This is a repository copy of *What is the Commons Worth? Estimating the Value of Wikimedia Imagery by Observing Downstream Use*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/132898/

Version: Accepted Version

Proceedings Paper:

What is the Commons Worth? Estimating the Value of Wikimedia Imagery by Observing Downstream Use[*]

Kristofer Erickson[†]   Felix Rodriguez Perez[‡]
Jesus Rodriguez Perez [§]

July 1, 2018

Abstract

The Wikimedia Commons (WC) is a peer-produced repository of freely licensed images, videos, sounds and interactive media, containing more than 45 million files. This paper attempts to quantify the societal value of the WC by tracking the downstream use of images found on the platform. We take a random sample of 10,000 images from WC and apply an automated reverse-image search to each, recording when and where they are used 'in the wild'. We detect 54,758 downstream uses of the initial sample and we characterise these at the level of generic and country-code top-level domains (TLDs). We analyse the impact of specific variables on the odds that an image is used. The random sampling technique enables us to estimate overall value of all images contained on the platform. Drawing on the method employed by Heald et al (2015), we find a potential contribution of USD $28.9 billion from downstream use of Wikimedia Commons images over the lifetime of the project.

Keywords: Wikimedia Commons, peer production, economic valuation, Creative Commons, public domain, curation, open licensing

1 Introduction

Established in 2003, the Wikimedia Commons (WC) is a significant volunteer-led repository of free-to-use public domain images. As of March 2018 it contained 45,583,565 files, of which 43,039,140 were images (Wikimedia Commons, 2018). Every illustration or photograph contained in the WC – referred to in copyright law as a 'work' – is available on a free and open basis. This is because either the original term of copyright protection in the work has expired, or the creator of the work has made it available under an open license. As of March 2018 the most commonly used open license on the WC was CC-BY-SA 3.0, which allows use for any purpose, including commercially, as long as the user provides credit to the original author of the work and continues to offer it under the same open license. Other commonly used licenses on the WC allow free use without the viral share-alike clause or the attribution requirement. This feature makes the WC very different from commercial image libraries where copyright law normally forbids unauthorised use and distribution of works.

Given the size and scope of the WC, there has been surprisingly limited empirical investigation of its economic and societal impact. Indeed, much of the cross-disciplinary scholarly work available has tended to use the WC as a valuable site for data-mining and other experimental research, or as a case study in collective governance (Dobusch, 2012; Vaidya et al. 2015). Searching for scholarly articles on the topic of the WC is also hindered by the fact that many scholarly scientific papers contain citations to illustrations and images available on the WC, vastly increasing the amount of false positives in search results.[1] The WC is clearly an important resource for science and

---

[1]This is a problem faced generally by copyright scholars, since most academic papers include the word 'copyright', overwhelming keyword searches.

humanities researchers. But does it have a wider societal impact, and if so, can we attempt to quantify the size of its potential influence?

This paper attempts to characterise the downstream use of image files contained on the WC by performing an automated reverse-image search on a sample of 10,000 randomly-selected image files. We record information about the images prior to the search (image size, quality, license parameters) as well as information about the URLs where images appear (quantity of downstream uses, domain type, language of target page).

We find an overall quantity of 54,758 downstream uses of images from our sample. We estimate a series of logistic regressions to study variables that are significant in the odds of uptake of WC images. Overall, we find that license type is a significant factor in whether or not an image is used outside of the WC. Public domain files and licenses (those without attribution or share-alike clauses) are associated with increased odds of downstream use. This is consistent with other economic studies of the public domain (Buccafusco and Heald, 2013; Erickson, 2018). We also find that for commercial use, prior appearance of the file elsewhere on Wikipedia has a significant positive effect, suggesting that human curation and selection are important in promoting key images to widespread use. We suggest further experimentation using a purposive sample of 'quality' and 'valued' images to test for the impact of human curation on the WC.

