The_Tower_of_Babel.jpg: Diversity of Visual Encyclopedic Knowledge Across Wikipedia Language Editions

Shiqing He
University of Michigan
Ann Arbor, MI 48109

Allen Yilun Lin
Northwestern University
Evanston, IL 60208

Eytan Adar
University of Michigan
Ann Arbor, MI 48109

Brent Hecht
Northwestern University
Evanston, IL 60208

Abstract

Across all Wikipedia language editions, millions of images augment text in critical ways. This visual encyclopedic knowledge is an important form of wikiwork for editors, a critical part of reader experience, an emerging resource for machine learning, and a lens into cultural differences. However, Wikipedia research–and cross-language edition Wikipedia research in particular–has thus far been limited to text. In this paper, we assess the diversity of visual encyclopedic knowledge across 25 language editions and compare our findings to those reported for textual content. Unlike text, translation in images is largely unnecessary. Additionally, the Wikimedia Foundation, through the Wikipedia Commons, has taken steps to simplify cross-language image sharing. While we may expect that these factors would *reduce* image diversity, we find that cross-language image diversity rivals, and often *exceeds*, that found in text. We find that diversity varies between language pairs and content types, but that many images are unique to different language editions. Our findings have implications for readers (in what imagery they see), for editors (in deciding what images to use), for researchers (who study cultural variations), and for machine learning developers (who use Wikipedia for training models).

In the past decade, the computing community has demonstrated the importance of understanding the similarities and differences between Wikipedia language editions (e.g., (Pfeil, Zaphiris, and Ang 2006; Adafre and De Rijke 2006; Hecht and Gergle 2009; 2010; Callahan and Herring 2011; Bao et al. 2012; Hecht 2013)).This research has shown that Wikipedia language editions represent encyclopedic knowledge in highly diverse ways and that this diversity has substantial effects on both (1) human readers and (2) Wikipedia-based artificial intelligence systems.

Research has shown that readers around the world gain a different understanding of concepts by reading different language editions, due in part to each edition's "cultural contextualization" of concepts (Hecht and Gergle 2010; Hecht 2013). Additional work has identified that a large number of AI systems that use Wikipedia as world knowledge ingest the cultural perspectives of the language edition on which they are operating. Indeed, very early evidence of what is now known as "algorithmic bias" (e.g., (ACM US Public Policy and Europe Councils 2017; Angwin et al. 2016; Knight 2017)) was first observed by examining the same AI technology operating on different Wikipedia language editions (Hecht and Gergle 2010).

Research on language edition *diversity* has focused on comparing article text and links across language editions. However, the encyclopedic knowledge in Wikipedia is multimedia in nature and, in addition to text and links, consists of images and other media. In this paper, we seek to extend the literature on Wikipedia language edition diversity to the critical medium of *imagery*. More formally, this paper aims to measure and characterize the similarities and differences in *visual encyclopedia knowledge* across language editions.

Recent work on Wikipedia and other user-generated content domains suggests that images should be associated with significantly *less* diversity, as, (1) images require no lexical translation (e.g., Hale, 2012) and (2) due to Wikimedia Commons, a shared resource for images (and other files). On the other hand, work in cultural psychology, design and related fields suggests that different cultural groups (including different "language-defined cultures") might make different choices when it comes to visual representations of concepts (Hecht and Gergle 2010).

In this paper, we focus on images to explore this seeming divergence in the literature by asking two research questions: **"What is the *diversity of visual encyclopedic knowledge* across language editions of Wikipedia?"** In doing so, we seek to first understand whether the language editions use the same or different images in their articles (RQ1a: *Language edition-level diversity*). Critically, we also investigate the extent to which two articles in different languages about the same concept tend to use the same or different imagery (RQ1b: *Within-concept diversity*). For instance, do all articles about the concept known in English as "Chocolate" use the same set of pictures to represent and describe Chocolate? Our second research question directly targets the opposing hypotheses in the literature. We ask (RQ2): **"How does the diversity in visual encyclopedic knowledge compare to the diversity of textual encyclopedia knowledge?"** Here, we compare our findings from RQ1 to those of Hecht (2013) on textual diversity in 25 language editions of Wikipedia.

To address these questions, we analyzed millions of images across the Wikipedia Commons and across 25 large language editions. In addition, we constructed a user-friendly system, *WikiImgDive*, to support qualitative analysis and exploration of image usage diversity across languages. This system provides multiple perspectives on diversity for concepts across languages. We offer the tool at http://whatsincommons.info/icwsm/ for interested readers to interactively engage with our data and results.

Overall, our results suggest that there is extensive diversity in visual encyclopedic knowledge across the language editions. Language editions tend to use highly diverse sets of images overall, with over 67% of images appearing in only one language edition and only 142 appeared in all 25 editions in our study. Moreover, we see that the same concept is often represented (visually) quite differently. Our results suggest that, on average, a person viewing a concept in one edition will see a relatively high percentage of unique imagery compared to a person viewing the same concept in a different edition. Indeed, for most language edition pairs, the average overlap for the same concept is well under 50%, and the maximum value for two editions is only 63.6%.

Our analyses also lead to a clear conclusion for our second research question (RQ2) which contrasts textual diversity to image diversity. Here we find that the visual diversity is, on average, greater than the observed textual diversity when considering two articles in different language editions describing the same concept.

Our results have implications for a number of constituencies. For readers, our findings suggest that two Wikipedia readers who speak different languages will see substantially different visual representations for the same concept. This may argue for additional tools and mechanisms that surface these differences. For Wikipedia editors, one implication of our results is the need for new sociotechnical mechanisms that would allow for more cross-language edition image sharing (if editors wish to engage in more of such sharing). Finally, more and more AI systems are consuming visual encyclopedic knowledge from Wikipedia in addition to textual knowledge, a trend that will likely accelerate thanks to advances in computer vision technology and major new initiatives (Wikimedia Foundation 2017). Our results suggest that these systems will be impacted as much or more by algorithmic bias as those that consume textual knowledge.

Related Work

Text Diversity in Wikipedia

Previous research on Wikipedia content diversity has only focused on text. This body of work has collectively demonstrated that textual content about the same concepts is highly diverse across different language editions (Adafre and De Rijke 2006; Pfeil, Zaphiris, and Ang 2006; Hecht and Gergle 2009; 2010; Callahan and Herring 2011; Bao et al. 2012; Hecht 2013). For example, Adafre and de Rijke (2006) showed that very few similar sentences (measured by word overlaps in machine-translated text) could be found on pages about the same concept (e.g., "Rice Pudding") in different language editions. Similarly, Hecht and Gergle (2010) and Bao et al.'s work (2012) identified (and visualized) a large number of unique named entities when comparing pages about the same concept in different languages. On a semantic level, Hecht and Gergle (2009) showed that all Wikipedia language editions have self-focus bias in that (1) the concepts that are covered tend to be those that are of interest to the corresponding language-defined culture and (2) concepts that are covered in multiple languages are covered in a fashion that tends to emphasize their relationship to the corresponding language-defined culture.

