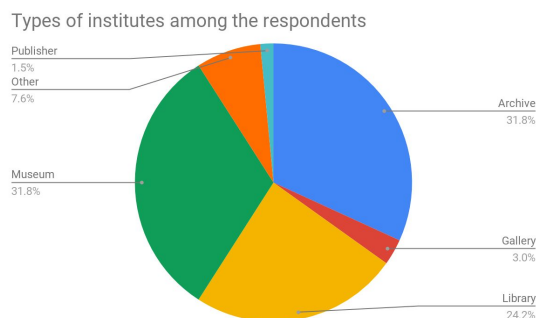
6. Integrate other authority files

It is also recommended for Wikimedians to work on integrating other structured data like thesauri and authority files of other GLAMs and heritage institutions. It is believed that this will increase trust in the structured data.

Detailed Survey Results

Respondents

The survey had 38 respondents. Most of these are from an institution that combines functions of an archive, museum and/or library. While 'Gallery' is a part of the well known GLAM acronym, it is hardly used to self-identify an heritage institution outside of the UK. Some respondents were identified as publishers, the remaining respondents indicated that they represented sector institutes.



A substantial part of the respondents were from Sweden, Finland, the Netherlands and the UK. This is not surprising as the project team for this research consists of people from these countries. Furthermore the survey was only available in English, meaning that it might only have attracted respondents that feel comfortable writing in that language.

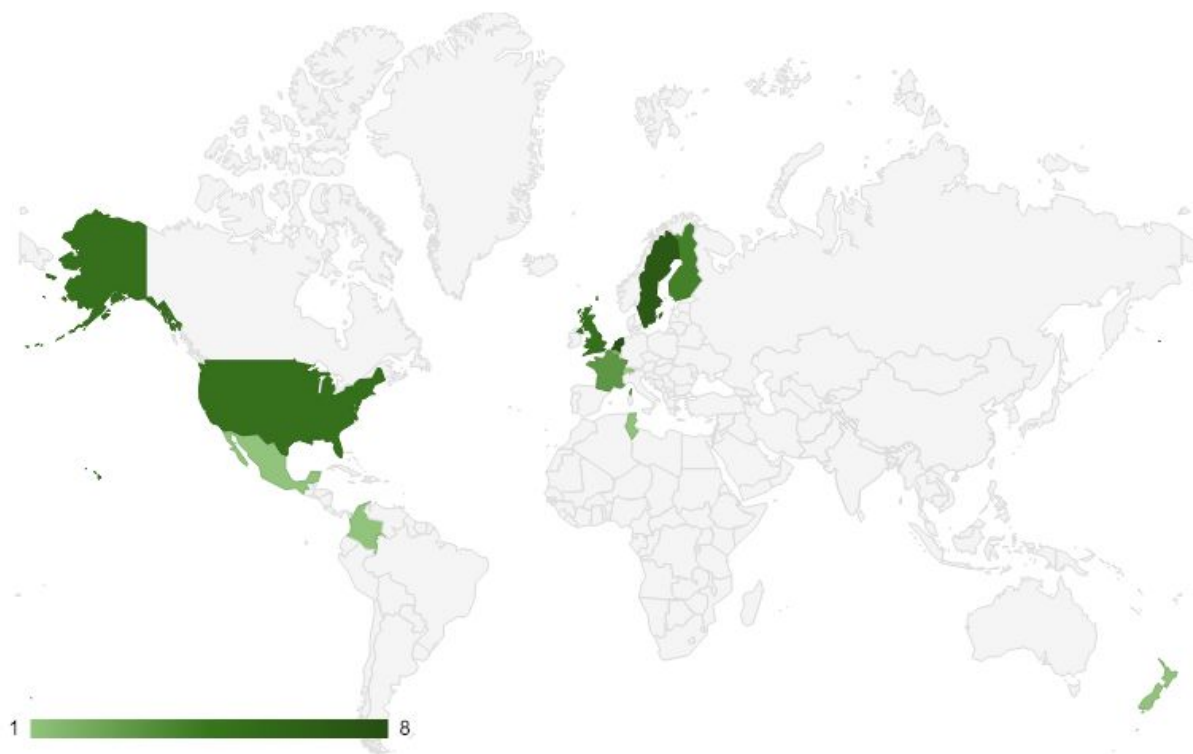


Fig. a world map indicating the origin of participating institutions in the survey.

Respondents were sourced using various social media outreaches like Twitter, Facebook, Slack, mailing lists, and direct mailings. The research team is confident that invitations to participate in the study reached a sufficient part of the OpenGLAM community.

The community is the audience of this research and the project, which might skew the research data. Participants were sourced from the OpenGLAM community, as such they are all likely to be aware and active of Wikimedia Commons and its possibilities.



