Visual Enrichment of Collaborative Knowledge Bases

Miriam Redi
& Special Guest Magnus!

WIKIMEDIA FOUNDATION
We Love Images

Photo by Dominik Scythe on Unsplash
Visual Language
perception, comprehension and production of visible signs

Images allow to explain, enrich and complement knowledge without language barriers

“Rain Gutter”
We Love Collaborative Knowledge Bases

Photo by Jebulon
Wikidata

Wikidata is an international and thus multilingual project. While English is the default interface language, the project is intended to be used by, and useful for, users of every language with MediaWiki internationalization support.
Wikidata already has the most images

https://docs.google.com/spreadsheets/d/1G2P89WW_GAkqShvaWdhkICN66WZB0synLCV9X0h6-lg/edit?usp=sharing
But Wikidata is still missing images!

ALL

- Entities without images: 94.7%
- Entities with images: 5.3%

PEOPLE

- People with images: 16.5%
- People without images: 83.5%

SPECIES

- Species with images: 7.9%
- Species without images: 92.1%
Smart Tools for **Wikidata Visual Enrichment**

Photo by [Jess Watters](https://unsplash.com) on [Unsplash](https://unsplash.com)
Visual Enrichment: Problem

- Worst-case scenario: users willing to add images to Wikidata might need to manually search for the right image using different tools from over 40M free-licensed commons images.
- We can help reducing the search space by
  - pulling images from reliable sources
  - using content-based analysis
Visual Enrichment: Pipeline

1. Discovering related images from different sources
2. Ranking images according to Relevance and Quality

Item without P18 - ‘Has image’
1. Finding Free-Licensed images
From: linked pages, Commons, Flickr
1. **Finding Free-Licensed images**
   Around 75% of people entities have images linked to them

There are so many of them.. How to reduce the amount of images users should select from?
2. **Ranking images**

Not all image candidates gathered are good candidates for the item.

**RELEVANCE:** does the images depict the entity?

**QUALITY:** is the image of high photographic quality?
2. **Ranking images** *(relevance)*

**RELEVANCE:** does the images depict the entity?

- Candidates are already biased towards relevance
- **Temporary solution:** match image name/desc with **item label** (plus face detection)
- **Content-based techniques** hard to use due to the singularity of each entity

**ONGOING:** semantic distance between item description and objects recognized *(word2vec)*
2. **Ranking images** (quality)

**QUALITY:** is the image of high photographic quality?

Not all relevant images are actually ‘good’ images

---

**High Quality**

Photo: [Jee & Rani Nature Photography](https://commons.wikimedia.org/wiki/File:Monarch_butterfly_on_flower.jpg) on Commons

**Lower Quality**

Photo: [Vinayaraj](https://commons.wikimedia.org/wiki/File:Monarch_butterfly_on_flower.jpg) on Commons
2. **Ranking images (quality)**

Solution: **COMPUTATIONAL AESTHETICS**

- Branch of Computer Vision
- Exploits visual information to detect quality
- Uses **Supervised Learning** Techniques
- Needs images **annotated** as being ‘High’ or ‘Low Quality’
2. **Ranking images (quality)**

**DATA:** exploiting the richness of the Commons

High Quality: 160K  Quality Commons

Lower Quality: 160K  Random Commons
2. **Ranking images** *(quality)*

**FRAMEWORK: Convolutional Neural Network** Google Inception-v3[1]

2. **Ranking images** *(quality)*

RESULTS: ACCURACY on 5-Fold Cross Validation

Adding side features:
- Image size
- Description Length
3. Combining Relevance and Quality

- **UNSUPERVISED**: Quality based re-ranking of the top X relevant images

- **SUPERVISED**: Joint learning to rank relevance and quality values
Offline Evaluation: The Distributed Game

DATA FOR EVALUATION

- 66 K ITEMS with 1+ image candidates
- User selection of the item image
- Classes:
  - 'person': Q5,
  - 'taxon': Q16521,
  - 'church': Q16970,
  - 'railway station': Q55488,
  - 'mountain': Q8502,
  - 'building': Q41176

https://tools.wmflabs.org/wikidata-game/distributed/#game=10
**Ranking Evaluation**

How good are the models proposed at ranking the selected image in the top-X of the candidate images?