Some research has highlighted the effect of this encyclopedic knowledge diversity on the many algorithms that rely on Wikipedia content. For instance, early reflection on "algorithmic bias", showed that well-known Wikipedia-based semantic relatedness measures (e.g., (Gabrilovich and Markovitch 2007; Milne and Witten 2008)) adopt the cultural perspectives of the language edition they are using as world knowledge, outputting different results depending on the language edition (Hecht and Gergle 2010; Hecht 2013).

Hecht's work (Hecht 2013; Hecht and Gergle 2010; Bao et al. 2012) provided the most comprehensive evaluation of textual content diversity on Wikipedia language editions and theorized this phenomenon as *cultural contextualization*. He designed a methodological framework (detailed below) that allows systematic evaluation of the text diversity in the top 25 language editions. Our work adapts this methodological framework to comprehensively analyze the 25 editions for image diversity. Utilizing the same set allows us to compare our image diversity results with their text counterparts.

Images on Wikipedia

Imagery on Wikipedia has largely been ignored in previous research. The few studies that have examined Wikipedia images focused on their use for building image databases (Tsikrika, Kludas, and Popescu 2012). The intended target for these databases is in training and testing image retrieval algorithms (e.g., (Lau et al. 2006; Tsikrika, Popescu, and Kludas 2011; Aletras and Stevenson 2013; Benavent et al. 2013)), making these algorithms vulnerable to biases as discussed above. The most prominent research that centered on Wikipedia's images themselves is from Viegas (2007). Viegas studied the peer-production process of Wikipedia images and found that the collaborations around images are quite distinct from those that tend to occur around textual content. Rather than image-related editing, our work investigates the **diverse (or not) usage** of Wikipedia images across language editions.

Image Diversity Versus Text Diversity

Whether one might expect more or less diversity in visual encyclopedic knowledge is unclear. One important development related to images and Wikipedia is the growth in prominence of the Wikimedia Commons, an image (and file) hosting site that can be used by any language edition and hosts over 46 million files (as of April, 2018). Though some policies (e.g., generating category names) are English-focused, the Commons is intended to be multilingual. This, in theory, should afford more cross-language image sharing, thereby reducing diversity. Many language editions have adopted the policy of favoring the Commons as the media file host, though this is neither universal nor applied rigorously within language editions. This leads to the hosting structure illustrated in Figure 1. Additionally, tools such as

Google Translate may aid in translation work (e.g., for captions and filenames). Research on online communities has found that translation work may be more common than previously thought (Hale 2015). The affordances of the Commons in driving cross-language sharing may also be bolstered by the fact that *visual* knowledge does not often require active translation (with the notable exception of images that contain text, such as information visualizations). Images should, therefore, be easy to "translate" into another language edition (Hale 2012).

The factors described above would appear to predict *less* diversity: images are easier to copy/translate, and Wikipedia encourages centralization through the Commons. Indeed, in a particularly illustrative instance, a single editor inserted the same photo of a train across hundreds of language editions' articles for "Train." Over time, some editors changed this photo, but it remains in over 100 articles.

On the other hand, experience with cultural preferences for certain imagery might lead to more diversity. Work on cross-cultural design for websites has demonstrated different color and layout preferences in different countries (Cyr, Head, and Larios 2010; Reinecke and Bernstein 2011). Extensive work on cultural psychology (Kitayama and Cohen 2010) has also demonstrated different cultural preferences for imagery (Miyamoto, Nisbett, and Masuda 2006). From the perspective of Wikipedia, while the Commons may subtly encourage image sharing, it certainly does not require this as a policy. Anyone can add an image as long as it satisfies the criteria that it "could be used" for educational purposes (Wikimedia Foundation 2018). Thus, there are often many suitable images from which a local editor could pick (e.g., there are over 235 images in the topmost "Chocolate" category in the Commons). As such, cultural preferences and the diverse pool of *available* images may suggest extensive image diversity.

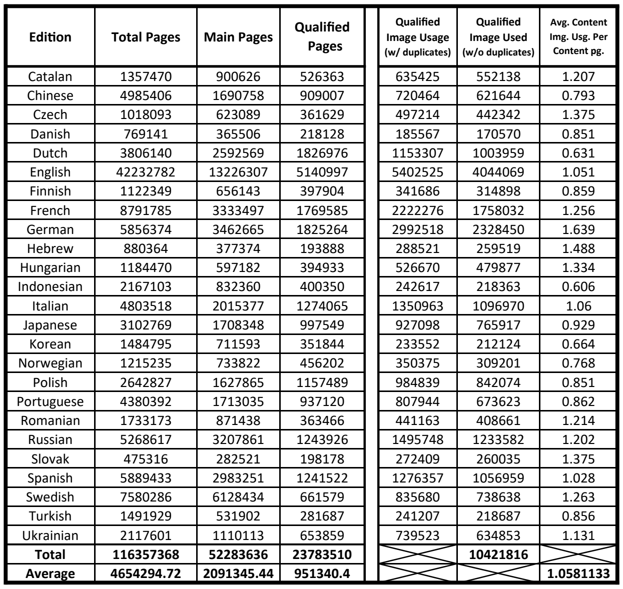
Put together, we have two well-motivated but opposing hypotheses as to the relationship between text and image diversity. Cultural differences may lead to similar (or more) diversity in image use. Conversely, given the ease of sharing images and the specific socio-technical innovations of the Commons, image diversity may be far less than that for text. Our analyses help determine for which hypothesis there is more empirical support.

Methodology

Datasets

We utilized the publicly available Wikipedia data dumps (WikiMedia-Meta-wiki 2017) from June 1st, 2017. To enable the comparison between the diversity of *visual* encyclopedic knowledge and *textual* encyclopedic knowledge, we selected 25 language editions to match the languages selected in (Hecht 2013) (see Table 1 for a complete list). These editions correspond roughly to the 25 largest language editions in 2009 (Hecht and Gergle 2009; Hecht 2013) by number of articles. Most of these editions remain in the top 25 in this respect.

For each of the 25, we collected article text and referenced images. We filtered articles to include only (a) those in the main article namespace (in contrast to discussion pages, user pages, etc.) and (b) only "content" articles (in contrast to disambiguation and redirect pages).