Maarten Zeinstra
@mzeinstra



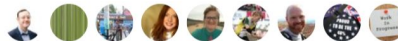
Work at a #GLAM? Participate in the #[Wikimedia](#) #[Commons](#) Data Roundtripping project. Fill in our survey to help design a tool to retrieve structured metadata from Wikimedia Commons to enrich your collection:



Returning additions and corrections of Wikimedia ...
An increasing amount of Galleries, Libraries, Archives, and Museums (GLAMs) upload media to Wikimedia Commons. These media files help to enrich the experience...
docs.google.com

9:47 AM - 3 Dec 2018

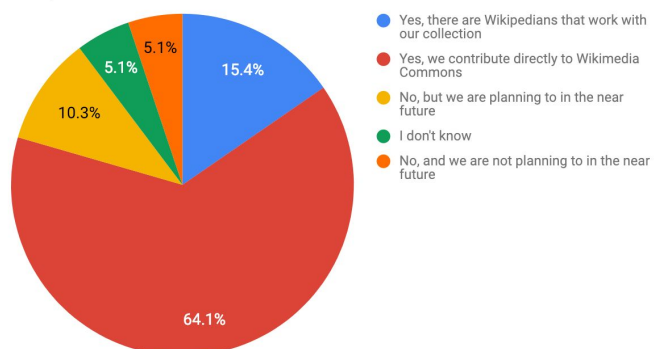
40 Retweets 33 Likes



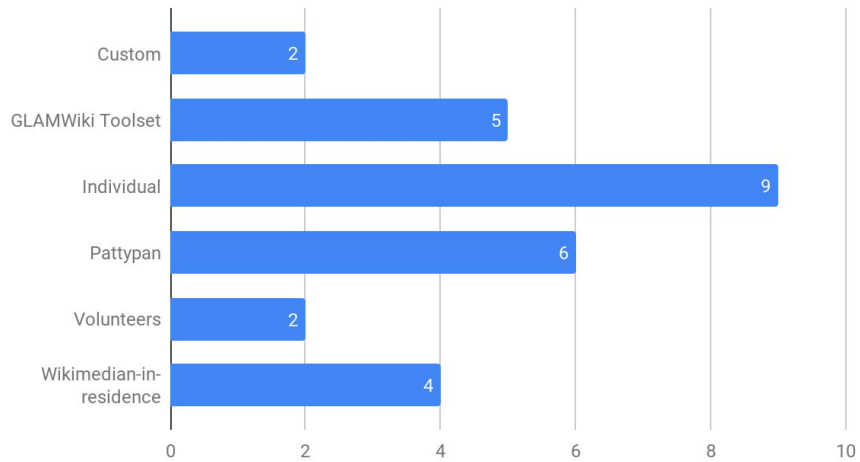
Contributors to Wikimedia Commons among the respondents

Almost two-thirds of the respondents already directly contribute to Wikimedia Commons. Another 15% know or are in contact with Wikimedia volunteers that work with their collection information. A further 10% plans to make direct contributions to Wikimedia Commons in the near future. The remaining 10% does not know whether they are contributing to Wikimedia Commons or is not planning to contribute to Wikimedia Commons in the near future.

Respondents that contribute to Wikimedia Commons



Way of contributing media files to Wikimedia Commons



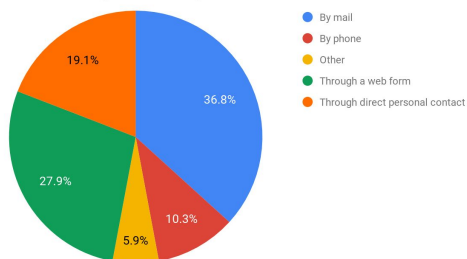
When those institutions that contribute to Wikimedia Commons are asked how they upload media files, almost 50% indicate that they use a mass uploading tool developed for Wikimedia Commons. The tools that were most used were [Pattypan](#) and the [GLAMWiki Toolset](#). Another nine institutions indicated that they individually upload media files to Wikimedia Commons. This can be interpreted as an indicator of technological readiness of these institutions to work with tools that handle mass information.

Direct user contributions

The first section of the survey focused on direct user contributions. These direct user contributions contain feedback received from the audiences of GLAMs via contact forms, emails, letters and calls. These contributions are usually based on information presented on the platforms that are controlled by the institutions themselves, like their own websites, and media platforms that don't allow for third party changes. This feedback can include recommendations to enhance or alter the metadata of the collection.

60% of the respondents indicated that they receive direct contributions from their audiences. Most of these institutions have departments or personnel that process these types of feedback. A couple of the respondents indicated that they have weekly team meetings to process this type of feedback.

Mediums through which respondents receives direct feedback



Respondents were asked which medium was used to retrieve this feedback. Most indicated that these were given by email and webforms. Direct personal contact and phone calls are less common. Direct user contributions can also come from other platforms like Wikimedia Commons talk pages, social media or image platforms that are owned and/or operated by

Information received from direct user contributions is not always adopted by the institutions. The reasons given for not adopting this feedback is a lack of verifiability of the source, technical issues, and a lack of resources. Specifically, over 25% of the respondents that receive direct user contributions cite a lack of verifiability of the source, another 25% indicate technical issues, including mismatched data standards as reasons for not adopting user contributed information. A lack of resources and a lack of trust both are indicated by 15% of the respondents.

Institutes that do make changes to their dataset based on these direct user contributions do so at least once a month.

