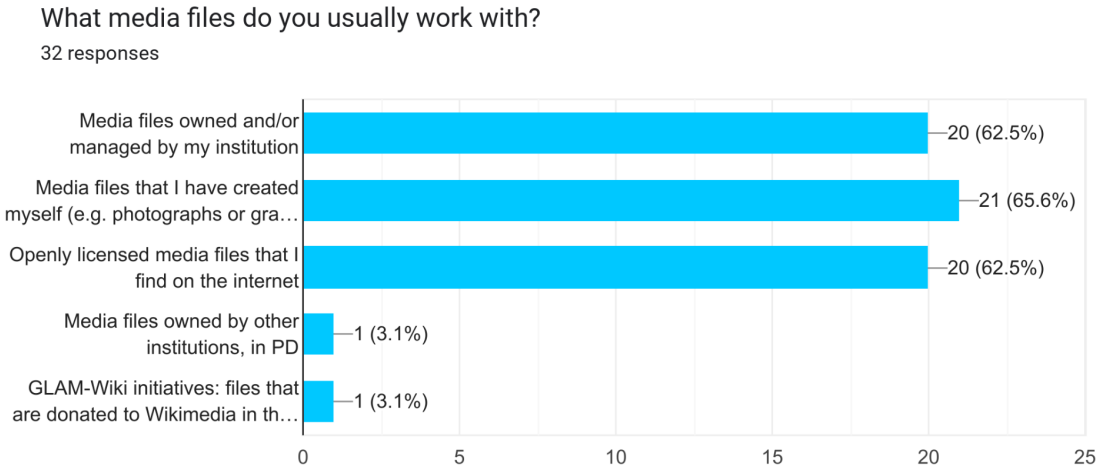
Analysis of the first OpenRefine SDC open survey

This survey is associated with the Wikimedia-funded project to develop an extension for OpenRefine that enables uploading files and editing structured data on Wikimedia Commons. You can read more about the project here: https://meta.wikimedia.org/wiki/Grants:Project/CS%26S/Structured_Data_on_Wikimedia_Commons_functionalities_in_OpenRefine
The survey questionnaire was prepared by Lozana Rossenova and Sandra Fauconnier. The survey was carried out between April 4th-14th, 2022. This results analysis is prepared by Lozana Rossenova. *32 responses in total*

Question 1

The first question was concerned with media file stewardship and provenance. It looks like **at least 60% of the survey participants are based at institutions**, but seemingly many of them also publish other media files, outside the parameters of institutional collections.



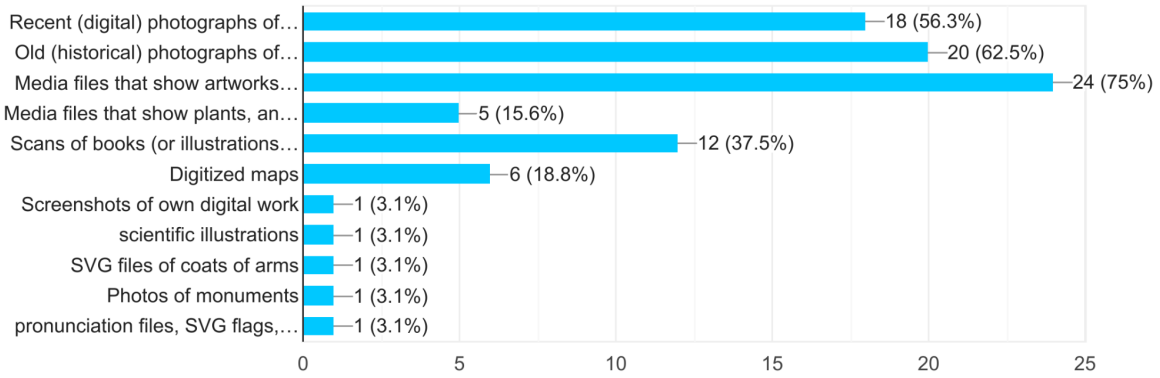
Transcription of possible answers:
Option 1) Media files owned and/or managed by my institution
Option 2) Media files that I have created myself (e.g. photographs or graphics that I have made)
Option 3) Openly licensed media files that I find on the internet
Opton 4) Other… (Users could freely enter their answers here)

Question 2

By far the most popular form of media upload is **representations of artworks and museum objects** (i.e. 75% of participants engage with this type of media), followed closely by historical photographs, and recent digital photographs.