Edition	Total Pages	Main Pages	Qualified Pages	Qualified Image Usage (w/ duplicates)	Qualified Image Used (w/o duplicates)	Avg. Content Img. Usg. Per Content pg.
Catalan	1357470	900626	526363	635425	552138	1.207
Chinese	4985406	1690758	909007	720464	621644	0.793
Czech	1018093	623089	361629	497214	442342	1.375
Danish	769141	365506	218128	185567	170570	0.851
Dutch	3806140	2592569	1826976	1153307	1003959	0.631
English	42232782	13226307	5140997	5402525	4044069	1.051
Finnish	1122349	656143	397904	341686	314898	0.859
French	8791785	3333497	1769585	2222276	1758032	1.256
German	5856374	3462665	1825264	2992518	2328450	1.639
Hebrew	880364	377374	193888	288521	259519	1.488
Hungarian	1184470	597182	394933	526670	479877	1.334
Indonesian	2167103	832360	400350	242617	218363	0.606
Italian	4803518	2015377	1274065	1350963	1096970	1.06
Japanese	3102769	1708348	997549	927098	765917	0.929
Korean	1484795	711593	351844	233552	212124	0.664
Norwegian	1215235	733822	456202	350375	309201	0.768
Polish	2642827	1627865	1157489	984839	842074	0.851
Portuguese	4380392	1713035	937120	807944	673623	0.862
Romanian	1733173	871438	363466	441163	408661	1.214
Russian	5268617	3207861	1243926	1495748	1233582	1.202
Slovak	475316	282521	198178	272409	260035	1.375
Spanish	5889433	2983251	1241522	1276357	1056959	1.028
Swedish	7580286	6128434	661579	835680	738638	1.263
Turkish	1491929	531902	281687	241207	218687	0.856
Ukrainian	2117601	1110113	653859	739523	634853	1.131
Total	**116357368**	**52283636**	**23783510**		**10421816**	
Average	**4654294.72**	**2091345.44**	**951340.4**			**1.0581133**

Table 1: Basic descriptive statistics for our datasets.

Comparing the 2012 size rankings from Hecht (2013) and our 2017 size rankings, we observed that Swedish had grown massively and disproportionately. It is now 11 times larger than it was in 2012. Investigating the cause of this growth, we found that it is mostly due to bot-created articles (e.g., the *Lsjbot* is responsible for over half of Swedish Wikipedia's articles). To prevent this signal from overwhelming those from more typical editing behaviors, we sought to remove these pages from consideration. The procedure we used was as follows: we found the Swedish Wikipedia category listing bots ("Kategori:Wikipedia:Robotar"), then removed all pages listed under the sub-category "Robotskapade artiklar", or "bot-created articles."

We define the term "**qualified pages**" as mainspace pages that are not redirect or disambiguation pages (or Swedish bot-created pages). As is shown in Table 1, the English edition has the largest number of qualified pages (around 5 million articles), followed by Dutch, German, French and Italian. Hebrew is the smallest edition in terms of qualified pages with around 190,000 qualified pages. In total, the filtering procedure described above resulted in about 24 million qualified pages. This collection forms our final *article* corpus, Wiki25Articles.

To connect articles in different languages about the same concept, we follow a standard practice in the literature (e.g. (Sen, Li, and Hecht 2014)) and turn to the article-to-concept mappings maintained in Wikidata (the structured data project that is closely affiliated with Wikipedia). These mappings make it trivial to identify that, for instance, "Chocolate" (English), "Schokolade" (German), and "Chocolat" (French) all describe the same concept. Figure 1 provides a visual overview of how this mapping works.

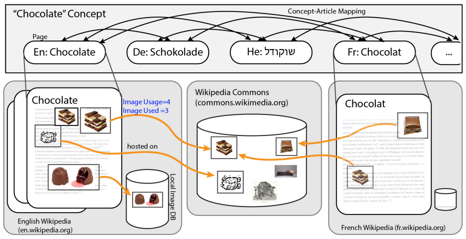


Figure 1: Examples of concept-article connections and the structure of image sharing. A concept (*"Chocolate"*), is represented by language-specific pages in different editions.

We used the version of this mappings that are manifest in the *languagelink* table in the Wikipedia dump, which also includes some legacy "interlanguage links," an earlier data structure for concept mapping in Wikipedia. To avoid some concept alignment issues (Bao et al. 2012; Hecht 2013), in the aligning stage, we removed "partial" links that link an article to a section of another article. For example, the English "Luke Skywalker" page has a mapping to the Luke Skywalker section of the German page "Figuren aus Star Wars." However, other alignment issues may remain (Bao et al. 2012), and this is an important direction of future work for the entire multilingual Wikipedia literature.

Template Filtering

We extracted image usage information from the Wikipedia database *imagelink*, which maps images to the articles that are linked to them (regardless of whether the images are hosted in the Commons or not). This database holds images explicitly added by editors to an article but also those added via *templates*. A template is "a Wikipedia page created to be included in other pages (Wikipedia-English 2012)." Templates usually contain repetitive material that might need to show up on any number of articles or pages (e.g., boilerplate messages, standard warnings or notices, infoboxes, and navigational boxes (Wikipedia-English 2012)). Templates include both administrative images as well as content-specific images. Administrative images are, for example, the icon used to designated articles that have achieved "Featured Article" quality status. Many of the images that appear on all 25 language editions were icons used in administrative templates. Content-specific images in templates vary widely, but one example comes from the page for the Canadian politician "John Grieve" (English) that uses the template *Liberal-Ontario-MPP-stub*. This template adds the flag of Ontario to every page that uses it. Other examples include an overview map (e.g., of the United States) that might appear on every page about a place in a given region (e.g., a state).

Template images presented a conceptual challenge for our research: are they part of each language editions' visual encyclopedic knowledge or not? This was exacerbated by the fact that template image usage is extraordinarily common, meaning trends in template image usage would overwhelm trends in manually-added imagery. Faced with a similar challenge in their analysis of links, Hecht and Gergle (Hecht and Gergle 2010) removed links in templates, and we followed their example in this study. Furthermore, in follow up work, Hecht (2013) considered both types of links and found few meaningful differences in overall trends).

We filter template images using a two-stage filter. First, images that are explicitly referenced in template pages are captured and removed. A portion of the remaining images, while not explicitly in templates, had template-like characteristics. For example, an "undo" icon was copied 295 times in the Turkish edition as part of a table on soccer teams (to indicate a player leaving the team that season). While the number of "implicit template images" appears low, they are used frequently and distort our statistics.

To remove these images, we applied a second filter based on frequency of use. A threshold was selected using data from the explicit template filter. Observing the frequency distribution for images caught by the explicit filter, we set a threshold of 85%. For example, in the English edition, 85% of explicit template images appear 22 times or more. In the Russian edition, 85% of explicit template images appear 17 times or more. We filter out images that are used more than this threshold. Figure 2 provides examples of removed and retained images. We refer to images that remain after this second stage as "**qualified images**."