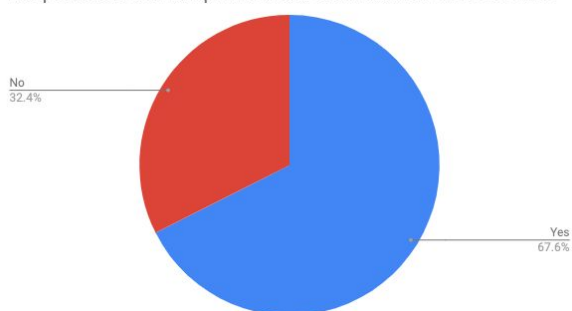
Sector collaborations

The GLAM sector exchanges metadata using [authority files](#), [thesauri](#) and other data instruments. These types of collaborations are usually hosted by one large institution like the OCLC or the U.S. Library of Congress. They host a centralised resource of structured information that can be imported or linked to from the institution's websites. GLAMs either link to or copy this information. Often they work in consortia that where they contribute information to authority files and thesauri.

These data instruments allow the GLAM sector to collaboratively work towards high quality metadata and lowers the resources needed to keep their collection information up-to-date. It allows institutions to share clean normalised data with each other. Two thirds of the respondents indicated that they use ontologies, biographies, authority files, and taxonomies in their collections.

A substantial part of these respondents update their internal data on a regular basis based on these shared resources. This either means that they use the resource directly or that they make a local copy of that resource on a regular basis.

Respondents that adopt metadata from outside their institute



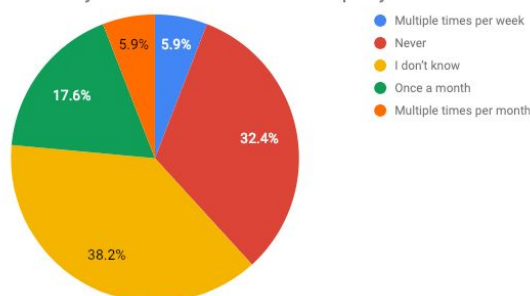
The topics of most of the authority files are for bibliographic information and geographic information. Resources that are mentioned by a majority of respondents are:

- [Getty authority files](#)
- [OCLC authority files](#)
- [Wikidata](#)
- [U.S. Library of Congress Authority Files](#)
- [Kulturnav](#)

All of these resources have a multitude of authority files. Local standardised [geonames](#), shared bibliographic information about artists, etc. were also mentioned by the respondents.

In comparison to direct user contributions the reason for not adopting this type of data from sector collaboration is not rooted in lack of verifiability but lack in (technical) resources. This type of collaboration within the GLAM sector are starting to become more commonplace, but require a lot of (technical) collaboration and coordination within the institutions and within the sector.

How often all respondents change information based on feedback that you receive from these third party contributions?



However, when asked if the respondents would adopt information from crowdsourcing platforms, respondents said that they don't often change internal metadata based on information from these third parties. About 30% of respondents indicated that they adopted information on a regular basis.

Third party contributions

Next to direct user contributions and sector collaboration, GLAMs also get metadata contributions from third parties. These general third party contributions occur when working in crowdsourcing project like [Zooniverse](#) or transcription services. These can include crowdsourcing projects that 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. Meaning that a large majority of institutions do not participate in crowdsourcing initiatives. Those that do mostly used custom developed solutions that work only for their institutions. Independant platforms like Zooniverse and e-manuscripta.ch are also mentioned.

Crowdsourcing platforms that are mentioned by the respondents include

- [doedat.be](#)
- [Flickr](#)
- [Hetvolk.nl](#)
- [ba.e-pics.ethz.ch](#)
- [smapshot.heig-vd.ch](#)
- [e-manuscripta.ch](#)
- [Kuvakokoelmat.fi](#)
- [Waisda](#)
- [Notes from Nature](#)
- [Atlas of Living Australia](#)
- [Youtube](#)
- [Zooniverse](#)

Additionally, Wikimedia projects are also mentioned.

Wikimedia Commons as a third party contributor

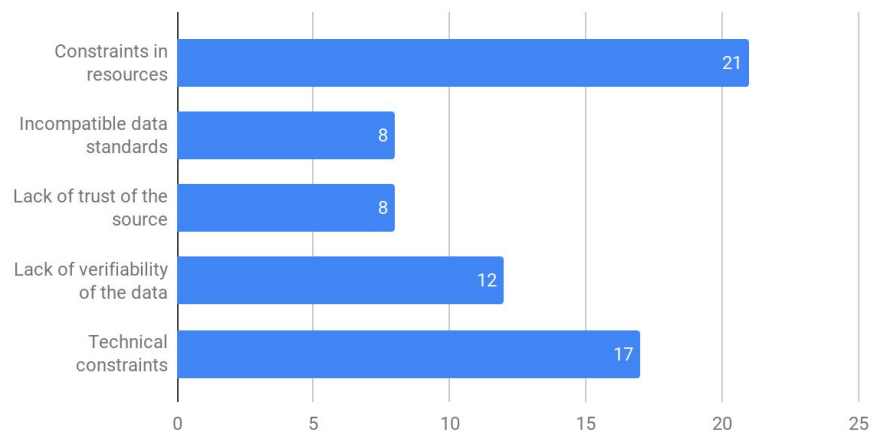
In 2019 Wikimedia Commons will add functionality for structured data in addition to describing media files in Wikitext as is currently available. This will allow GLAMs to query Wikimedia Commons and get structured data back that can be used by the GLAMs to enhance their data.