What do the media files that you upload to Wikimedia Commons usually show? What are they generally about?

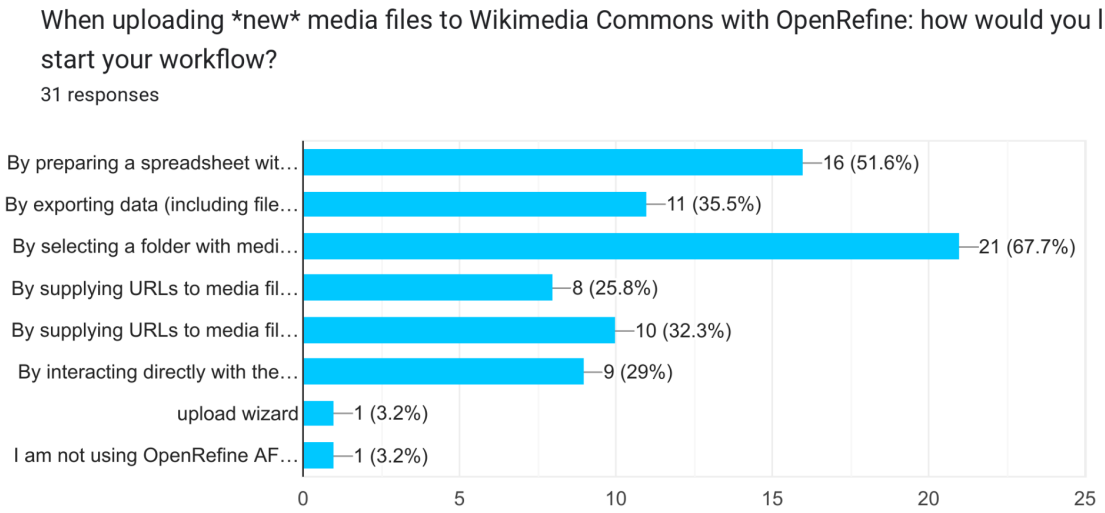
32 responses



Transcription of possible answers:
Option 1) Recent (digital) photographs of people, events, food, buildings, vehicles, other objects...
Option 2) Old (historical) photographs of people, events, food, buildings, vehicles, other objects...
Option 3) Media files that show artworks and/or objects in museums or cultural institutions (paintings, sculptures, prints, drawings, and all other types of GLAM collection items); including public domain artworks
Option 4) Media files that show plants, animals or other living organisms
Option 5) Scans of books (or illustrations in books)
Option 6) Digitized maps
Opton 7) Other… (Users could freely enter their answers here)

Question 3

This question wanted to explore mental models for the initial workflow steps prior to creating a new project in OpenRefine. The most popular mental model of starting the media upload process is to **start from a folder with media files** on one's own machine (>65% of participants ticked this option). However, many users also ticked the option to **start by preparing a spreadsheet** with all the relevant data (>50%). Many (roughly a third) also selected various options that start from an institutional repository (or other external web resource) and involve some level of data exporting or URL gathering, etc., to be then imported into OpenRefine.



Transcription of possible answers:
Option 1) By preparing a spreadsheet with all the data (including file paths for the media files) and uploading that
Option 2) By exporting data (including file paths for the media files) in a standard format (e.g. XML, a csv file) from my institution's repository and uploading that
Option 3) By selecting a folder with media files from my local computer
Option 4) By supplying URLs to media files stored in my institutional repository
Option 5) By supplying URLs to media files stored elsewhere on the web (e.g. Flickr)
Option 6) By interacting directly with the API of my institutional repository to select what data to include in a new OpenRefine project
Opton 7) Other… (Users could freely enter their answers here)