The paper proceeds as follows: we first review work on economic value and incentives in peer-produced resources, with a focus on the role of intellectual property licensing on wider commercial usage. We then describe the approach and research methods used in our analysis of WC images and discuss the results. We suggest one

method for calculating social welfare represented by downstream use of WC images and report the result. We close by offering suggestions for further research and policy considerations from the findings generated by this preliminary study.

2   Background and Related Work

Two important economic questions emerge from the study of online peer production. The first question relates to the incentives that animate participation of volunteers in the creation of public goods; the second question relates to the overall societal effects of the availability of peer-produced resources. A significant amount of scholarly research from management and organisational studies has addressed the first question; there has been limited investigation of the second. In this section, we briefly review both literatures with a focus on the role that intellectual property might play, both internally to peer production and externally in wider societal usage.

2.1   Incentives

One enduring question in studies of commons-based peer production has been where volunteer labour comes from. In his seminal legal analysis of the copyright public domain, James Boyle bracketed the question by suggesting it didn't matter, as long as evidence, such as the presence of Wikipedia, showed that it would happen – *'E pur si muove'* (Boyle, 2003). In certain legal scholars' view, maintaining a vibrant public domain in creative works is important for enabling the existence of innovative commons, regardless of how individual communities operate internally (Lessig, 2006). Early observations of open source software communities suggested that communitarian

and altruistic incentives were important for participants, alongside economic incentives (Hars and Ou, 2001). Copyleft, which encourages openness by requiring all modified and extended code to be made freely available, is associated with an ideology of communitarian sharing (Stallman, 2002). In contrast, Von Hippel and Von Krogh (2003) suggested that volunteer participation could still be explained by economic incentives, because contributing to private-collective innovation offers strategic benefits not available to free riders. In a large-scale review of research on open software, Von Krogh et al. (2012) suggested that social norms, self-regulating institutions and communities could be important factors in sustaining open practices. In their study of management concerns for firms that participated in open source communities, Dahlander and Magnusson (2008) found that open licensing could be an impediment to commercialisation if private incentives clashed with open source norms. Similarly, in a case study of engagement with open source communities by mobile phone manufacturer Nokia, Stuermer et al. (2009) found that the requirement to protect certain proprietary corporate information disrupted community development. Overall, the literature on private-collective innovation suggests that while both firms and individuals may derive benefits from participation in collaborative projects, the open licensing environment sometimes presents a challenge to commercialisation.

2.2   Economic impacts

In characterising the public domain, Boyle (2003) identified anecdotal examples of successful commons-based creative production. But what is the overall volume of such activity, and what are its effects on society? Public domain status has been found

to increase the availability of works that would otherwise not circulate due to copyright. For example, Buccafusco and Heald (2013) found that audiobooks made from public domain bestsellers published between the years 1913–22 were significantly more available than those made from copyrighted bestsellers during the ten-year period after 1923. Pollock et al. (2010) analysed the economic contribution of the public domain in a variety of mediums using historical datasets. When calculating the welfare benefit represented by copyright term expiry, the authors counted the marginal increase in sales represented by wider availability of works. They found that public domain status reduced the mean price of printed books by 5-15% at retail, but increased their circulation. By combining price with usage estimates, the authors were able to estimate the net social welfare represented by the expiration of copyrights. Another study by Heald et al. (2015) attempted to measure the social value of public domain imagery using data on page-level Wikipedia visitorship combined with equivalent license pricing from Getty Images. The authors selected a sample of biographical pages across a time period which included in-copyright and public domain photographs. Subject pages accompanied by a freely available public domain image were found to draw an additional 22% traffic usage. Based on industry standard advertising rates for equivalent commercial websites, the authors estimated a consumer surplus for the availability of public domain photographs of between USD \$208M and USD \$232M annually. In this study, we extend the methodology used by Heald et al. to assess the economic value represented by free availability of images from the WC. We do this by first detecting instances of use and then applying the standard Getty editorial license rate as a guide.