All respondents indicated that they are interested in data from Wikimedia projects. Almost two thirds of the respondents have indicated that they are very interested, whereas a third indicated that they were moderately interested.

When prompted about the type of metadata they were most interested in from media files contributed to Wikimedia Commons, respondents indicated that were most interested in new metadata that can be received from Wikimedia Commons, followed by changed metadata, translations, and added categorisation and classifications.

The most frequently given answer to why institutions do not adopt information from Wikimedia Commons or information from other crowdsourcing platforms is constraints in (human and financial) resources and technical constraints. These types of constraints are more blocking than constraints of lack of trust or lack of verifiability which is given as a reason for not adopting direct user contributions.

Respondents reasons for not adopting data from Wikimedia Commons



Technical capabilities

As seen in the previous section the biggest hurdle to use third party metadata from other platforms is a lack of human or financial resources or a lack of technical resources. The latter means that it is difficult to get that information into the collection management system of the institutions. The survey asked about these collection management systems in order to determine if these can be used to overcome this barrier.

However, almost all respondents use a different collection management system. It is therefore not efficient to look into developing specific means to get third party metadata into the collection management systems of the GLAMs. Collection management systems that are used more than once are:

- [Axiell](#), either adLib or EMu
- [ALMA](#)
- [Memorix \(Maior\)](#)
- [MuseumPlus](#)
- [Primus](#)
- [TMS](#)

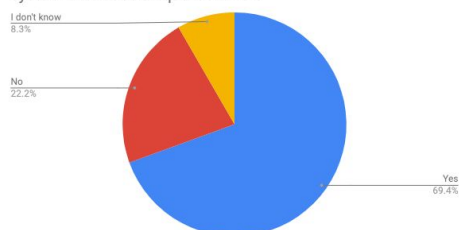
Note, it is interesting to recall that the respondents are from groups of countries, each country had dominating collection management systems. Memorix (Maior), and AdLib is for example used in institutions from the Netherlands. MuseumPlus is seen in French and German speaking institutions, etc.

Since almost no two institutions use the same collection management system has the consequence that any tool that is going to be developed for the GLAM sector should not be specifically build to match the technical capabilities of a collection management system.

There are overlapping capabilities of Collection Management Systems that can be employed to make integration of third party resources simpler. These are using unique identifiers and allowing for bulk import of third party metadata.

Unique identifiers

Percentage of respondents that have collection registration system which use unique identifiers



More than half of the institutions use public unique identifiers for their individual records. A small 10% does not have enough knowledge about unique identifiers to answer this question. These unique identifiers can be used to map metadata on Wikimedia Commons to metadata in the collection management system of the institution.

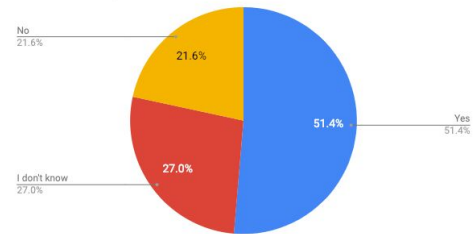
Bulk import

More than half of the institutions allow bulk import of data into their collection management systems. The extent of these capabilities is however unclear. This due to the differences in understanding by the respondents in the capacities of their Collection Management Systems.

After asked whether the can ingest bulk data into their systems, respondents were asked to describe their data standards. Often the respondents indicated that data that is to be ingested includes manual labor and custom written scripts, or is done directly on a database using SQL. Some indicated that they had REST-APIs and SOAP-APIs, where OAI-PMH is the only API standard that is mentioned. OAI-PMH is however only an export API, having that standard does not mean that the institutions can import data using that standard. In terms of data formats CSV, XML, JSON are mentioned. The data standards that are used are Dublin Core, MARC and EAD (over XML).

Again it is important to note that not all respondents were familiar with the technical capabilities of their collection management system. Meaning that the above indication of used standards is not completely representative for this survey.

Respondents that have collection registration systems that allow bulk import



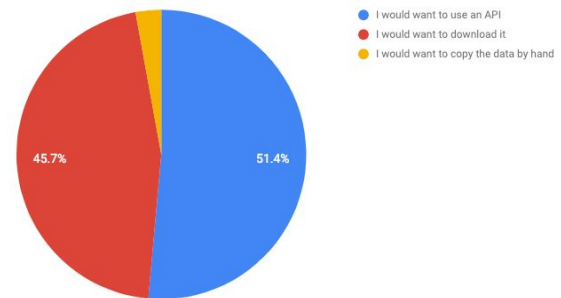
Tracking changes on Wikimedia Commons

Finally, respondents were asked how they currently track which changes are made to their contributed collections to Wikimedia Commons and how they ideally want to track these changes and be notified of them. This information helps to determine the tooling that project can create for these institutions.

