Structured Data on Wikimedia Commons and knowledge equity

Wikimania 2018, Cape Town
Etherpad

https://is.gd/sdequity

![QR Code](https://is.gd/sdequity)
What is happening?
When?
Upcoming work
Benefits
How can this advance knowledge equity?
Structured Data on Commons
2017-2019

adding metadata on Commons in a **structured & machine-readable** format making Commons files easier to view, search, edit, organize and re-use, in many languages
What are we doing?
Describing media files with Wikidata...
What we're working on

We worked / are working on
● Multi-content revisions
● Designs and prototypes for
  ○ file pages
  ○ UploadWizard
  ○ search
● Prepare for GLAM pilot projects

~ October 2018:
  multilingual captions for files

~ Early 2019:
● 'depicts' and other structured data
Summary

<table>
<thead>
<tr>
<th>Description</th>
<th>English: A wireframe sculpture of a bee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>2016-07-19</td>
</tr>
<tr>
<td>Source</td>
<td>Own work</td>
</tr>
<tr>
<td>Author</td>
<td>Wadzup</td>
</tr>
</tbody>
</table>

Licencing

I, the copyright holder of this work, hereby publish it under the following license:

This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International licence.

You are free:
- to share — to copy, distribute and transmit the work
- to remix — to adapt the work

Under the following conditions:
- attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that you endorse the work)
- share alike — You must license your new work under the same license

File history

Click on a date/time to view the file as it appeared at that time.

<table>
<thead>
<tr>
<th>Date/Time</th>
<th>Thumbnail</th>
<th>Dimensions</th>
<th>User</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>current</td>
<td>14:26, 19 July 2016</td>
<td>2,000 x 2,000 (636 KB)</td>
<td>Wadzup</td>
<td>contribs with UploadWizard</td>
</tr>
</tbody>
</table>

File usage

There are no pages that link to this file.

MediaInfo:Minigami

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>British English</td>
<td>No label defined</td>
<td>No description defined</td>
</tr>
<tr>
<td>English</td>
<td>File the structured data on Commons bee, an amigurumi</td>
<td></td>
</tr>
</tbody>
</table>

Statements

| Add statement |
Nighthawks is a 1942 oil on canvas painting by Edward Hopper that portrays people in a downtown diner late at night. It is widely considered Hopper's most famous work, and one of the most recognizable paintings in American art.
Metadata

**depicts**

- diner (Q1226)
- Greenwich Village (Q2053)

**material used**

- oil paint (Q6955)
- canvas (Q4269)
Upcoming, now and later in 2018-19

Properties for structured data on Commons

- Community consultation runs NOW. Participate!
- The basic properties that will be available for media files
Structured copyright and licenses

- Needed by services like search.creativecommons.org
- Powering better attribution and re-use
- Alignment with other structured copyright schemas, like rightsstatements.org
GLAM pilot projects

- Diverse and representative
- Support and documentation
Support for tool developers
The project Structured Data on Wikimedia Commons (2017–19) converts the free media files on Wikimedia Commons to a structured and machine-readable format, so that they become easier to view, search, edit, organize and re-use. To achieve that, the Commons backend is migrated to Wikibase, the same technology as used for Wikidata.

What is this?

Wikimedia Commons holds a lot of (meta)data about the media files it hosts. Structuring this data more and making it machine-readable has many benefits: it makes it easier to view, search, edit, curate, use and re-use the files on Commons.

The Structured Data on Commons project is an effort to make this happen. In early 2017, this project was funded by the Alfred P. Sloan Foundation. For more information about the grant, see the documentation here.


https://commons.wikimedia.org/wiki/Commons:Structured_data
Benefits for Wikimedia Commons?

(With some examples of what people are currently already doing with Wikidata...)
Make Commons multilingual...

... so that it can be searched in other languages than English

Marten van Valckenborgh: The Tower of Babel (1595), Gemäldegalerie Alte Meister, Public Domain

Henri Meunier: Au théâtre, before 1922, Public Domain
'tableware' in Japanese - テーブルウェア
A gilling machine, from Popular Science Monthly Volume 39, 1891. Unknown author, Public Domain
By becoming compatible with structured data used by other websites, we can be part of more discovery platforms.
Structured metadata and APIs to improve re-use
More value and potential of Commons' media files beyond Wikipedia

... where Wikidata enables multilingual search and browsing
Linked data - Wikimedia's data connected with other datasets around the world
tableware (Q851782)

plates used for setting a table and serving food.

External sources

- AAT: 300043196
- BabelNet: 00075834n
- BNCF Thesaurus: 566
- EAGLE: objtyp/lod/1780
- GND: 4124863-6
- Iconclass notation: 41C3d1
- JSTOR topic: tableware
- PACTOLS thesaurus: portM4tR8ZGTww
- PSH: 10251
- Quora topic: Dishes-and-Tableware
- US National Archives: 10842939

https://tools.wmflabs.org/reasonator/?q=851782
Wikidata's WikiProject
Sum of all Paintings...

