An Exploration of Wikipedia Data as a Measure of Regional Knowledge Distribution

Fabian Stephany ✉[a,*], Fabian Braesemann[b,*]

[a] *Vienna University of Economics and Business, fabian.stephany@wu.ac.at, ORCID ID 0000-0002-0713-6010*
[b] *Oxford Internet Institute - University of Oxford, fabian.braesemann@oii.ox.ac.uk, ORCID ID 0000-0002-7671-1920*

This is an Accepted Manuscript reprint of a proceedings article published by Springer Lecture Notes in Computer Science. The original source of publication is:

Stephany F., Braesemann F. (2017) An Exploration of Wikipedia Data as a Measure of Regional Knowledge Distribution. In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol 10540. Springer, Cham.
DOI: 10.1007/978-3-319-67256-4$_4$

Available online: `https://link.springer.com/chapter/10.1007/978-3-319-67256-4_4`

Abstract

In today's economies, knowledge is the key ingredient for prosperity. However, it is hard to measure this intangible asset appropriately. Standard economic models mostly rely on common measures such as enrollment rates and international test scores. However, these proxies focus rather on the quality of education of pupils than on the distribution of knowledge among the whole population, which is increasingly defined by alternative sources of education such as online learning platforms. As a consequence, the economically relevant stock of knowledge in a region is only roughly approximated. Furthermore, they are abstract in content, and both capital-, and time-consuming in census. This paper proposes to explore Wikipedia data as an alternative source of capturing the knowledge distribution on a narrow geographical scale. Wikipedia is by far the largest digital encyclopedia worldwide and provides data on usage and editing publicly. We compare Wikipedia usage worldwide and edits in the U. S. to existing measures of the acquisition and stock of knowledge. The results indicate that there is a significant correlation between Wikipedia interactions and knowledge approximations on different geographical scales. Considering these results, it seems promising to further explore Wikipedia data to develop a reliable, inexpensive, and real-time proxy of knowledge distribution around the world.

JEL classification: C 55, C 82, I 21
Keywords: Mining of Big Social Data, Wikipedia, Knowledge Geographies

---

1. Introduction

The stock of knowledge is a strong determinant for economic growth and the vast majority of economic forecasting models includes some kind of approximation for it. Most commonly, enrollment rates in tertiary education and standardized international education scores, such as the Programme for International Student Assessment (PISA) are used to capture the distribution of

*Both authors contributed equally to this work.

knowledge in economic modeling. While the availability of these standard measures has increased over the last decades, these indicators are usually not available below the country level. Moreover, due to the complex process of data collection, they often lag significantly behind in time.

When it comes to relating knowledge or know-how to economic growth, particular knowledge in specific domains is most relevant. Scholars have argued that, from an economic point of view, increased enrollment in some subjects show higher return on investment than others. Most prominently, the acquisition of knowledge in natural and computer science, mathematics, as well as in engineering, seems to boost economic growth to a sizable extent [1]. The increasing digitalisation in many economic sectors suggests that this trend will continue in the future. At the same time, it remains questionable if assessments like PISA can reflect the complex distribution of knowledge across different disciplines. The capability of performing well in reading and calculus is certainly a basic requirement for the acquisition of further skills, but it does not represent the much higher stock of knowledge that is relevant for innovation and competitiveness in todays' high-tech economies. Additionally, international test scores only focus on a small reference group, in most cases students, while neglecting the knowledge distribution among the rest of the population.

At the same time, the education sector gets more and more digitized [2]. Massive open online course (MOOC) providers like *Coursera* and *Khan Academy* bring high-level education to people all over the world. While these technologies indisputable help individuals to acquire important knowledge, classical measurements like enrollment rates do not capture their effect on the stock of knowledge in a region.

As an alternative to other knowledge measurements, this work therefore proposes the exploration the world's largest on-line encyclopedia, Wikipedia, as a source for retrieving data about knowledge distribution on a global and narrow geographical scale.

Wikipedia data has the following properties that are appealing to use it as a source to measure knowledge: first, its predominance as general reference work makes it a first stop for many who seek information online; it is thus an important provider of knowledge.[1] Secondly, the content production is a collaborative project. Everyone is invited to contribute to the articles. Indeed, millions of users worldwide provide content to the different language versions of Wikipedia.[2] Thirdly, while consuming Wikipedia articles is a process that increases knowledge, editing articles is even more associated with personal learning [4, 5, 6]. Similar to the online learning activities described above, the contribution to Wikipedia content can thus be considered as a learning process related to domain-specific knowledge. Finally, Wikipedia data are publicly available. Usage data are published every month on country level.[3] The edit history of all articles can be collected via an SQL interface.[4] Since the articles are semi-hierarchically organised into categories, it is possible to assign edits into different domains of knowledge. Taking all this together, Wikipedia appears to be a valuable data source that allows to map the global distribution of knowledge. While it takes the macro-perspective of standardized knowledge assessments, it features the timeliness and

[1] Wikipedia provides over 40 million articles in 250 languages worldwide and is ranked among the top-ten most popular websites, see [3].