To evaluate this strategy, we sampled 40 images for manual evaluation. In an initial analysis by two authors, we found a high Cohen's Kappa (.95) for determining whether an image was a template image (regardless of whether it was implicit or explicit). Given this agreement, one researcher analyzed 200 images (half labeled as templates by the classifier and half as qualified images) and determined the template filter accuracy of our approach to be 96%.

Table 1 summarizes the results of the filtering process. In total, there are 10.4 million qualified images within our dataset (WIKI25IMAGES). English has the largest number of qualified images. German has the highest image-per-page ratio. On average, every qualified German article uses 1.64 qualified images. The Indonesian Wikipedia has the lowest average image per-page-ratio (0.606). There are notable differences in this "image concentration" ratio across language editions, hinting at diverse image selection preferences.

Coding with *WikiImgDive*

In addition to our large datasets, we also constructed a number of tools to support qualitative coding and detailed inspection of cross-language image use. These tools, which we combined into *WikiImgDive*(Figure 3), are available at `http://whatsincommons.info/icwsm/`. In addition to supporting our own research, *WikiImgDive* was designed to be a user-friendly way to explore our dataset and results on a concept-by-concept basis.[1] Specific features include the ability to visualize images used to describe a specific Wikipedia concept (e.g., Chocolate) in the 25 selected editions. In our research, we used this tool to analyze concepts to determine which images are used uniquely, or by many editions. Images are displayed in a sortable grid (e.g.,

[1] *WikiImgDive* current uses direct calls to Wikimedia's API, which structures data differently than the raw data used in our analyses. Full documentation is available on our webpage.

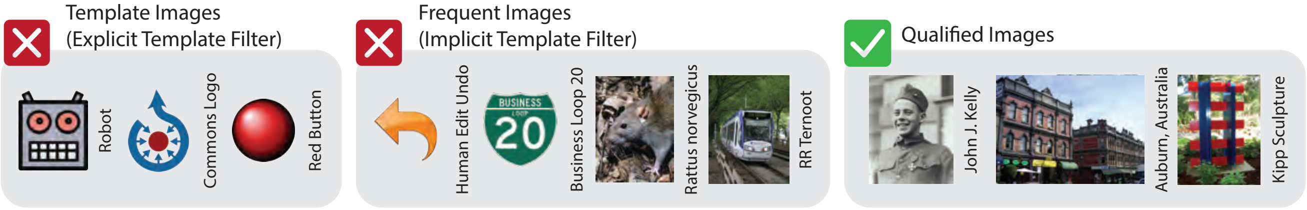
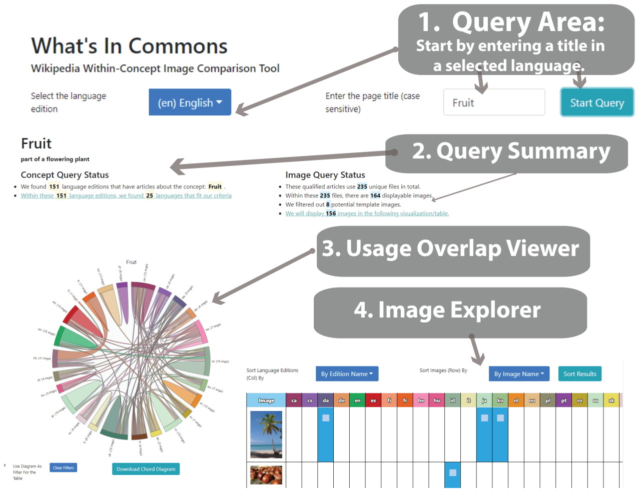


Figure 2: Examples of our two-stage template filter. Icons used in explicit templates (left panel) are removed. Frequent images (middle panel) are also removed. In this case, the "human edit icon," and the rat were manually copied in tables (of soccer-related tables in Turkish and species lists in Ukrainian). The road sign was manually added to English articles about Texas highways and the train was used in a gallery duplicated for multiple train stations in the Netherlands. The rightmost panel contains examples of qualified images.



What's In Commons

Wikipedia Within-Concept Image Comparison Tool

Select the language edition: (en) English

Enter the page title (case sensitive): Fruit

Fruit

part of a flowering plant

Concept Query Status

- We found 151 language editions that have articles about the concept: Fruit .
- Within these 151 language editions, we found 25 languages that fit our criteria

Image Query Status

- These qualified articles use 235 unique files in total.
- Within these 235 files, there are 164 displayable images.
- We filtered out 8 potential template images.
- We will display 156 images in the following visualization/table.

Figure 3: Overview of *WikiImgDive*.

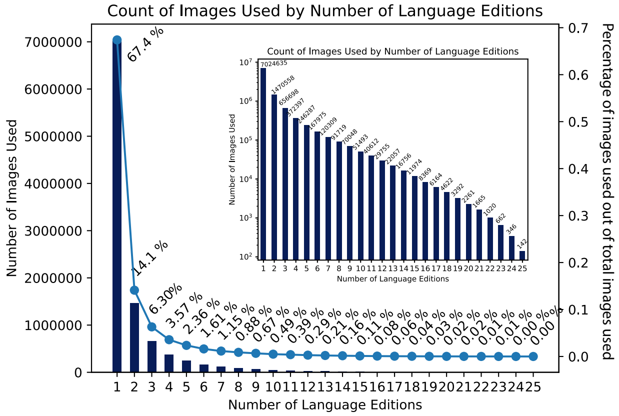


Figure 4: The distribution of image counts by the number of language editions in which they are used. The y-axis on the right is used for a line chart that shows the percentage of images that are used in x number of editions. The inset plot is the same figure, but with a logged y-axis.

one can sort by language editions with the most images about a specific concept, or images that are most often used). *WikiImgDive* also presents a chord diagram that provides a high-level view of image use overlap for a particular concept.

Results

We present our results by first focusing on the images used across *entire language editions* (RQ1a). We then discuss the amount of visual encyclopedic diversity at the within-concept level (RQ1b) and compare our image diversity results to those of text (RQ2).

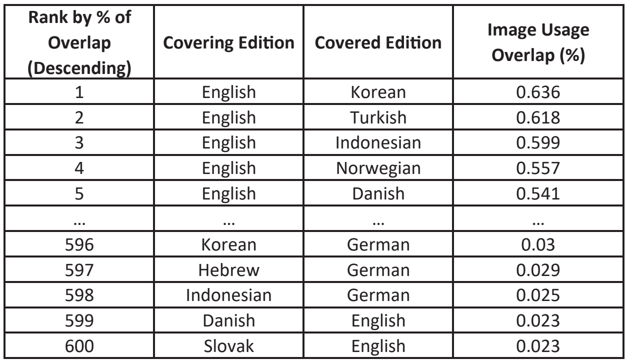
Language Edition-Level Image Diversity (RQ1a)

Figure 4 shows the image coverage distribution across all 10.4 million images. In their paper, Hecht and Gergle referred to concepts that only have articles in one of the 25 editions they studied as "single-language concepts." Analogously, Figure 4 (leftmost point) shows that "single-language images" make up 67.4% of all images. That is, over 67% of images appear in only one of the 25.