All the museum collections which have paintings in them with the count per location
http://tinyurl.com/zqjz6q2
Querying & bringing together (thematic) information that may not have been combined before (across & also outside heritage collections!)

https://commons.wikimedia.org/wiki/Self-portraits_of_women
How can this advance knowledge equity?

(With some examples of what people are currently already doing with Wikidata...)
More potential for community improvements to media files

... even if they're minimally described when uploaded.
Better crowdsourcing with data that can be given back to an institution in a format that works for them.

**Birdwatching Rijksmuseum**

On Sunday October 4th, World Animal Day, the Rijksmuseum will host a birdwatching day in collaboration with Naturalis Biodiversity Center, Wikimedia Netherlands and the COMMIT SealincMedia project.

---

**Rijksmuseum – Naturalis Biodiversity Center – Wikimedia Nederland & COMMIT/ SEALINCMedia present a unique birdwatching event**

Birds are everywhere. In your own garden, in nature, and also in art. Among the Rijksmuseum’s 12 million collection objects are many prints, paintings and artefacts that have bird species depicted on them. Among the 37 million objects in the Naturalis collection are many birds from all over the world that have been collected in the last 200 years, as well as historical drawings of plants and animals in which many birds are depicted.

**Wikimedia Commons**

Some of the depicted birds are easily identified. Others require a trained eye to determine which species the artist has pictured. The Rijksmuseum and Naturalis are looking for experienced bird watchers and other avian enthusiasts to join an expedition through their digital collections and help the museums identify bird species in works of art.

The Rijksmuseum and Naturalis are currently in the process of donating large parts of their digitized collections of bird images to Wikimedia Commons, Wikipedia’s open multimedia library. Participants of the birdwatching day are challenged to collaboratively identify as many bird species as possible.
Multilinguality ... also for small languages!

What is depicted in images in the collections of the National Library of Wales? Wikidata query in English...
What is depicted in images in the collections of the National Library of Wales? Wikidata query in Welsh, French, Korean and Arabic
Equity in vocabularies
Actively add more diverse terms, people, places... to Wikimedia projects

Let's imagine and work together...
Annual photographic celebration of Africa’s cultures and diversity
65% of people are visual learners

Africa is the least covered region on Wikimedia projects
Wiki Loves Africa

- Annual photographic and media competition.
- Celebrates all that is quintessential about the African continent.
- Each year, submitted images illustrate a specific theme that highlights individual passions and daily encounters, collective idiosyncrasies and universal humanity.
- Run on a continental level with local events organised by participating usergroups and volunteers.
2014: Cuisine

2015: Cultural Fashion and Adornment

2016: Music and Dance

2017: People at Work

2019: Play!
2014: Cuisine

- 873 people contributed
- 6,116 photographs entered
- 2,420,791 page views
- 969 images used (15.44%) March '18

2015: Cultural Fashion and Adornment

- 722 people contributed
- 7,500 photographs entered
- 10,482,196 page views
- 609 images used (8.25%) March '18
2014: Music and Dance

- 836 people contributed
- 7,844 photographs entered
- 1,206,727 page views
- 319 images used (4.04%) March '18

2017: People at Work

- 2,473 people contributed
- 18,294 photographs entered
- 2,929,945 page views
- 1,124 images used (6.14%). March '18
Challenges

Majority of users are first time:

- **File names** are tricky
- **Descriptions** are difficult
- **Categorisation** is completely foreign and therefore usage is relatively low unless actively tasked by a member of the team (some kind of AI would definitely help here ;-) )
- Images are “dumped” with the country tying images together, but nothing else.
- **Cleaning up** what suits criteria is manual and normally done by one person.
- **Small teams** - very little local team input (confidence)
- **Photo Essays** - keeping them together and automatically representing
- Need to **curate** the process more ...

2017: People at Work
1st prize winner
• Berria newspaper now is publishing the web version under cc-by-sa.

• The have published more than 16,000 infographics in their history.

• Now they are uploading them to Commons, with the tags they use in their newspaper (i.e. Africa).
Everything is uploaded in svg format, so i18n is easier.

Lots of infographics are about very specific subjects.