5

3   Approach and Methods

3.1   Data collection

We used the MediaWiki API query command to gather a random sample of 10,000 image pages from the WC database in February 2018. In the Wikimedia Commons database, each page is assigned a "random index", which is a random floating point number uniformly distributed between 0 (inclusive) and 1 (exclusive). Because Special:random returns the next article whose random index is greater than the selected random number, the size of 'gaps' between index numbers will bias selection so that certain pages have a higher probability of being selected if different samples are taken repeatedly.[2] This means that the MediaWiki function has limitations if used in repeat studies, but we consider the randomisation sufficient for the purposes of this study, which uses one-shot rather than panel data. For each page returned by our query, we recorded relevant variables (see Table 1). We extracted further information for each file using the API commands *imageInfo*, *globalusage*, *extlinks*, *revisions* and *pageimages*. The main variables of interest were image size, author, source, license type, and linked usage elsewhere on Wikipedia. We also recorded the URL, filename and image description as text strings. Data collection stopped after we reached 10,000 unique images, having first removed duplicate entries.

In the second phase of data collection, we made use of the Selenium open source browser automation framework to repeatedly search for downstream uses of image

---

[2]See /Wikipedia:FAQ/Technical#random. According to the documentation, this command returns results by checking a randomly-generated double-precision floating point number against a randomly-assigned index number for each page contained on the Commons. Some pages will have a larger gap before them in the random index space, and so will be more likely to be selected.

Table 1: Summary statistics for main variables drawn from 10,000 image pages on Wikimedia Commons

Variable	MIN	MAX	MEAN	SD
Any external use	0	1	.348	.476
Any non-commercial	0	1	.304	.46
Any commercial	0	1	.267	.442
Total uses	0	395	5.48	19.78
Total commercial	0	331	2.99	11.722
Total non-commercial	0	129	2.49	8.93
Age of image (years)	0	14	4.4	2.995
Image size (square)	12	8074.5	1324.8	947.3
Format non-jpeg	0	1	.057	.233
Uploader's own work	0	1	.47	.499
Quality image	0	1	0	.062
Originated on Flickr	0	1	.15	.356
Used on Wikipedia	0	1	.171	.376
PD licenses	0	1	.234	.423
Attribution licenses	0	1	.161	.368
Viral licenses	0	1	.582	.493

files.[3] Using this tool, we subjected the URL of each image file to a reverse image search using the public Google web interface. This was accomplished by running a script to complete fields for each query, emulating a human search. The results of the reverse image search were recorded with each case being a URL returned by the search. This process yielded 54,758 URLs. For each returned hit, we recorded the rank in the search results, and extracted the domain information for each URL, recording it in a separate field. We carried out human review to further sort TLDs according to their overall type (country code or generic top-level domain) as well as purpose (commercial TLDs compared to .GOV, .EDU, etc.). Usage results obtained in the second phase of data collection were merged with the first dataset by matching back to unique image IDs. This enabled us to record as a continuous variable the total number of results returned for each original image. We then performed a series of regression analyses, first with any use as a binary outcome variable and then with commercial and non-commercial use as the outcome variable, reported below.

4   Discussion

We found that 34.8% of images in our sample were used externally at least once. This figure does not include previous use on Flickr if an image was obtained from that website. It includes all other detected external uses from URLs besides the page on WC where the original image was hosted. The most frequent user was Wikipedia: some 17.1% of images in our sample were also featured in Wikipedia articles. The