Interaction with the information

First the respondents were asked what they want to do with this information. It is clear from the respondents that they want to be able to download the metadata from Wikimedia Commons, either via an API or directly as a data file. Only very few want to be able to copy data by hand. However for preview purposes this should be possible as well.

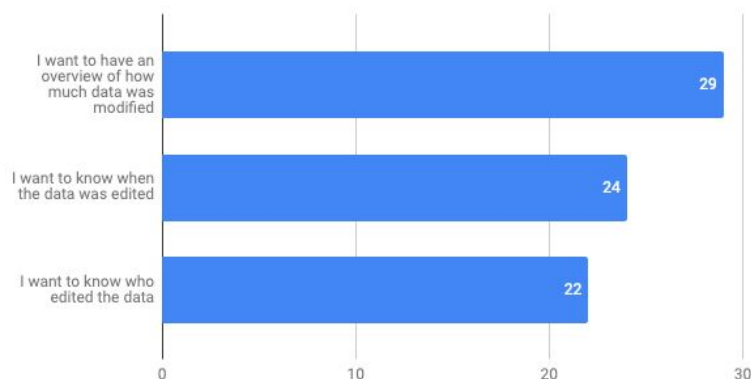
How respondents want to be able to retrieve structured data from the commons



Notification of changes

Respondents could select multiple answers when asked how respondents want to know about the changes made to datasets that are on Wikimedia Commons. Most respondents were interested in knowing that the data was modified, followed by when it was modified. Least interest was who modified the data. Additionally two respondents would like to some sort of confidence or anti-vandalism mechanism to be visible (e.g. if the person editing is a frequent editor).

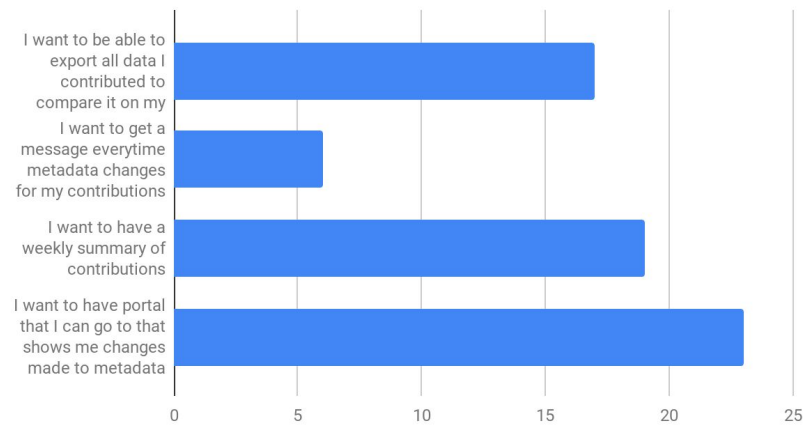
What respondents want to be know about changes made to data in their collections on Commons



Respondents want different ways of being notified when data changed in their collections. About a third want to have a portal where they can see changes to their collections. An additional 30% would like to see a weekly summary of changes to their data. A quarter is content with only having a download functionality and comparing data with their own collection information. A few would like to be informed with every change of data in their dataset, when they occur. One

respondent indicated that they were using the existing watchlist RSS feed functionality on Wikimedia Commons to retrieve changed data.

How respondents want to be informed of changed metadata



When asked for further suggestions for additional features some respondents suggested adding statistics to the views and attention to their collections on Wikimedia Commons.

Detailed Qualitative Research results

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. Addendum two contains an interview script that was used to interview some key stakeholders in Sweden and the Netherlands.

An additional five people were interviewed from three institutions:

- Multiple archivists, [Musikverket](#)
- Head of collection, and a Media producer, [Nordiska Museet](#)
- Program manager, [Nederlands Instituut voor Beeld en Geluid](#)

One institute indicated that they receive about 10 to 15 requests for corrections per month about their collections. These mostly include suggestion for new spellings and to de-anonymize works (i.e. identify the creator or subjects of a work). Next to emails some of the institute have online webforms allow users to communicate possible mistakes. These forms are turned into requests queues that follow the same process as mails.

Most interviewee indicate that they had no formal procedure of processing these mails or other feedback. However they all indicate that each change needs to be verified, and that process can take up to an hour per change. When contributors indicated the source of their change or are themselves an authoritative source this amount of time is decreased.

Another interviewee indicated that they had a dedicated departement for metadata management and that when it comes to changing or adding information about persons a process with additional verification starts. These types of information are then put into shared authority files that is publicly available.

Another interviewee indicated that they tracked changed on Wikimedia Commons category via the build-in RSS function. This allowed the interviewee to see all changes made to files in that category. The interviewee indicated that they would however disregard a lot of these changes as they were not important for their work. Other interviewees were not aware of this functionality, some had little interest in the changes that they were presented with in relation to their donations to Wikimedia Commons (mostly categorisations).

One institution started their media donations more than 10 years ago. Over these years they have not been very interested in the unstructured data that is currently available on Wikimedia Commons. The institution is very interested in – and already experimenting – with the structured data that is available via wikidata.