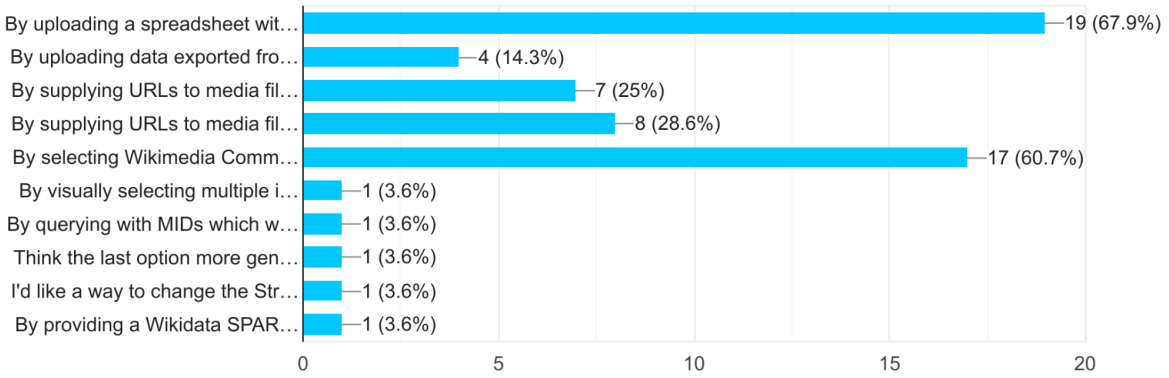
Almost all of these options – apart from the one involving a spreadsheet being prepared – will **require some additional UX adjustments in OpenRefine**. The 'starting from a folder with files' workflow is particularly challenging, since there is currently no way to just select a set of media files to then start a new project in OpenRefine, and there is also no way to attach additional metadata to these files (except for manual entry) within the current UX framework of OpenRefine. We will need to decide if we want to accommodate this model for user interaction at all and what new UX patterns will need to be developed, or alternatively if we do not accommodate this – how to still meet user expectations, e.g. via better onboarding materials, etc.

Question 4

This question focused on workflows with media files that are already uploaded to Commons. The two most popular answers in this question were connected to workflows we're already working towards accommodating fully: namely **starting from a well prepared spreadsheet** (>65%), or **starting from a Wikimedia category(ies)** and 'pulling' associated files (>60%). Most of the other answers that got close to a third of participants' ticks, include workflows that we also already considered, e.g. starting from an XML metadata file, or supplying a list of URLs pointing to the source files.

When *editing* media files that are already available on Wikimedia Commons: how would you like to start your workflow with OpenRefine?

28 responses



Transcription of possible answers:
Option 1) By uploading a spreadsheet with all the data (including file names for the media files) and then using OpenRefine to check the file names against Wikimedia Commons and match to corresponding media
Option 2)By uploading data exported from my institution's repository (including file names for the media files) in a standard format (e.g. XML) and then using OpenRefine to check the file names against Wikimedia Commons and match to corresponding media
Option 3) By supplying URLs to media files stored in my institutional repository and then using OpenRefine to check the file names against Wikimedia Commons and match to corresponding media
Option 4) By supplying URLs to media files stored elsewhere on the web (e.g. Flickr) and then using OpenRefine to check the file names against Wikimedia Commons and match to corresponding media
Option 5) By selecting Wikimedia Commons Categories and getting back a list of files tagged with those categories; OpenRefine already matches these files to corresponding media in Wikimedia Commons
Opton 6) Other… (Users could freely enter their answers here)