[2] According the [3] the English language version alone has more than 30 million registered editors and additionally a large number of not-registered editors.

[3] `http://bit.ly/PageViewsPerCountry`

[4] `http://bit.ly/WikiSQL`

availability of online data sources.

This article explores the feasibility of Wikipedia data to measure the stock of knowledge on different geographic scales by relating usage and edit data to common measures globally and in the United States.

2. Literature Review

As outlined in the introduction, the primary underlying assumption of this work is that engagement on Wikipedia, via usage and editing, reflects learning and the knowledge base of individuals and, as a consequence, of societies as a whole. Previous studies about topical editing on Wikipedia have found strong evidence for the assumption ([4], [5], [6]) that editing Wikipedia articles deepens the contributors' understanding of the articles' topic. Other scholars confirm that editors feel confident with the subject they are contributing to, as reported by Collier and Bear [7]. The authors also report some of the results of the Wikipedia user survey that has taken place in 2008. From 22,000 readers of the English Wikipedia that took part in the survey, more than 50 % are students. The participating contributors report on average 14.7 years of education, which is slightly higher than the overall U. S. or U. K. average. Thus, the relation between domain knowledge and editing behaviour has been proven relevant on an individual level, but no large scale investigation on the relation between Wikipedia and knowledge measures has been undertaken so far.

Most studies that employ Wikipedia data, utilize the network of different language versions for each article. Examples are studies that focus on similarities between cultures [8], [9], [10] or the global influence of languages [11].

Additionally, the global scope of the online encyclopedia has encouraged the investigation on research questions with a geographical background. Yasseri et al. [12], for example, examine the geographical and linguistic similarities and differences between controversial topics on Wikipedia in so-called "edit wars". Their findings show that Wikipedia should be considered as more than "just" an encyclopedia, for it contains also information about socio-spatial patterns of interest groups, which can be observed as they develop over time. A similar study has been conducted by Borra et al. [13] who provide a tool that gives insights into the development of controversial topics on Wikipedia.

Similarly, Wikipedia activities can be used to mirror geographical circumstances. For instance, Graham et al. [14] show, in analysing all geocoded articles of the English Wikipedia, that regional imbalances in broadband connections match the activity rates on Wikipedia. Analog to the physical world, the researchers map the participation patterns on Wikipedia. They underline the importance of access to knowledge, via Wikipedia, and advocate the democratization of the digital space in order to overcome existing "analog" inequalities. Other studies on participation and the geographical distribution of Wikipedia articles find spatial clusters in the knowledge production which lead to a digital underrepresentation of certain parts of the World ([15], [16], and [17]).[5]

In summary, past contributions on Wikipedia have proven that a second use of its' data can help to answer specific research questions from the social sciences. However, to our best knowledge,

[5]Visualiszations of these information geographies can be found here: [18], [19].

there are no large-scale investigations on the applicability of Wikipedia data as a measure of knowledge distribution available so far.

3. Scope and Methodology

3.1. Access to Wikipedia and Knowledge Measures

Wikipedia is by far the largest online encyclopedia. The English Wikipedia alone counts roughly 3,000 views per second. Surely, the share of people that have regular access to the internet and as a consequence may visit Wikipedia varies significantly around the world. While large parts of the population in high-income countries use the internet on a daily basis, this is not the case for many developing countries.