Institutions that have been publishing data on Wikimedia Commons solely as part of their distribution channels might not see Wikimedia Commons as a source of data yet.

One institute indicated that they might ask a Wikimedian-in-Residence (WiR) to help with getting the new structural metadata out of Wikimedia Commons and into their own collection management systems to see if this new type of information will become relevant for them.

Wishes from the interviewees

After having explained the new functionality and potential for use of structured data on Wikimedia Commons, interviewers and interviewees discussed what kind of functionality they like to see made available using the new functionality of Wikimedia Commons.

Interviewees indicated that they have more structured data in their collection information system than that is currently attached to their contributed media on Wikimedia Commons. They indicated that next to tools for retrieving changed metadata they would like to see a tool that allows them to upload and update structured data to Wikimedia Commons as well. They indicated that they put a lot of effort in researching Information like names, birth and death dates, alternative names, pseudonyms. This information is not being added to their Wikimedia Commons contributions, as the process is currently too limited.

There was also a wish to maintain different spellings of a person names on Wikimedia Commons, as that would make matching easier with their internal collections. It also matches the internal practices of the GLAMs. This includes aliases and pseudonyms of depicted persons.

When asked about different types of data the interviewees would be interested in, the interviewees described functionality like categorisation of the content of data or semantic tags of content (e.g. what animal does this image depict, what play is this a photograph of, who were the actors in this photograph, where is this photograph taken, when is this photograph taken, etc.).

Another request by the interviewees is to have source information available with the statement. This is foreseen functionality of the structured data on Commons project. Having a confirmation that certain added information is confirmed trustworthy by a reputable source would be helpful in the adoption of this information.

Next to information about the content and context of the image, there was a general interest in having all these statements translated into different languages. This would be helpful to have the institute reach new audiences. This was not shared by all interviewees, some institutions have only national audience and have not much interest in translations.

For audiovisual material interviewees were interested in time-coded annotations, subtitles and translations of subtitles. The audiovisual archive indicated that there are already computer programs that can already automatically add some of this metadata to their collection registration system, but that precision is sometimes lacking.

Campaigns

During the interviews the ideas of using structured data on Wikimedia Commons as part of crowdsourcing campaigns. Where the community is asked to focus attention on describing or adding metadata to a specific dataset. This idea fits in the current collaboration between Wikimedian and the GLAM sector in projects like [Wiki Loves Monuments](#).

Most examples of these were campaigns directed at translating metadata and identifying subject matter in media files. There seems to be little interest in other types of metadata that could be crowdsourced. It could be argued that this is because these are very large categories of metadata that can be crowdsourced. Even though the homogenic group of interviewees might not see value in other types of metadata, this does not mean that this interest does not exists at other institutions.

Wikimedia as a an authority

Interviewees indicated that adopting existing authority files and thesauri helps to bring trust to Wikimedia Commons. Likewise, translating concepts into several languages will add value to the platform. These could help lowering barriers for the adoption of Wikimedia Commons.

Being able to link to the structured data would help the adoption of the type of information that becomes available via the structured data on Wikimedia Commons. This means that the structured data can become a external reference within their dataset that they can consult or communicate to third parties. This would increase the usability of the data without having to adopt the information into the system of the collection itself. An example of this is the [Europeana Annotation API](#) and the Enrich Europeana platform.

Some interviewees indicated that they do not intend to integrate data from Wikimedia Commons into their own systems. They described using a separate data layer instead that would be hosted next to the core data of their collection registration system.

While other were able to confirm that they have mass ingestion processes available for integrating third party metadata, they also argued that they were more likely adopt a link to the structured metadata rather than to duplicate the information. This would be lower mental barrier for the adoption of this information than creating a local copy of the data and merging it into the collection management systems.

Next steps

This research report is the result of the first phase of the '[Wikimedia Commons Data Roundtripping](#)' project. The next phases of the project are:

- Developing and testing a prototype tool that supports GLAM-collections managers in identifying, reviewing and retrieving added or changed metadata to media files.
- Report on lessons learned and recommend future actions.

Prototype tool

A data roundtripping web application will be developed. The initial focus of the tool will be to create an overview of altered metadata, new translations and references to authority files. These tools will be tested in three pilots:

- a translation pilot,
- a authority identifier pilot, and
- a contextual information pilot.

Translation pilot – Musikverket

During this pilot around 1.200 glass plate photographs will be uploaded to Wikimedia Commons. A simple translation tool will be created based on the prototype tool. During the pilot a small translation campaign is organised to translated the descriptions from Swedish to other languages. Musikverket will be able to import the translations into their system. The results of this pilot will be documented in the project documentation.

Authority identifiers pilot – Nationalmuseum

The data roundtripping web app features functionality to query the institution's contributions. Nationalmuseum will choose the properties (fields) which it wants to download as a csv (Comma Separated Values) file. The museum will explore the downloaded data during the pilot. The results of this pilot will be documented in the project documentation.

Contextual information – Nordic Museum

The Nordic Museum has contributed media 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. Contributions are retrieved using the prototype tool. The institution can explore the data and import into their system if they choose to. Results of this pilot will be documented in the project documentation.