As the number of editions increases, the image coverage decreases quickly. Only 14.1% of images appear in 2 language editions, and 0.016% of images appear in 21 editions. Cumulatively, only 6.25% of images appear in more than five editions. The inset in Figure 4 displays the same information with logged y-axis. In analogy to Hecht and Gergle's "global concepts" (which defines global in terms of the 25 language editions), we find only 142 "global images" used across all 25 examined editions.

To get a sense of the content of global images (what few exist), we conducted a qualitative coding exercise. We used an affinity diagramming approach in which we 1) went through each global image and labeled it with a description of its content and 2) grouped similar images to come up with common themes. We found 5 major groups: *Portrait/people* (53%); *Icon* (19%); *Landscape* (9%); *Object* (6%); *Map* (6%); *Book* (4%); *Creature* (3%). The Icon group contained non-template-introduced icons that had relatively low usage count within any editions (effectively, this represents a large portion of the small amount of noise in our template filter).

Overall, the high percentage of single-language images and very low percentage of global images indicate that ex-



Rank by % of Overlap (Descending)	Covering Edition	Covered Edition	Image Usage Overlap (%)
1	English	Korean	0.636
2	English	Turkish	0.618
3	English	Indonesian	0.599
4	English	Norwegian	0.557
5	English	Danish	0.541
…	…	…	…
596	Korean	German	0.03
597	Hebrew	German	0.029
598	Indonesian	German	0.025
599	Danish	English	0.023
600	Slovak	English	0.023

Table 2: Edition pairs with most and least image overlap (%)

tensive visual encyclopedic knowledge diversity is present in Wikipedia at the language edition level.

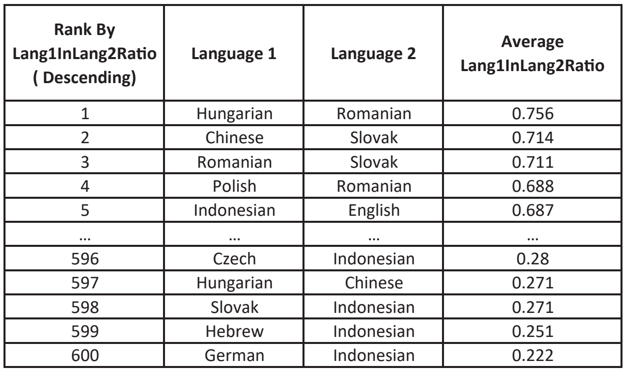
Pairwise Comparisons Between Languages: Examining image usage through a pairwise comparison framework can reveal interesting patterns in the diversity observed above. Using French and German as an example pair for this analysis, we took the set of images used in each of these two editions and calculated the image overlap between these sets, i.e., what percent of images used in French are also used in German, and vice versa.

Table 2 displays the language edition pairs with the highest and lowest pairwise overlap percentages. The table reveals that of all 600 language edition pairs (25 × 24), the most overlap between any two pairs is 63.6 % (the percentage of images used in the Korean that are also used in English). This means that, given any two language editions, at least 36.4% of the images are unique to one of the language editions. More generally, the English Wikipedia has high coverage of other language editions, likely due to its size. The same is true of other large language editions (in terms of qualified count), such as Dutch, German, and French. Small editions such as Hebrew, Slovak, Danish and Turkish have relatively low coverage of other language editions.[2]

Image Diversity Within Concepts (RQ1b)

The analyses above address the question "what images are used or not used in each language edition?" (i.e., RQ1a). In this section, we present within-concept diversity analyses that address the question "how are images used to describe the same concept in different languages?" (i.e., RQ1b). In other words, we compare how a concept (e.g., Chocolate in English) is visually represented in each language edition in which it has an article (e.g., Chocolat (French), Schokolade (German), and so on).

Overall, we identified 15 million qualified pages with at least one corresponding page in another language. Within this corpus, we filtered out page pairs in which at least one page does not have any qualified image. After the filtering process, 7.6 million page pairs remained in our dataset.



Rank By Lang1InLang2Ratio (Descending)	Language 1	Language 2	Average Lang1InLang2Ratio
1	Hungarian	Romanian	0.756
2	Chinese	Slovak	0.714
3	Romanian	Slovak	0.711
4	Polish	Romanian	0.688
5	Indonesian	English	0.687
…	…	…	…
596	Czech	Indonesian	0.28
597	Hungarian	Chinese	0.271
598	Slovak	Indonesian	0.271
599	Hebrew	Indonesian	0.251
600	German	Indonesian	0.222

Table 3: The maximum and minimum pairwise RL_1L_2 values. For example, on average, 75.6% of images used in the Hungarian Wikipedia are also used in the same-concept article from the Romanian Wikipedia.

To represent the image overlap between page pairs, we adopt (and adjust for images) the *RatioOfLang1InLang2* (RL_1L_2) metric from Hecht (2013). RL_1L_2 measures the overlap between two articles in two language editions about the same concept. Whereas Hecht used this metric to compare "bags of links" between two pages, we use it to compare the images used on two pages. Adapted to our image-specific context, RL_1L_2 is defined as follows:

$$\begin{aligned}&\text{RatioOfLang1InLang2}_{img}\\&=\frac{\text{images of lang1}\cap\text{images of lang2}}{\text{lang1 images}}\end{aligned}\tag{1}$$

To better understand how RL_1L_2 works, let us consider a French article (*Lang1*) and an English article *Lang2*, both about some concept *A* (*Á* in French). In our example, the English article has three images and the French article uses only two images, and one appears on both pages. In this hypothetical scenario, the French-English pair has the RL_1L_2 of 0.5, i.e., 50% of images used in "Á" (French) article are included in the "A" (English) article. We note that images are only counted once in the rare event that they appear multiple times in an article.

We computed the average metric for the 600 pairs. Table 3 presents the highest and lowest average values of RL_1L_2 (see supplement for the full table).

Overall, the average RL_1L_2 ranges from 22% to 75.6%. On average, *at least* 24.4% of images on an article in one language will be unique relative to an article about the same concept in another language. For some language pairs, this rises to almost 80%.

Hungarian-Romanian had the highest RL_1L_2, providing support for the notion that cultural context may play an important role in image selection. Edition size also appeared to be a factor, with English, French, Polish, and Spanish having relatively high overlaps of other editions. We also observed that Indonesian appears to use substantially different images than other language editions. Because we have already filtered out pages that have no images at all, this indicates that

[2] See our page, `http://whatsincommons.info/icwsm/` for the full table.