73% of the times the selected image is in the top 3!
Tools for Enriching Wikidata

Magnus Manske
WikiShootMe

- Map-based tool
- Shows Wikidata, Wikipedia, Commons images, and other sources on a single map
- Allows direct upload/adding of images to Wikidata
- Mobile friendly

https://tools.wmflabs.org/wikishootme/
WD-FIST (WikiData Free Image Search Tool)

- Checks items without images
- SPARQL, Categories, lists, etc.
- Uses associated Wikipedia pages, Commons free-text search, GPS etc. to find candidate files
- One-click interface to set an image, two clicks for other media properties
- Can be pre-filled from other tools (e.g. WikiShootMe) or Wikipedia (e.g. Listeria lists)

https://tools.wmflabs.org/fist/wdfist
The Distributed Game

https://tools.wmflabs.org/wikidata-game/distributed/#game=10
Ongoing Online Evaluation: 
**File Candidates**

- Pre-computed file-to-item candidates
- “Expensive” queries (e.g. full-text searches)
- Multiple sources (Commons, Flickr, ...)
- Quick, unified display to user
- Specific topics
- Single-click transfer to Commons
- Single click to set as item image

---

https://tools.wmflabs.org/fist/file_candidates/
Back-port images to Wikipedia

Last update: Mon, 19 Feb 2018 03:47:04 +0000

Pages are listed if they have an image on Wikidata, but no "page image", as determined by Wikipedia.

Total: 42960 pages with image candidates.

Mensch

2. Yuto Nakamura (Nordischer Kombinierer) [Q48460502] : Summer Grand Prix Competition Planica
4. Nawaf Salam [Q638463] : AmbNawaf Salam addressing UN General Assembly.jpg [Commons]
5. Wu Changshuo [Q715400] : Wu-Chang-Shi.jpg [Commons]
7. Imru' al-Qais [Q1051776] : Hujjatul-Wada.jpg [Commons]
9. João I. (Kongo) [Q1522957] : Jean Roy de Congo.jpg [Commons]
10. Lê Thành Tông [Q1771050] : Lê Thành Tông.jpg [Commons]
11. Carl Wilhelm Walther [Q2057675] : Carl walther.jpg [Commons]
12. David Taylor (Snookerspieler) [Q2069811] : Alexhiggins1968.jpg [Commons]
15. Raniero Felice Simonetti [Q2363553] : CardinalRanieroSimonetti.jpg [Commons]
18. Frank Lovejoy [Q2557512] : In a Lonely Place - trailer - 10 - Frank Lovejoy.png [Commons]
19. Alison Skipworth [Q2837168] : Alison Skipworth in The Casino Murder Case trailer.jpg [Commons]

Pages on German Wikipedia where Wikidata has an image but Wikipedia does not

Using unused files

- Many great images already on Commons, some by donation
- Takes a category on Commons, and shows all files *not* used on any Wikipedia/Wikidata/etc.
- Might even **spark ideas for new articles or items**

https://tools.wmflabs.org/magnustools/unused_files.php
Flickr firehose

- Flickr version of “Recent Changes”
- Shows latest images on Flickr with a free license
- Upload button to Commons
- Gets us new images for the other tools

https://tools.wmflabs.org/magnustools/flickrfree.php
Finding Free-Licensed images

- Are Flickr images already on Commons?
- **Computer vision** can detect **exact and near-duplicate images**.
- In a small scale experiment, we found that only **0.1%** of these images are already on Commons.
Future Paths

- Improve machine learning models for image recommendation (help reducing the backlog!)
- Leverage efforts on Wikidata to make Wikipedia more visual
- Acquire more free-licensed visual data from external sources

Inspire readers and editors through images
Thank you :)
Visual Thinking