---

[3]Selenium is a browser automation library that may be used for any task that requires automating interaction with the browser. See: https://github.com/SeleniumHQ/selenium
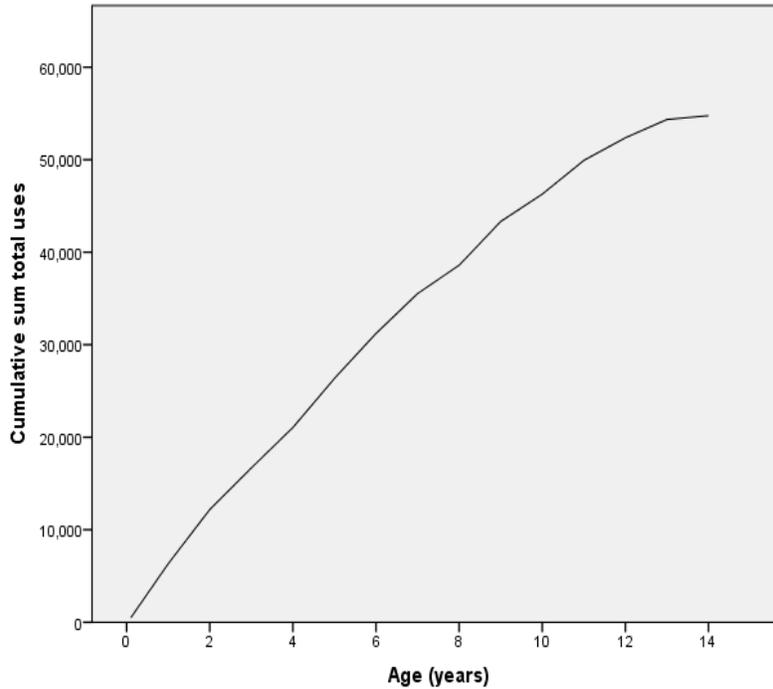
Figure 1: Cumulative sum of 54,758 total uses by years since upload.

mean age of uploaded images was 4.4 years. External use varies with age as shown in Figure 1. Newer images (those uploaded more recently than the mean of 4.4 years ago) accounted for 48.4% of all detected uses. Smaller images account for more of the observed uses, with those below the mean size of 1325 pixels square accounting for 63.9% of all detected uses. Most images hosted on the WC were in JPEG format. In our sample, only 5.7% of images were in a different format, with the most common alternative formats being .PNG and .GIF.

Nearly all of the images in our sample (98.8%) were accompanied by copyright license information. Approximately 47% of images in our sample were the authors' own work, in which case the uploader was prompted to choose the appropriate license.
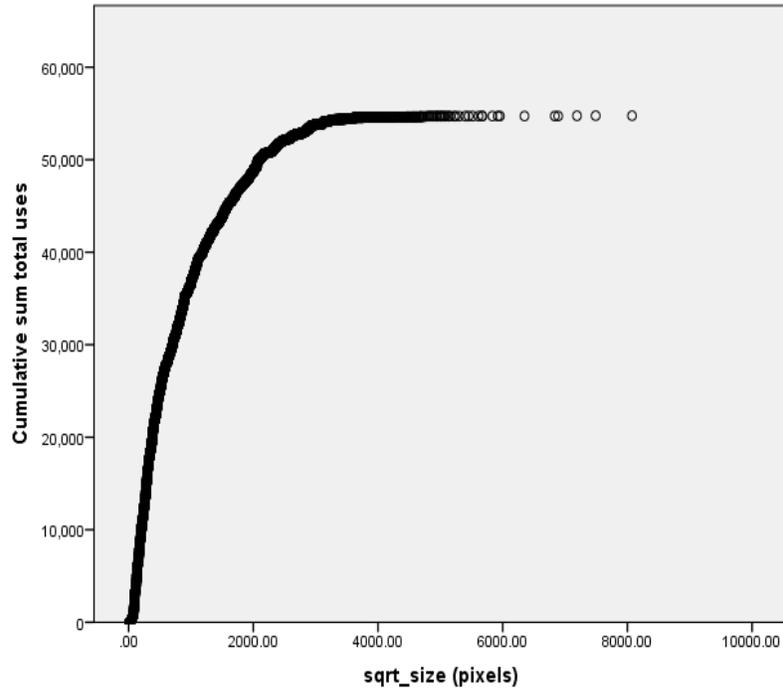
9

Figure 2: Cumulative sum of 54,758 total uses by years since upload.

Some 15% of the sample consisted of freely-licensed images that were pulled from commercial hosting site Flickr (functionality available in the WC Upload Wizard from December 2012), with licensing information automatically accompanying those files across to the WC. In other cases, the uploader specified a license at the point of upload, or reproduced the licensing information in the case of third-party images (such as those marked PD-old for out-of-copyright works).