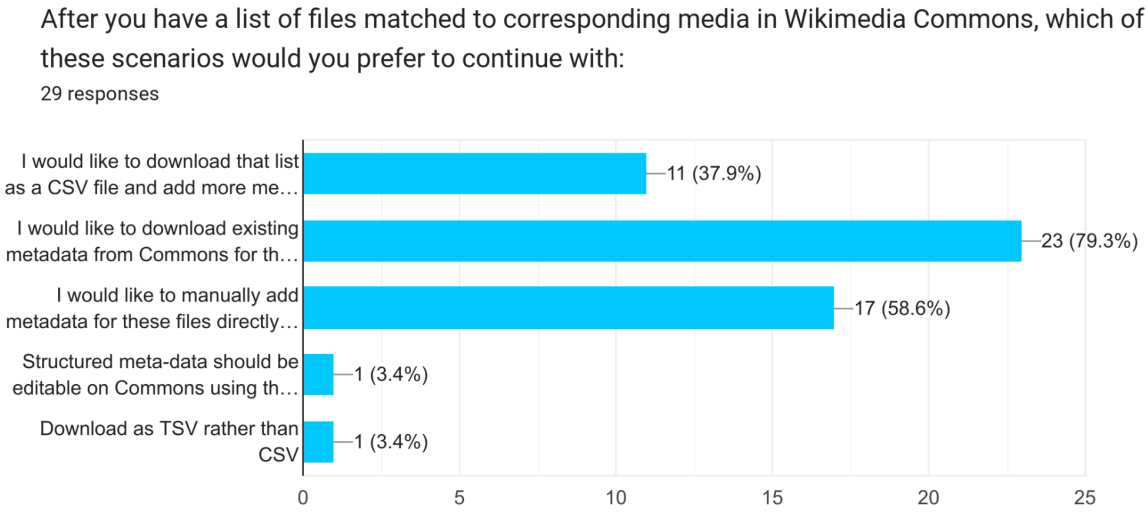
There were however, five participants also requested several **alternative options** connected to e.g. bulk metadata editing (with tools like cat-a-lot and AutoWikiBrowser being mentioned vs data attached to individual files); **reconciliation based on MIDs**, not just file names or URLs (though this could potentially be addressed by the user preparing the list of MIDs as a CSV / XML file); visual selection (though this would be hard to accommodate in the current project scope); and also **more general querying** of the MediaWiki API or the SPARQL endpoints of Wikimedia Commons or Wikidata, though presumably this could also happen outside OpenRefine and the results of the queries could then be supplied to OpenRefine in the form of a structured data file.

Question 5

This question looked at processes following reconciliation. The processes that we are already working towards supporting fully, i.e. **parsing wikitext for reconciled media files**

and then restructuring for upload as SDC, received most votes (~80%); with the entirely **manual approach** that does not require additional UX work from us, coming second (~60%). A **Pattypan-like solution**, which involves an extra spreadsheet download step and doesn't necessarily require any extra work from us, also received significant votes (>35%).

One user provided an additional perspective, which seemed to express frustration with the fact that SDC is not currently editable by standard workflow tools in use by the Commons community and requires entirely manual efforts, which become untenable. **We should aim to emphasize:** 1) the potential to automate and replicate same metadata (when appropriate) for large batches of media files via OpenRefine, too; and 2) the fact that OpenRefine will be able to handle not only SDC specific workflows, but general Commons upload workflows, too, with SDC being added as part of that complete end-to-end workflow, not something that requires significant extra effort.



Transcription of possible answers:
Option 1) I would like to download that list as a CSV file and add more metadata to it locally before uploading again to OpenRefine to proceed with uploading metadata to Commons.
Option 2) I would like to download existing metadata from Commons for these files and then parse and convert unstructured text into structured metadata, so that I can then upload these new structured data statements back to Commons.
Option 3) I would like to manually add metadata for these files directly in OpenRefine before uploading to Commons as structured data statements.
Opton 4) Other… (Users could freely enter their answers here)

Question 6

Six participants responded positively to this question asking them if they'd be willing to share files with sample data from their actual datasets for our internal testing purposes. We can follow up with them.