Addendum 1. Questionnaire

The questionnaire informs to the capabilities of GLAMs to adopt data from third parties. The questionnaire does not limit itself to user contributed information in Wikimedia projects to be able to identify reasons for not adopting this information that are outside the possibilities of Wikimedia to solve with a technological solution (e.g. lack of trust).

Introduction

An increasing amount of Galleries, Libraries, Archives, and Museums (GLAMs) upload media to Wikimedia Commons. These media files help to enrich the experience of Wikipedia articles and increase the reach of heritage collections. The metadata of these media files are often augmented and enriched by the volunteers of Wikimedia projects. These enrichments often do not find their way back to the contributing GLAMs.

The Swedish National Heritage Board (www.raa.se) in collaboration with museums in Sweden are researching and prototyping a tool that makes it easier to extract this enriched metadata. This questionnaire indexes the use of third-party metadata in collection management systems in order to determine the scope of the tool that is going to be developed.

The questionnaire will take approximately 10 minutes to fill in. If you have additional information or suggestions feel free to mail the researcher (Maarten Zeinstra) at info@ip-squared.com.

Privacy statement

The information that you contribute to this questionnaire will only be used for research purposes for the duration of the project. We will not keep or distribute information about you or your institute unless you have given us prior consent to do so.

1. General Questions

1. What is your name?
2. At which institute do you work?
3. What is your role in the institute?
4. What type of institute is this? (select all that apply)
 - a. Gallery
 - b. Library
 - c. Archive
 - d. Museum
 - e. Other...
5. What type of collections does your institute hold?
6. What is your mail address?
7. Does your institution contributes media to Wikimedia projects?
 - a. Yes, we contribute directly to Wikimedia Commons
 - b. Yes, there are Wikipedians that work with our collection
 - c. No, but we are planning to in the near future
 - d. No, and we are not planning to in the near future
 - e. I don't know
8. If your institution contributes media to Wikimedia projects, please provide relevant links to your categories, projects, contributions, etc.
9. If your institutions contributes media to Wikimedia projects, please describe how you contributed these media files.

2. Individual contributions

Almost all institutions receive feedback from their audiences via contact forms, emails, letters and calls. This feedback can include recommendations to enhance or alter the metadata for your collection. These questions collect information about how often you get this feedback, what you do with it and what keeps you from adopting that information.

1. Do you receive feedback from audiences containing suggestions how to enhance or alter metadata?
 - a. Yes
 - b. No
 - c. I don't know
2. How often do you receive information from direct contributions? (choose most applicable)
 - a. Never
 - b. Once a month
 - c. Multiple times per month
 - d. Multiple times per week
 - e. Daily
 - f. I don't know
3. Do you have specific departments or individuals in your institutions that deal with these kinds of feedback?
 - a. Yes
 - b. No
 - c. I don't know
4. How do you receive comments on the metadata that you have about your collection (select all that apply):
 - a. By phone
 - b. By mail
 - c. Through a web form
 - d. Through direct personal contact
 - e. other
5. How often do you change information based on feedback that you receive from individuals? (choose most applicable)
 - a. Never
 - b. Once a month
 - c. Multiple times per month
 - d. Multiple times per week
 - e. Daily
 - f. I don't know
6. What reasons play a role for not adopting this information (select all that apply):
 - a. Lack of trust of the source
 - b. Lack of verifiability of the data
 - c. Constraints in resources
 - d. Incompatible data standards
 - e. Technical constraints

- f. Other ...
- 7. Do you have any additional descriptions about how you deal with direct user contributions?

3. Sector collaborations and crowdsourcing

The GLAM sector exchanges metadata using authority files, thesauri and other data instruments. This allows the sector to collaboratively work towards high quality metadata and lowers the resources needed to keep your collection information up-to-date.

1. Do you incorporate metadata that is created outside your institute?
 - a. Yes
 - b. No
 - c. I don't know
2. If so, what type of datasets do you use?
 - a. Authority files
 - b. Thesauri
 - c. Linked data
 - d. Other instruments, like
3. Please describe the datasets/thesauri that you use
4. How often do you change information in your own collection registration system based on feedback that you receive from these collaborations? (choose most applicable)
 - a. Never
 - b. Once a month
 - c. Multiple times per month
 - d. Multiple times per week
 - e. Daily
 - f. I don't know
 - g. N/A, we use the shared resource itself
 - h. Other
5. What reasons play a role for not adopting this information (select all that apply):
 - a. Lack of trust of the source
 - b. Lack of verifiability of the data
 - c. Constraints in resources
 - d. Incompatible data standards
 - e. Technical constraints
 - f. Other ...
6. Do you contribute to sector collaborations?
 - a. Yes
 - b. No
 - c. I don't know
7. If you contribute to sector collaborations, please describe your contributions
8. Do you have any other remarks about your work in sector collaborations to improve your metadata?

4. Third party contributions

Some institutions use general third party contributions in their digitisation and quality assurance projects. These can include crowdsourcing projects that help to index, transcribe, etc. collection information. Over time, metadata that you contributed to Wikimedia Commons is enhanced by the community. They either add translation, fix errors, transform media, etc.