Figure 3 shows the share of different licenses used in our sample. This information was obtained by extracting the license data from the individual image pages on the WC. The license establishes the uses that can be made of the image without requiring direct permission from the creator of the work. Creative Commons Attribution
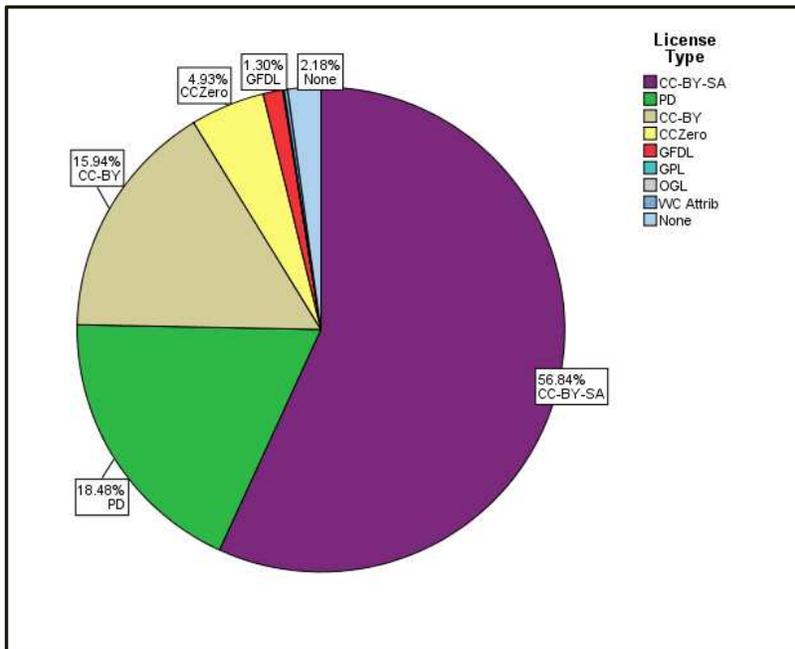
10

Figure 3: License types used

Share-Alike (CC-BY-SA) was the most commonly selected license for images (56.8%), followed by PD marks and other public domain dedications (18.5%) and Creative Commons Attribution licenses (15.9%). Less commonly used licenses included the GNU Free Documentation License (GFDL) and software documentation marked with GPL licenses. Some 2.2% of files did not have a valid license, either because they were not accompanied by adequate information, or because the chosen license conflicted with the information provided (e.g. an attribution license with no known author).

We examined how images were used downstream of the WC by automatically searching for matches and recording information about the domains where images were found. The reverse image-search process detected 54,758 external uses, excluding original WC pages. Individual uses were categorised according to the top-level domain
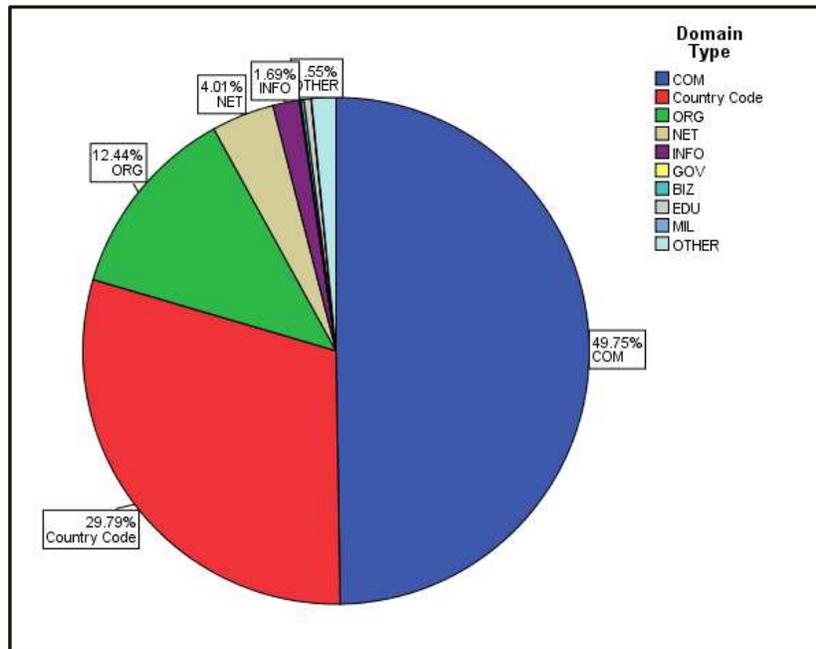
11