Question 7

We received three examples to the question regarding links to public APIs or institutional repositories that participants use to source data from:
https://kulturnav.org/
https://archive.org
https://museudaabolicao.acervos.museus.gov.br/acervo_museologico/?view_mode=cards&perpage=12&paged=1&order=ASC&orderby=date&fetch_only=thumbnail%2Ccreation_date%2Ctitle%2Cdescription&fetch_only_meta=

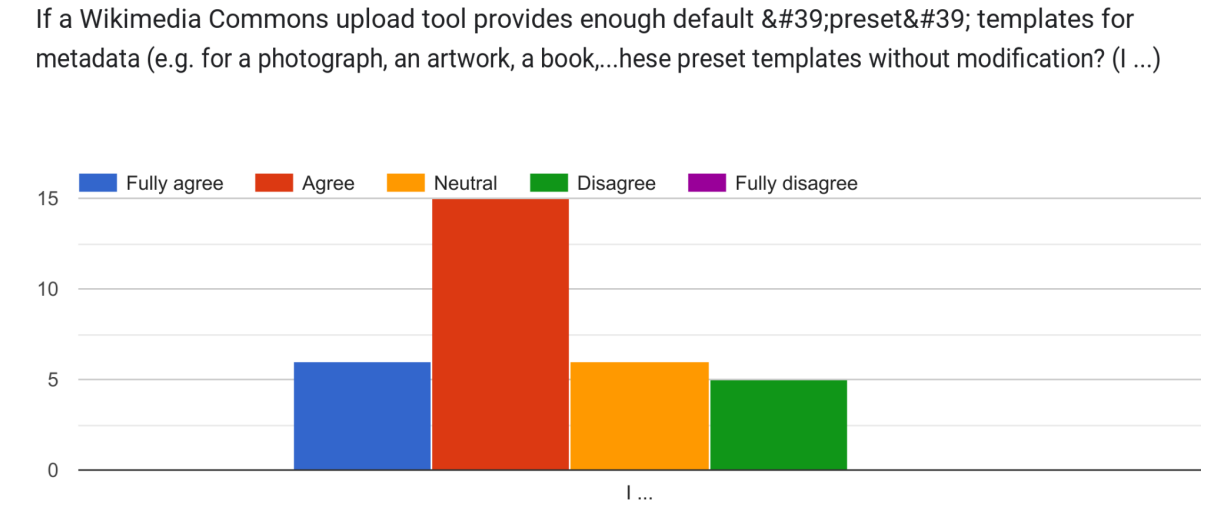
Some users could not share links publicly, so we can follow up with them.

Section 2 - Metadata

The following questions concern how participants want to engage with metadata and related templates on Wikimedia Commons. Only some of these are directly related to UX issues under the control of the OpenRefine tool suite. Some are more generally related to UX on Wikimedia Commons and will of most use to communicate directly to the WMF team that works on the Commons platform.

Question 8

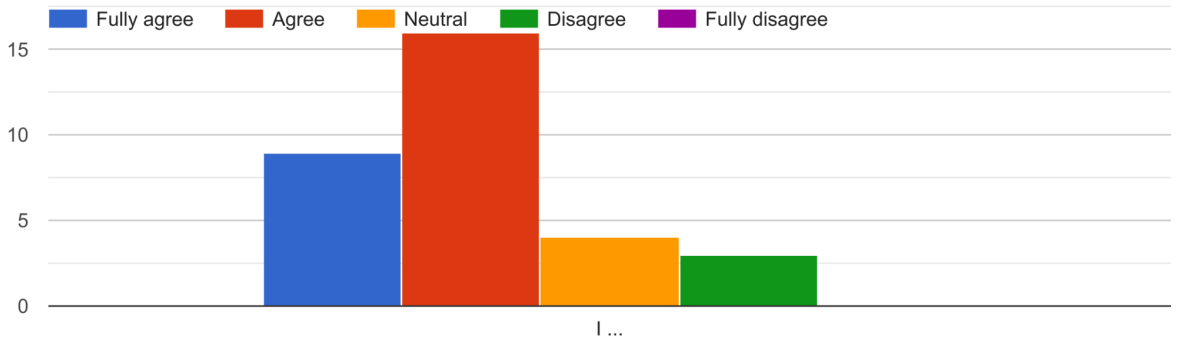
Most participants (21 votes in total) seem to agree that using a limited number of metadata templates without further modification would be enough to meet their needs. This indicates that our current plan for the first version of the SDC extension should meet the needs of at least 60-70% of use-cases.