(i.e. unemployment in the Basque Country in August 2015)

Some others are highly valuable for everyone.
- Tags and topics can benefit broadly using P180.
- Language translations can be easier once they start to be better organized.
- Infographics have text, let’s dream a future with automatic svg translation from Wikidata entities.
Digitized publications on Commons from the Indian Wikimedia communities
## Statistics of Indic Wikisource

- **11 languages**
- **123 readable books**

### Books/Validated Books updated on 26 Feb 2018

<table>
<thead>
<tr>
<th>Language</th>
<th>Index File/Book</th>
<th>Readable Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese[1]</td>
<td>14</td>
<td>00[2]</td>
</tr>
<tr>
<td>Malayalam[9]</td>
<td>242</td>
<td>00[10]</td>
</tr>
<tr>
<td>Odia [13]</td>
<td>54</td>
<td>02 [14]</td>
</tr>
<tr>
<td>Sanskrit[15]</td>
<td>166</td>
<td>00[16]</td>
</tr>
<tr>
<td>Tamil[17]</td>
<td>2169</td>
<td>30 [18]</td>
</tr>
<tr>
<td>Punjabi[21]</td>
<td>14</td>
<td>00[22]</td>
</tr>
</tbody>
</table>

(As of 26 Feb 2018)
Workflow
- Copyright free/PD books
- Digitization
- Wikimedia Commons
- Wikisource
  - OCR
  - Proofreading
  - Validating
Steps for Indexing the books on Wikisource
Structured data and Wikisource

- It will make the Indexing of the books on Wikisource more easier. Will save more time.
- Extract the data related to the work of the author/publisher
- Activate Listeria bot when a new file will be added to Commons it will auto-update the list.
<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Sports discipline</th>
<th>country for sport</th>
<th>Stmts</th>
<th>WM SL</th>
<th>IAAF</th>
<th>Wikidata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aila Saeed Mohammed</td>
<td>5000 metres 1500 metres</td>
<td>United Arab Emirates</td>
<td>19</td>
<td>6</td>
<td>235683</td>
<td>Q16211257</td>
</tr>
<tr>
<td>2</td>
<td>Annu Rani</td>
<td>javelin throw</td>
<td>India</td>
<td>16</td>
<td>6</td>
<td>256879</td>
<td>Q17403394</td>
</tr>
<tr>
<td>3</td>
<td>Anu Raghavan</td>
<td>400 metres hurdles</td>
<td>India</td>
<td>11</td>
<td>1</td>
<td>261260</td>
<td>Q33104653</td>
</tr>
<tr>
<td>4</td>
<td>Aya Ota</td>
<td>shot put</td>
<td>Japan</td>
<td>7</td>
<td>0</td>
<td>292528</td>
<td>Q33111803</td>
</tr>
<tr>
<td>5</td>
<td>Ayako Jinnouchi</td>
<td>1500 metres</td>
<td>Japan</td>
<td>16</td>
<td>4</td>
<td>204522</td>
<td>Q11658164</td>
</tr>
<tr>
<td>6</td>
<td>Ayako Kimura</td>
<td>100 metres hurdles</td>
<td>Japan</td>
<td>29</td>
<td>8</td>
<td>259312</td>
<td>Q4830847</td>
</tr>
<tr>
<td>7</td>
<td>Bùi Thị Thu Thảo</td>
<td>long jump</td>
<td>Vietnam</td>
<td>11</td>
<td>4</td>
<td>294015</td>
<td>Q32177640</td>
</tr>
<tr>
<td>8</td>
<td>Cecilia Yeung</td>
<td>high jump</td>
<td>Hong Kong</td>
<td>16</td>
<td>4</td>
<td>308354</td>
<td>Q29698185</td>
</tr>
<tr>
<td>9</td>
<td>Chen Qiaoqiu</td>
<td>pole vault</td>
<td>People’s Republic of China</td>
<td>9</td>
<td>2</td>
<td>298726</td>
<td>Q33104570</td>
</tr>
<tr>
<td>10</td>
<td>Chen Yang</td>
<td>discus throw</td>
<td>People’s Republic of China</td>
<td>19</td>
<td>7</td>
<td>256869</td>
<td>Q26404027</td>
</tr>
<tr>
<td>11</td>
<td>Darya Maslova</td>
<td>5000 metres 10,000 metres</td>
<td>Kyrgyzstan</td>
<td>17</td>
<td>7</td>
<td>291067</td>
<td>Q20745026</td>
</tr>
<tr>
<td>12</td>
<td>Debashree Mazumdar</td>
<td>4 x 400 metres relay</td>
<td>India</td>
<td>16</td>
<td>6</td>
<td>271416</td>
<td>Q26265953</td>
</tr>
<tr>
<td>13</td>
<td>Dihani Lekamge</td>
<td>javelin throw</td>
<td>Sri Lanka</td>
<td>6</td>
<td>0</td>
<td>235683</td>
<td>Q33112318</td>
</tr>
<tr>
<td>14</td>
<td>Dutee Chand</td>
<td>100 metres 200 metres 4 x 100 metres relay</td>
<td>India</td>
<td>24</td>
<td>17</td>
<td>279590</td>
<td>Q15020846</td>
</tr>
</tbody>
</table>

- Extract the data as you want to show
- Customise it according to the interest of local readers.
- Digital catalogue.

Images from State Government of Odisha
Enkosi!