Figure 4: Detected uses by TLD

where the use was detected, as summarised in Figure 4. Some 49.7% of the detected uses were found on .COM domains, while .ORG domains (including Wikipedia) made up 12.44% of uses. Some 29.8% of uses were detected on country-level domains (such as .ca or .ru). A human reviewer further categorised uses according to whether they were found on a commercial domain (.COM) or a commercial country code (such as .co.uk). Within the original sample of 10,000 images, some 26.7% were found to have at least one commercial use, while 30.4% were found to have at least one use that we deem non-commercial (any remaining ccTLDs and non-commercial generic TLDs). Further analysis could improve the determination of commerciality, as TLDs can only provide a rough guide. Alternative approaches include crawling the target URLs for the presence of ad code, or analysing text collected from the page headers.

Next, we performed a series of binary logistic regressions taking any detection of use as the dependent variable (1=use detected, 0=otherwise). Table 2 presents the results of 6 regressions using 3 different DVs: Any detected use, non-commercial use and commercial use.

The first model includes only the main control variables (image characteristics and age). Variables of interest are license type (attribution-style, share-alike, or public domain) as well as human curation ('Quality' tag designation). In all models the estimates are shown as odds ratios, with values greater than 1 indicating an increase in the odds of use and values lower than 1 indicating reduced odds.

The age of an image slightly increases the odds of use on commercial and non-commercial domains, suggesting an effect related to discovery time. Larger image size reduces the odds of use. 'Quality' images have singificantly increased odds of use, as do unusual file formats (non-JPEG files). An image's origin on Flickr does not significantly effect its odds of use (Flickr is counted as a commercial external use, so this variable is excluded from other models). By contrast, inclusion on Wikipedia significantly increases the odds that an image is used commercially. We interpret this to indicate the importance of Wikipedia in providing context and exposure to prospective users.

Compared to public domain files, those issued with an attribution or share-alike requirement have significantly reduced odds of being used externally. Attribution and share-alike requirements reduce the odds of commercial use more than for non-commercial use, suggesting that these are important impediments for prospective commercial users.

13

Table 2: Binary logistic regressions with dichotomous measures of external use as the outcome variable

Model	(1)	(2)	(3)	(4)	(5)	(6)
DV	ANY USE	ANY USE	NON-COMMERCIAL	NON-COMMERCIAL	COMMERCIAL	COMMERCIAL
AGE	1.161***	1.159***	1.165***	1.164***	1.151***	1.078***
	(0.008)	(0.008)	(0.008)	(0.015)	(0.008)	(0.010)
IMAGE.SIZE	0.877***	0.881***	0.888***	0.891***	0.877***	0.906***
	(0.014)	(0.014)	(0.014)	(0.002)	(0.015)	(0.018)
NOT.JEPG	1.331**	1.228**	1.316**	1.308**	1.465***	1.163
	(0.097)	(0.098)	(0.099)	(0.098)	(0.099)	(0.121)
OWN.WORK	1.109**	1.268***	1.324***	1.278***	1.109*	0.779***
	(0.046)	(0.052)	(0.053)	(0.057)	(0.055)	(0.068)
QUAL.IMAGE	2.278**	2.311**	2.202**	2.185**	2.168**	1.597
	(0.335)	(0.335)	(0.338)	(0.338)	(0.347)	(0.464)
FLICKR				0.879		
				(0.080)		
WIKIPEDIA						28.022***
						(0.076)
ATTRIBUTION		0.729***	0.727***	0.755***	0.633***	0.578***
		(0.072)	(0.075)	(0.078)	(0.077)	(0.092)
SHARE.ALIKE		0.654***	0.670***	0.677***	0.621***	0.560***
		(0.056)	(0.058)	(0.058)	(0.059)	(0.070)
CONSTANT	1.513**	1.806**	1.228	1.205	1.459*	0.800
		(0.200)	(0.205)	(0.206)	(0.210)	(0.252)
OBSERV	10,000	10,000	10,000	10,000	10,000	10,000
PSEUDO R2	0.097	0.105	0.106	0.104	0.108	0.423

