In Finland (and Estonia), Wikimedia Finland has been batch uploading images from Finna.fi to Commons and, before that, to the crowdsourcing site Ajapaik.ee. Finna is a Finnish national library aggregator service for archives. To make this possible, we have indexed photos using image hashes to match images in the source repository to images already in the Wikimedia Commons.

Image hashes are identifiers where visually similar photos will get similar identifiers. One can use them to compare the similarity of pictures by checking how much the identifiers differ to detect duplicates, match photos in different services, find similar photos, etc. Longer explanation can be found from the link:

Our focus is on robust checks to determine if two images are identical. We have used image hashes to prevent duplicates when uploading files and to prevent the wrong photos from being updated when reuploading and updating metadata. For this, roughly 25M of 100M Wikimedia Commons images are indexed with image hashes, and we have a public API for the database in Toolforge.

In 2024, to make the data more widely accessible, we started gradually storing the hashes and Structured data on Commons. First, we added hashes photos from Finna photos. The following targets are photos from Estonia, Europeana, Flickr, and the rest. Another target is adding a function to pywikibot to check if the image is in Wikimedia Commons.

Overall target is to make an index for all Wikimedia Commons images and robust system for duplicate detection to prevent importing duplicate photos, making it easier for uploading photos and metadata from external repositories and for tracking how photos are used.

FinnaUploadBot is one of our bots if you want to see what we are currently doing.

Links:

- https://commons.wikimedia.org/wiki/User:FinnaUploadBot
- https://github.com/Wikimedia-Suomi/ImageHash-Toolforge
- https://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html

Short summary:

We are using imagehashes to create robust checks for determining if two images are identical to prevent duplicates when uploading files and to prevent the wrong photos from being updated when reuploading and updating metadata.