Language (Average % of Images Covered by English)		
Indonesian (0.687)	Catalan (0.599)	Finnish (0.551)
Korean (0.67)	Romanian (0.596)	Swedish (0.551)
Spanish (0.634)	Norwegian (0.593)	Russian (0.547)
Polish (0.63)	French (0.59)	Danish (0.546)
Turkish (0.627)	Japanese (0.578)	Czech (0.539)
Hungarian (0.618)	Chinese (0.575)	German (0.536)
Portuguese (0.61)	Dutch (0.565)	Hebrew (0.519)
Italian (0.607)	Ukrainian (0.565)	Slovak (0.436)

Table 4: The RL_1L_2 for each language edition where *Lang2* = English

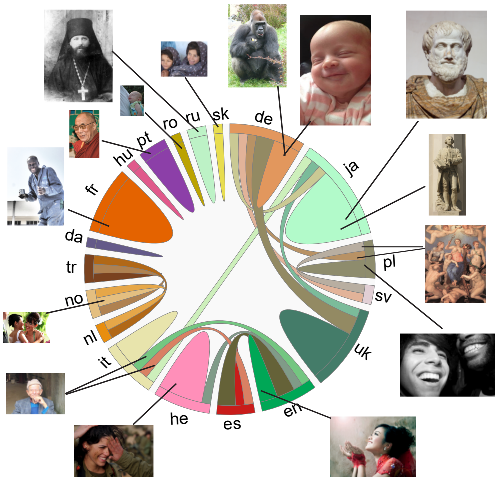


Figure 5: Sample images used in the concept “Happiness”

Indonesian Wikipedia editors may have particularly unique image preferences.

Our results for the language pairs that include English reveal a few additional interesting insights. Because English is the largest language edition (by number of articles), there is a (false) assumption among many Wikipedia readers that English contains most if not all of the content in other language editions (i.e., the “English-as-Superset” assumption (Hecht and Gergle 2010; Bao et al. 2012; Hecht 2013)). Table 4 shows that, as is the case for text, the English-as-Superset assumption is highly problematic for images: people who only look at images on the English Wikipedia about a given concept miss out on a great deal of visual content about that concept in other languages. At best, for a given concept, an English article will on average only contain around 65% of the images in another article about the same concept (e.g. Korean, Spanish). At worst, this average figure will be around 50% (e.g. Hebrew, Slovak).

One potential confound in comparing articles about the same concept across language editions is the issue of “sub-articles” (Lin et al. 2017). Sub-articles occur when a “parent article” has been split into multiple articles due to excessive length and related factors. For example, the English parent article “United States” has been split into multiple sub-articles, e.g. “History of the United States”, “Languages of the United States”. If an image that appears on the parent article in one language edition appears on a sub-article in another language edition, that would add complexity to the interpretation of the above results. The same would be true if an article has one or more sub-articles in one language edition and none in another language edition. While it is reasonable to consider the differential placement of the image across parent and sub-articles to be a form of diversity, it is a different, more nuanced type of diversity than that which we have discussed thus far (i.e. that one language edition chose to visually represent a concept differently than another language edition).

To gain an understanding of the role of sub-articles on diversity measurements, we calculated an upper-bound for their effect. We assumed if any image missing from an article A in *lang2* but that appears in *any* article in *lang2* (call this article B), then B is a sub-article of A (and we treat both A and B as part of the concept A). This is a rather extreme upper-bound for several reasons. First, while many high-interest articles have sub-articles, few lower-interest articles do, and these lower-interest articles make up the bulk of the article distribution (Lin et al. 2017). Second, there are many articles that are relevant to A but are not sub-articles of A and use a given image, especially in large language editions.

Even with this extreme upper-bound for the sub-article effect, substantial within-concept diversity persisted. The average increase in our RL_1L_2 metric (Table 3) when considering this upper-bound was only around 0.08 (both median and mean). Moreover, the maximum RL_1L_2 value increased to only 0.79 and the minimum to only 0.30. As such, we can conclude that the bulk of the within-concept diversity observed cannot be traced to sub-articles (full table online).

To provide a lower-level view of within-concept diversity, we return to a visualization generated by *WikiImgDive*(using our standard, non-sub-article, metrics). Figure 5 displays a *WikiImgDive*’s chord diagram for the concept of “Happiness,” as well sample images as surfaced by *WikiImgDive* (e.g. the Russian orthodox priest is from the Russian language edition, and the gorilla is from German). *WikiImgDive*’s chord diagrams provide an effective way to identify low-diversity concepts. Crossing chords between language pairs indicate image usage overlap. Chords without any connection indiciate images that are uniquely used in the article instance of that language edition. Figure 6 provides an illustration of chord diagrams across the range of low-diversity to high-diversity concepts. The *Wiki* concept (an meta-article about Wiki systems) shows very little diversity as most images are shared. *Paris* and *Car* show moderate diversity, with some shared images, but many unique ones (in particular, within the Czech and French editions). “Science” shows the most diversity (nearly all editions use unique images).

Visual versus Textual Diversity (RQ2)

Having discovered substantial image diversity both across language editions and within concepts, we can focus on answering: *“How does the diversity in visual encyclopedic knowledge compare to the diversity of textual encyclopedia knowledge*?” Here, we compare our findings to the textual diversity patterns identified by Hecht (2013).

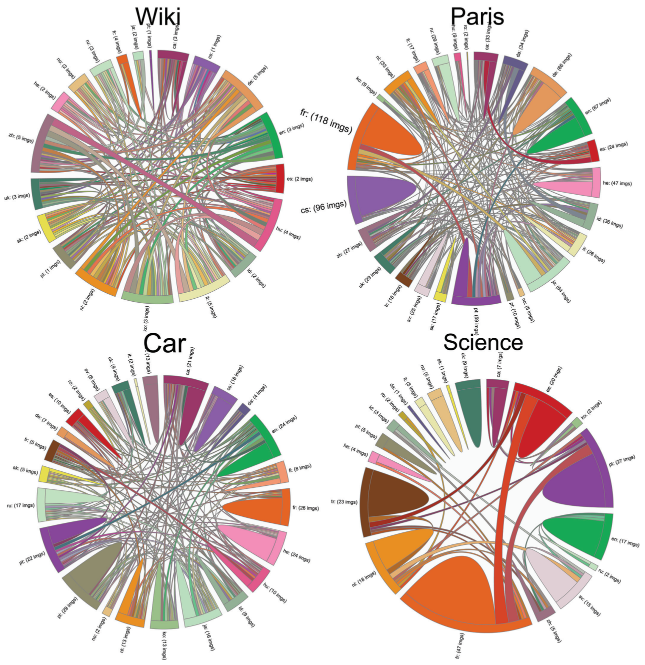


Figure 6: Examples of concepts that have different image diversity patterns. These range from "Wiki" (low-diversity) to "Science" (high-diversity).