Notes: Odds ratios displayed, SE in parentheses, Pseudo R2 is Nagelkerke's R2, *** p<0.001, ** p<0.05, * p<0.1

4.1 Estimating value

Having established the usage rate of Wikimedia images and identified some of the variables influencing downstream use, it is possible to make estimates of the overall economic value of the project. One method of establishing the value represented by free and open projects is to compare them with equivalent commercial offerings. Consumer willingness to pay (WTP) for commercial services can be used as a benchmark to establish the consumer surplus represented by free and open alternatives such as the WC (Heald et al. 2015).

For commercial comparison, a relevant market is the image licensing industry. Visual content company Getty Images, which holds one of the largest commercial image catalogues in the world, reported 2017 revenue of USD $868 million (Seattle Times, 2016). Image libraries acquire copyright in images from photographers and sell to customers which include advertising agencies, press publishers and corporate communications departments. The image library business model relies on economies of scale to reduce search costs and increase choice for buyers. Significant investment goes into maintaining a user-friendly, searchable platform of images to help prospective customers find exactly what they are looking for. Digitalisation has offered new opportunities for companies like Getty to find and license images to consumers; it has also given rise to alternative ways of curating and distributing photographs and images.

Since the downstream use of images from WC measured in our study is limited to digital uses located on the web, we apply the pricing of web images only. Getty currently offers a 1-year, royalty-free license for commercial editorial use of a digital image from its editorial catalogue for USD $175. We use the editorial rather than

more expensive 'creative' rate because editorial more closely approximates the usage observed for our sample, which includes political figures, landmarks and public events. Based on the mean commercial usage rate of 2.99 for our sample, we can estimate the total commercial use for the entire WC catalogue as [43,039,140 * 2.99 * $175] or approximately USD $22.5 billion over the lifetime of the project. This figure does not include use on non-commercial TLDs such as .ORG or generic country code TLDs, which make up the other 45.4% of the observed uses. Getty and other image libraries are happy to license non-commercial use of their images, typically for a reduced price. The lowest license price we could obtain for non-commercial editorial use of an image from Getty was USD $60. Using the same operation but with different price and mean usage of [43,039,140 * 2.48 * $60] we obtain a figure of USD $6.4 billion for the additional non-commercial uses. Both estimates include total use over the 14-year period since the establishment of the WC.

Our valuation approach is limited by several assumptions. We assume that downstream users would be willing to pay the equivalent of a commercial license to use an image if no free alternative was available from the WC. We also make the assumption that images licensed by Getty are aesthetically equivalent to the free images used from WC. We attempt to address these issues by taking the lowest available license rate from Getty for editorial use, where aesthetic differences are judged to be less significant. Our approach could be improved with greater information about the nature of downstream use (for example if advertising code or e-commerce functionality is present on the page). More information about aesthetic differences between Getty and WC might be obtained by having human reviewers score images.

5 Conclusions

This paper has tracked downstream digital use of images hosted on the WC. We find a mean rate of online use of 5.48 uses per image. Using commercial TLDs as a proxy for commercial use, we estimate a mean commercial usage of 2.99 per image. The odds that a given image from the WC will be used is significantly influenced by the license type issued by its uploader. Images with attribution and share-alike licenses have significantly reduced odds of being used externally compared to images fully in the public domain.

One aim of our paper is to propose a method to assess the economic contribution of volunteer produced, openly licensed content. We offer this approach as an alternative to traditional copyright industry accounts of economic value. Studies of commons-based projects such as this could be helpful to evaluate future policy proposals that may reduce the availability of works in the public domain.