Question 9

At the same time, most participants (25 votes in total) also agree that being able to add custom fields to a format structure is important to them. Indicating that a default template is not really the only type of metadata they plan to fill in. Still, this use-case will be met by the first version of the SDC extension as currently planned.

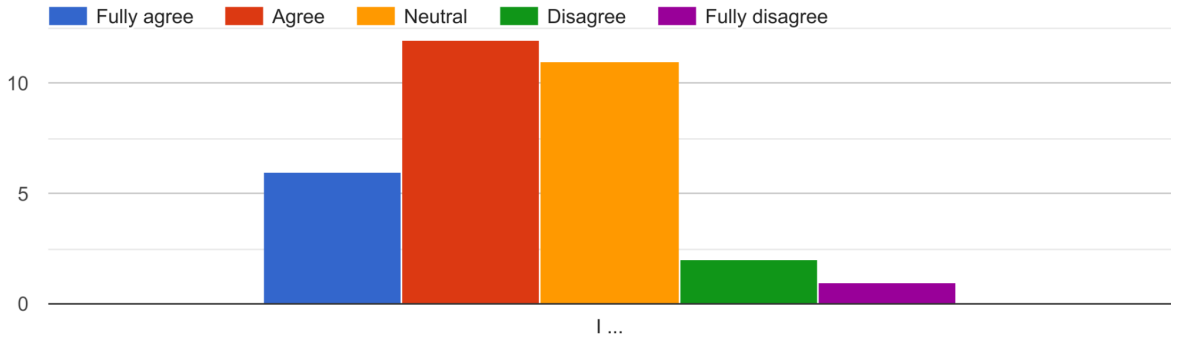
> Do you want to be able to control and customize the (structure of the) metadata that is uploaded and shown about a media file on Wikimedia Commons? ...at (e.g. a painting) that fits the content? (I ...)



Question 10

This question shows a nearly even split in opinion between the need to keep metadata “unmodified” in Commons vs allowing users to modify as they wish (the latter being the typical Wikimedian’s mental model). Many participants voted neutral in this question.

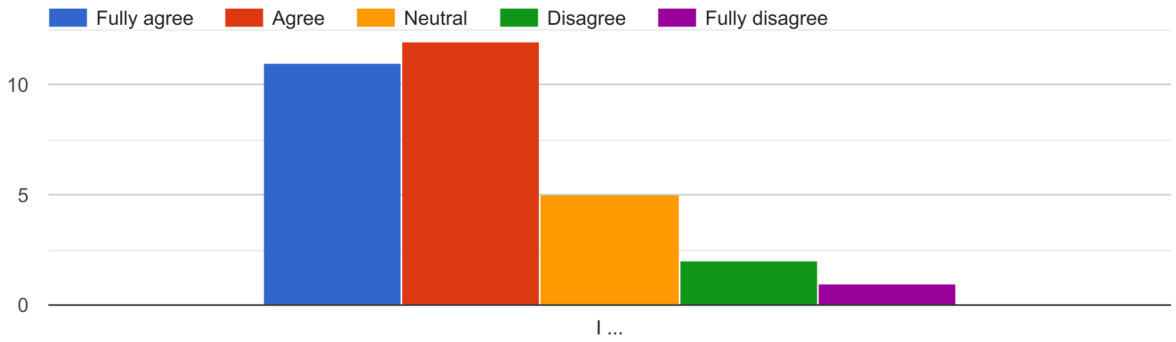
> Is it important to you that Wikimedians must clearly see that specific information (metadata) about your media files is as originally provided, and has not been modified? (I ...)



Question 11

This question shows overall strong agreement in terms of allowing users to modify metadata as they wish, when it comes to files uploaded outside an institutional context.