1. Do you use crowdsourcing platforms to help in digitisation process or metadata proces?
 - a. Yes
 - b. No
2. If so, please name and describe the crowdsourcing platforms you use?

Wikimedia and other third parties have duplicates of your metadata or new metadata based on objects that are in your collection. Over time these duplications change. Additional metadata may have been added, small corrections might have been contributed.

1. To what degree are you interested in metadata from Wikimedia projects (e.g. Wikipedia and Wikimedia Commons)?
 - a. Very interested
 - b. Moderately interested
 - c. Not interested
2. Do you currently keep track of changed metadata from your Wikimedia Commons contributions?
 - a. Yes
 - b. No
 - c. No, but I want to
3. In what kind of data would you be interested? (select all that apply)
 - a. New metadata (e.g. links to other works, creators, titles, geographical coordinates, depicted objects and people, etc.)
 - b. Altered metadata (e.g. different spelling, or fixing errors)
 - c. Translations of metadata into other languages
 - d. Added categorisations and classifications
 - e. Digital alteration of media files (e.g. restoration and crops)
 - f. other..
4. What reasons play a role for not adopting this information? (select all that apply)
 - a. Lack of trust of the source
 - b. Lack of verifiability of the data
 - c. Constraints in resources
 - d. Incompatible data standards
 - e. Technical constraints
 - f. Other ...
5. How often do you change or update information based on feedback that you receive from these third party contributions? (choose most applicable)
 - a. Never
 - b. Sometimes
 - c. Often
 - d. Always

6. Do you have any other remarks about third party contributions?

5. Ingesting third party information

This section asks questions regarding the technological readiness of your collection registration system on ingest information

1. What is the name of the main collection registration system software you use?
2. Does your collection registration system use unique identifiers for objects that are available to the public?
 - a. Yes
 - b. No
 - c. I don't know
3. Does your collection registration system allow for bulk import of metadata from external sources?
 - a. Yes
 - b. No
 - c. I don't know
4. If your collection registration system allows for bulk import of metadata, please describe the technical aspects of this import? (e.g. data standards, data formats, API standard, etc.)

Tracking changes on Wikimedia Commons

The Swedish National Heritage Board in collaboration with museums in Sweden are researching and prototyping a tool that to easier to track altered metadata of your contributions on Wikimedia Commons. The following questions are directly about this tool.

5. How would you like to be informed of changed metadata? (check all that apply):
 - a. I want to get a message everytime metadata changes for my contributions on Wikimedia Commons
 - b. I want to have portal that I can go to that shows me changes made to metadata
 - c. I want to be able to export all data I contributed to compare it on my own
 - d. I want to have a regular summary of contributions
 - e. Other:
6. What else would you like to know about the metadata connected to your contributions?
 - a. I want to know who edited the data
 - b. I want to know when the data was edited
 - c. I want to have an overview of how much data was modified
 - d. Other:
7. How would do you want to be able to retrieve the data?
 - a. I would want to download it
 - b. I would want to use an API
 - c. I would want to copy the data by hand
 - d. Other
8. Do you have any other remarks about retrieving (changed) metadata from Wikimedia projects?

6. Final Questions

Thank you for filling the questionnaire. The following questions are about our follow up of this questionnaire.

1. Can we mention you or your institute in our research report?
 - a. Yes
 - b. No
 - c. Other..
2. Can we contact you for further questions?
 - a. Yes
 - b. No
3. Do you want us to keep you informed about the research report and development of our tool?
 - a. Yes
 - b. No
4. Do you have any final remarks for us?

Addendum 2 Qualitative research questions

Profile the interviewee

1. What role do you have in your institution?
2. What kind of information do you process in your role?
3. What kind of collections does your institute hold (e.g. items, images, text, time-based media, born-digital, cultural history, art, science, etc.)

Collection information

4. When is new information added to your collection system? (e.g. when a new item is acquired, after an exhibition) Can you describe that process?
5. When do you update information in your collection?
6. How much time do you spend on updating existing data?

User Contributed information

7. Could you describe how you work with your collection information?
 - a. What is the role of user contributed information in your organisation's data?
 - b. What limitations do you find in incorporating those in your collection data?
 - c. What user contributed information are you especially interested in?
 - d. Do you work with authority files, ontologies, thesauri, etc.
8. What kind of information is difficult to get and would you like to have crowdsourced? (e.g. coordinates, image descriptions, transcriptions, etc.)
9. Can your systems bulk ingest data from other sources?
 - a. Could you talk us how that works? (e.g. what kind of data can be ingested? What technical standard is used?)
10. Do you use crowdsourcing platforms?
 - a. If you do, do the crowdsourcing platforms let multiply volunteers make the same suggestion/validate each others changes affect these barriers?

Wikimedia projects

11. How/do you work with Wikimedia projects?
 - a. What kind of data/content is shared, which projects (Wikipedia, Wikimedia Commons, Wikidata, Wikisource)?
 - b. What is the process of getting this information in Wikimedia, have you participated in that? (WiR, WMSE, internally...)
 - c. Do you have a category on Wikimedia Commons? Can you show us?
12. What information do you want to have about your objects on Wikimedia Commons?