Based on real-world pricing of image licenses from commercial provider Getty images, we estimate a total value of all online uses (commercial and non-commercial) of USD $28.9 billion. The actual societal value of the WC is likely considerably greater, and would include direct personal uses as well as print, educational and embedded software applications not detectable by our reverse image search approach. Getty routinely charges license fees of $650 or more for creative use (such as magazine covers), significantly higher than the rate for editorial use. Our valuation method could be improved with more information about usage rates of commercial stock photography as well as potential qualitative differences between stock and Commons-produced imagery.

The significance of 'quality' tagging for downstream uptake suggests that human

curation plays an important role in the overall value represented by the WC. These are internal mechanisms used by WC contributors to identify images of importance. Users can flag images in various ways, such as 'valued' 'quality' or 'featured' images. Such human flagged images make up a very small proportion of the overall WC, and our random sample did not capture enough of each to carry out a full analysis. In future, we suggest combining a purposive sample of quality and featured images to generate data on the value of human curation to the overall WC. This approach might also be used in combination with aesthetic comparison between WC and Getty images, to establish whether significant qualitative differences exist between professional content and images available in the Commons.

Acknowledgments:

References

James Boyle. 2003. The second enclosure movement and the construction of the public domain. *Law and contemporary problems* 66, 1/2 (2003), 33–74.

Christopher Buccafusco and Paul J Heald. 2013. Do bad things happen when works

enter the public domain?: Empirical tests of copyright term extension. *Berkeley Technology Law Journal* (2013), 1–43.

Wikimedia Commons. 2018. Media Statistics. `https://commons.wikimedia.org/wiki/Special:MediaStatistics`

Linus Dahlander and Mats Magnusson. 2008. How do firms make use of open source communities? *Long range planning* 41, 6 (2008), 629–649.

Leonhard Dobusch. 2012. Strategy as a Practice of Thousands: The case of Wikimedia. In *Academy of Management Proceedings*, Vol. 2012. Academy of Management Briarcliff Manor, NY 10510, 1–1.

Kristofer Erickson. 2018. Can Creative Firms Thrive Without Copyright? Value Generation and Capture from Private-Collective Innovation. *Business Horizons* (2018), 1–19.

A Hars and S OU. 2001. Working for Free?–Motivations of Participating in Open Source Projects, 00 (c), 1–9. In *34th Annual Hawaii International Conference on System Sciences (HICSS-34), Havaí.* 25–39.

Paul Heald, Kristofer Erickson, and Martin Kretschmer. 2015. The Valuation of Unprotected Works: A Case Study of Public Domain Images on Wikipedia. *Harv. JL & Tech.* 29 (2015), 1.

Eric von Hippel and Georg von Krogh. 2003. Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science* 14, 2 (2003), 209–223.

Lawrence Lessig. 2006. Re-crafting a public domain. *Yale JL & Human.* 18 (2006), 56.

Rufus Pollock, Paul Stepan, and Mikko Välimäki. 2010. *The Value of the EU Public Domain.* Technical Report. Faculty of Economics, University of Cambridge.

Richard Stallman. 2002. *Free software, free society: Selected essays of Richard M. Stallman.* Free Software Foundation.

Matthias Stuermer, Sebastian Spaeth, and Georg Von Krogh. 2009. Extending private-collective innovation: a case study. *R&d Management* 39, 2 (2009), 170–191.

Seattle Times. 2016. Corbis sale to Chinese company may be boon to Getty. `https://www.seattletimes.com/business/technology/corbis-sale-to-chinese-company-may-be-boon-to-getty/`

Gaurav Vaidya, Dimitris Kontokostas, Magnus Knuth, Jens Lehmann, and Sebastian Hellmann. 2015. DBpedia commons: structured multimedia metadata from the wikimedia commons. In *International Semantic Web Conference.* Springer, 281–289.

Georg Von Krogh, Stefan Haefliger, Sebastian Spaeth, and Martin W Wallin. 2012. Carrots and rainbows: Motivation and social practice in open source software development. *MIS quarterly* (2012), 649–676.