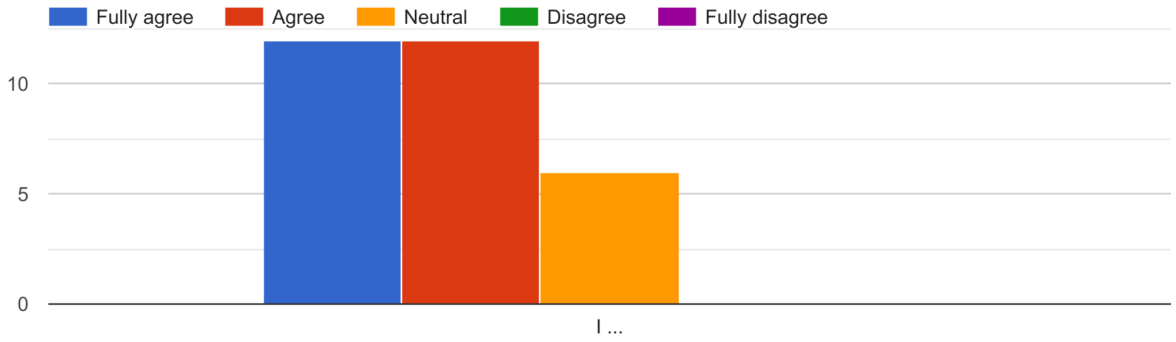
> Do you think that Wikimedians should be able to permanently and visibly modify, update and improve the metadata about media files on Commons that you have uploaded on your own behalf? (I ...)



Question 12

This question also shows overall agreement that the provenance of data – particularly when uploaded by an institution – is important to remain clearly visible.

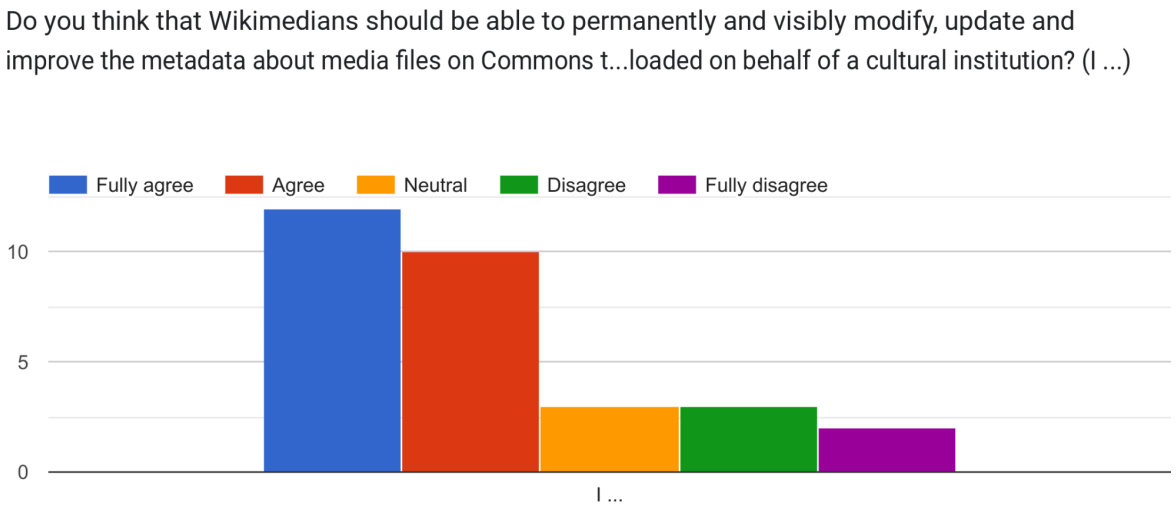
> If you are uploading or editing media files on behalf of a cultural institution, is it important that Wikimedians must clearly see which metadata has been provided by the institution? (I ...)



Question 13

Despite some moderate disagreement here, still most users generally agree that Wikimedians should be able to permanently change even data that has been uploaded by institutions. Staying in line with the general Wikimedian mental model that all data is

ultimately editable, and there are no authoritative single sources of truth that should remain immutable.



Final note

Several users expressed concerns over data privacy when using Google Forms; one user specifically expressed preference to use LimeSurvey in the future, so this is something we should definitely take under consideration. In addition, the option to upload files directly via the Google Forms seemed to require users to log-in via a Gmail account, prompting fears among some users that the form is not anonymous. I disabled this option subsequently, and we should probably not use it in the future – opting to get in touch with users separately.