However, in order to perform this comparison, we first had to overcome a challenge: Hecht computed his results based on a Wikipedia dump from 2012, whereas our data is from 2017. Although Hecht showed that his results stayed roughly constant between 2009 and 2012–suggesting that text diversity stays roughly fixed over time–we wanted to further ensure that this trend extended into 2017. As such, we used our Wikidata concept mappings to attempt to replicate a key result from (Hecht 2013): the distribution of concepts by the number of language editions in which they have articles. Figure 7 shows this distribution from (Hecht 2013) versus our data. As can be seen, the results in 2012 are nearly identical to those in 2017, suggesting that textual diversity remains fixed over time.

Language Edition-level Diversity vs. Concept Diversity: The distribution visualized in Figure 7–concepts by the number of language editions in which they have articles–is analogous to the language edition-level image diversity distribution in Figure 4. Put simply, the former considers how concepts are *instantiated* across languages editions whereas the latter considers how images are *used* across languages.

The analogous nature of these two distributions presents an opportunity to compare textual and image diversity. Figure 8 makes this comparison and shows a clear similarity for both distributions. This means, at least with respect to this lens, the degree of text and image diversity across language editions is roughly the same.

Figure 8 does show a few smaller-scale noteworthy differences between text and images. Whereas 67.4% of images only appear in a single language edition, 73.5% of concepts only have articles in a single language edition. However, this trend reverses at the other end of the distribution: 0.13% of concepts are "global concepts" while only 0.0014% of images are "global images." In other words, one is more likely to encounter a *concept* that has an article in all 25 language editions than an *image* across all 25.

Image vs. Text Within-Concept Diversity: Because we also used the RL_1L_2 metric in our within-concept diversity analyses, we can compare our within-concept image results to the within-concept text results reported by Hecht (2013). In other words, we can compare the average visual similarity of two articles about the same concept to the textual similarity of those articles.

To calculate textual similarity, Hecht utilized Wikification-enhanced "bags of links." This approach effectively represents each article using the set of concepts that are discussed in the text of the article and compares these representations. Our approach is comparable, except instead of concepts, we use images.

Figure 9 displays the results of a comparison between our within-concept image diversity analysis and the Hecht's analogous analysis for text. The figure reveals that for the vast majority of language pairs, there exists *more image diversity than textual diversity*. Only 22 of 600 language pairs exhibit more text diversity than image diversity. Even considering our extreme sub-article upper-bound, we see that approximately 500 of the 600 pairs still have more diversity in images than text.

Summary of Comparisons: Our results at both the language edition-level and the within-concept-level suggest that image diversity matches or exceeds text diversity. Despite the availability of affordances in the Wikimedia Commons for cross-language image sharing and the many other reasons for believing that textual diversity would be greater, it appears that editors of different editions still represent encyclopedic world knowledge with at least as much visual diversity as they do textual diversity.

Discussion

Drivers of Diversity

While we have identified the extensive image diversity across language editions, a rigorous understanding of what drives this diversity remains an open question (as is the case for text (Hecht 2013)). Using *WikiImgDive*, we identified evidence that one cause is "cultural contextualization." For instance, the article on the concept of Rice in the Indonesian Wikipedia includes a picture of a woman pounding rice near the Indonesian city of Bandung, whereas this image does not appear in other language editions. Similarly, the Catalan article on Chocolate is the only article featuring an image of the Catalan chocolate brand "Xocolata a la pedra." Similarly, in Figure 5, cultural contextualization is likely behind the picture of a Russian orthodox priest appearing exclusively in the Russian language edition. Evidence of cultural contextualization is even apparent in some of our higher-level findings, e.g. that Hungarian-Romanian is the least diverse within-concept language edition pair.

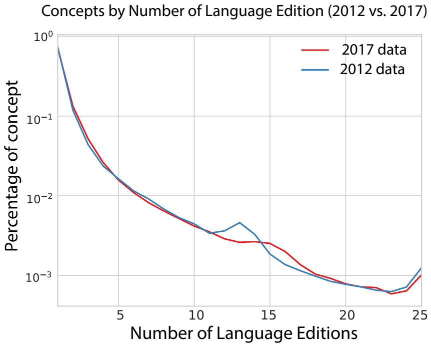


Figure 7: Concept-level diversity changed very little between 2012 and 2017.

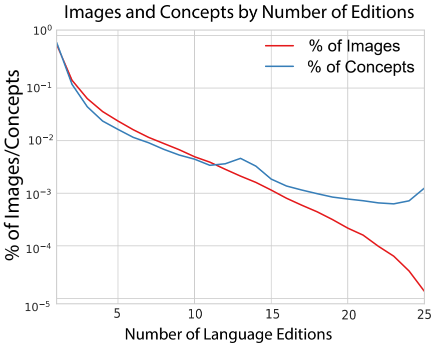


Figure 8: Comparison between Language-Edition Level Diversity (image) and Concept-Level Diversity (text).

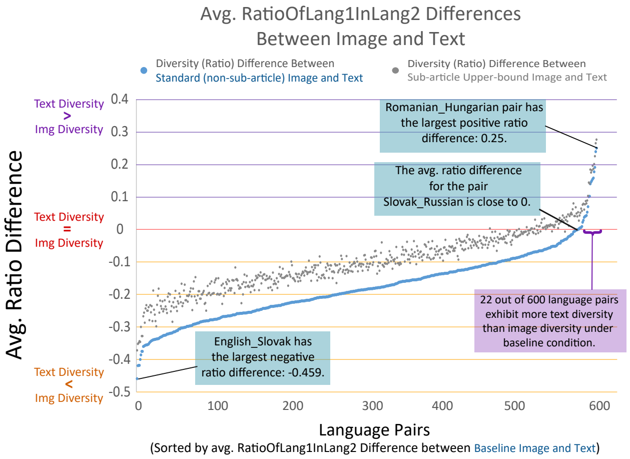


Figure 9: Comparison of RL_1L_2 for images and text.

One additional potential cause for the diversity are simple variations on the same image. These may be generated by cropping, making the background transparent, and so on. For example, the English and German articles on Albert Einstein use different (bit-level) profile pictures for Einstein, but these pictures are derivations of the same image (the German image cropped closer and in grayscale). While some variations may be driven by diversity in cultural aesthetic preferences, others come down to individual-level choices. To determine if simple variations–regardless of their cause–were a major factor behind the diversity we observed, we ran a small coding study. We randomly sampled one article pair from each of 100 language pairings (also chosen at random). Within these 100 articles, we detected the usage of image variants in only five article pairs (5%). As such, image variants are likely only a minor driver for diversity.

The sociotechnical design of the Commons may also play a role in diversity. One design choice in the Commons that may reduce image re-use is that images can be named and indexed in a variety of languages. Examining 5 sample images from each of the 25 languages (125 images total), we saw that 40 (32%) did not have English names. With English likely the most common shared language for editors, non-English file names may make it hard to find, and therefore reuse, images, thereby *boosting* diversity. A system such as PanImages (Colowick 2008) or the use of interlanguage links, may help address this issue while maintaining critical support for multi-lingual names.

Beyond 25 Languages

One area for future exploration is to expand our work beyond the 25 editions. A statistic such as "global images" which appeared in all 25 editions is quite rare (0.0014%). In fact, when considering all 287 languages we found that there are zero "truly global images." That is, no image is used in all language editions of Wikipedia. The most common image is of the bacteria *Gemmatimonas aurantiaca* (appearing in 173 editions). More generally–as is the case with the 25-edition dataset–the 287-edition dataset has a long-tail image usage distribution. Of the 9.8M unique (non-template) hosted images, 5.7M (59%) were single-edition images. Only 15.5% appeared on two editions and 7% on 3 (leaving 18% to appear on four or more editions). Additional analysis may reveal other patterns in the data.

Implications for AI

Given its scale and structure, the dataset that comprises Wikipedia's images is ideal for various AI applications. Indeed, with an eye towards improving support for automated systems, the Wikimedia Foundation recently received a major grant to increase improve the metadata and structure of the Commons (Wikimedia Foundation 2017).

However, our results suggest that Wikipedia's imagery should be used by AI systems with some caution: only considering images from one language edition will likely give these systems a limited, biased view of the world (as has been observed for text (Hecht and Gergle 2010; Hecht 2013)). Fortunately, the situation for visual encyclopedic knowledge is less dire than that for textual knowledge. It is relatively straightforward to gather all images about a given concept rather than those from a single language edition; the main challenge is knowing that this is important to do, and hopefully, this paper can help to address this challenge. Visual AI systems will still be liable to the biases that are manifest in some language editions having more imagery than others, but this is by definition a less serious problem than having to only consider a single language edition due to lexicographic issues.

Conclusion

In this paper, we have extended the literature on the diversity of encyclopedic knowledge across Wikipedia language editions to include images rather than just text. We found that there is a great deal of image diversity across language editions, with this diversity rivaling or even exceeding that found in text. Supplemental data and our live image diversity exploration system (*WikiImgDive*) are available at `http://whatsincommons.info/icwsm/`.

Acknowledgments

This work was funded in part by the U.S. NSF (IIS-1702440, IIS-1707319, CAREER IIS-1707296, and IIS-1421438). We thank Ceren Budak, Clifford Lampe, and Hariharan Subramonyam for their feedback.

References

ACM US Public Policy and Europe Councils. 2017. Principles for algorithmic transparency and accountability. Technical report.

Adafre, S. F., and De Rijke, M. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the Workshop on NEW TEXT*.

Aletras, N., and Stevenson, M. 2013. Representing topics using images. In *HLT-NAACL*, 158–167.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *Pro Publica*.

Bao, P.; Hecht, B.; Carton, S.; Quaderi, M.; Horn, M.; and Gergle, D. 2012. Omnipedia: bridging the wikipedia language gap. In *CHI'12*, 1075–1084. ACM.

Benavent, X.; Garcia-Serrano, A.; Granados, R.; Benavent, J.; and de Ves, E. 2013. Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection. *IEEE Tran. Multimedia* 15(8).

Callahan, E. S., and Herring, S. C. 2011. Cultural bias in wikipedia content on famous persons. *JASIST* 62(10).

Colowick, S. M. 2008. Multilingual search with panimages. *MultiLingual Computing* 19(2).

Cyr, D.; Head, M.; and Larios, H. 2010. Colour appeal in website design within and across cultures: A multi-method evaluation. *Int. J. of Human-Computer Stud.* 68(1):1 – 21.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, 1606–1611.

Hale, S. A. 2012. Net increase? cross-lingual linking in the blogosphere. *JCMC* 17(2):135–151.

Hale, S. A. 2015. Cross-language wikipedia editing of okinawa, japan. In *CHI'15*, 183–192. ACM.

Hecht, B., and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Communities & Technologies '09*.

Hecht, B., and Gergle, D. 2010. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI'10*, 291–300. ACM.

Hecht, B. J. 2013. *The mining and application of diverse cultural perspectives in user-generated content*. Ph.D. Dissertation, Northwestern University.

Kitayama, S., and Cohen, D. 2010. *Handbook of cultural psychology*. Guilford Press.

Knight, W. 2017. Biased algorithms are everywhere, and no one seems to care.

Lau, C.; Tjondronegoro, D.; Zhang, J.; Geva, S.; and Liu, Y. 2006. Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images. In *INEX Workshop*.

Lin, Y.; Yu, B.; Hall, A.; and Hecht, B. 2017. Problematizing and addressing the article-as-concept assumption in wikipedia. In *CSCW'17*, CSCW '17, 2052–2067. ACM.

Milne, D., and Witten, I. H. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *AAAI 2008*.

Miyamoto, Y.; Nisbett, R. E.; and Masuda, T. 2006. Culture and the physical environment. *Psychological Science*.

Pfeil, U.; Zaphiris, P.; and Ang, C. S. 2006. Cultural differences in collaborative authoring of wikipedia. *JCMC* 12(1).

Reinecke, K., and Bernstein, A. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM ToCHI* 18(2):8:1–8:29.

Sen, S.; Li, T.; and Hecht, B. 2014. Wikibrain: democratizing computation on wikipedia. In *WikiSym*, 27. ACM.

Tsikrika, T.; Kludas, J.; and Popescu, A. 2012. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia* 19(3):24–33.

Tsikrika, T.; Popescu, A.; and Kludas, J. 2011. Overview of the wikipedia image retrieval task at imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, volume 4, 5.

Viegas, F. B. 2007. The visual side of wikipedia. In *HICSS'07*, 85–85. IEEE.

Wikimedia Foundation. 2017. Wikimedia foundation receives $3 million grant from alfred p. sloan foundation... wikimedia blog. https://goo.gl/n93eUH.

Wikimedia Foundation. 2018. Commons:project scope/summary. `https://goo.gl/YfqtBY`.

WikiMedia-Meta-wiki. 2017. Data dumps. `https://meta.wikimedia.org/wiki/Data_dumps`.

Wikipedia-English. 2012. Help:template. Page Version ID: 528512056.