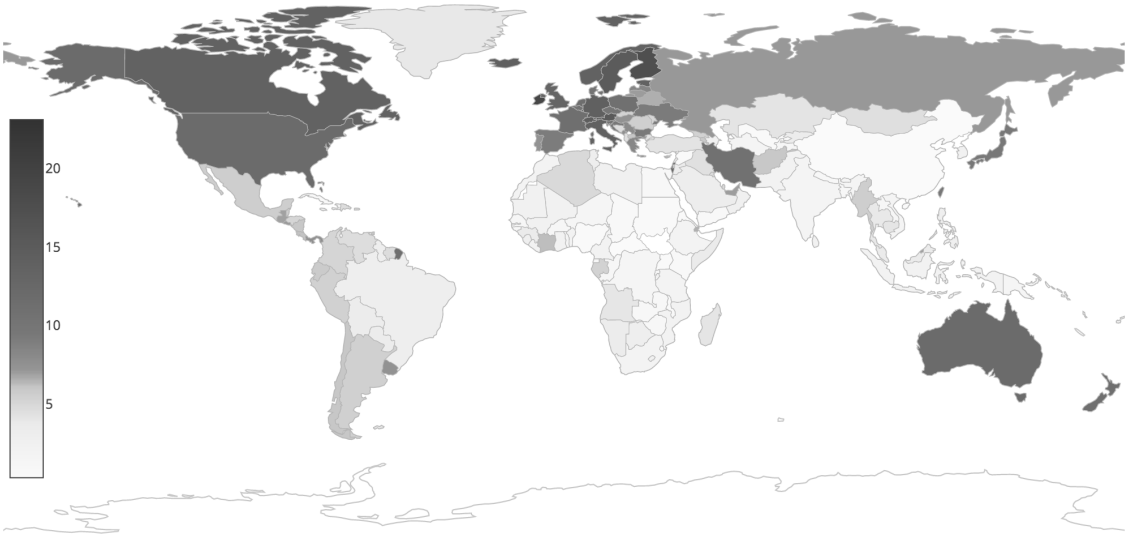

Figure 1: *Monthly Wikipedia clicks per internet user (averaged on the period 2011 to 2017) - Wikipedia is substantially more often visited in high income countries than in developing countries. China is an outlier with very few clicks per user.*

As shown in figure 1, the clicks to Wikipedia per internet user are highest for high-income countries, lower for emerging economies and lowest for poor countries, particularly in Africa. The map moreover shows that this pattern is largely consistent with only two major outliers. China, with more than 650 million internet users counts on average only 0.2 monthly Wikipedia clicks per user. This is a clear indication of the distorted internet access in China. The other outlier is Iran with more than 10 monthly Wikipedia clicks per user on average. These examples show that one needs to be cautious in making far-fetched claims on the general inference of results gained by the big data analysis of Wikipedia. However, the rates of usage across high-income countries indicate the prevalence of Wikipedia across the populations of these countries.

It is one of the main assertions in this article that the distribution of Wikipedia usage does not only reflect economic circumstances, but that it is rather related to differences in the stock of knowledge in the respective economies. If one consults Wikipedia especially in the process of

learning or acquiring knowledge about a specific topic, then it should be expected that classical measures like student enrollment rates are related to Wikipedia clicks. In order to investigate this hypothesis, we collect tertiary enrollment rates as well as economic control variables on a country level from the World Bank.[6]. The Wikipedia usage level is collected from the Wikimedia Foundation.[7]

3.2. Wikipedia Edit Data

While Wikipedia usage data is only available on a very aggregated level, information about the edits to Wikipedia can be collected in much more detail. This is true for the geographical dimension, as many editors are identified by their IP-address which can be geolocated on a narrow geographical scale (at least for many developed countries), as well as for the topic of contribution. The edits are displayed for each single article. The articles themselves are summarised into categories and sub-categories. Following the tree-like structure of the categories starting by a relatively broad one, it is thus possible to collect a lot of data on the geographical origin of edits in a specific domain.

To test the feasibility of the proposed approach, we concentrate in this article on one specific category of Wikipedia articles: all pages that are linked to the category "Computer Science" and its subcategories to a depth of two subcategory-layers.[8] Moreover, the analysis is limited to the English version of Wikipedia. In total, the edits of 12,669 Wikipedia pages that are linked to 203 subcategories of the category "Computer Science" have been collected via the SQL-interface provided by the Wikimedia Foundation.[9] This leads to 467,896 edits (date of data collection: 26th June 2017) with an IP-address. After geocoding the IP-addresses using the *freegeoip.net*-API, we end up with 60,326 individual IP-addresses in the United States that are considered as individual editors contributing to the domain "Computer Science".

4. Findings

The results on the relation between Wikipedia usage and tertiary enrollment rates on a country level are summarised in model (1) in table 1. Wikipedia click rates per internet user are available for the years 2011 to 2016. The World Bank provides information on tertiary enrollment for 127 countries. In total, 492 country-year observations are included in a simple regression model that relates the Wikipedia clicks per users to the share of urban population, GDP per capita, and the tertiary gross enrollment rate. Even on this very aggregate level, the regression shows a highly significant relation between tertiary enrollment, a classical knowledge measure commonly applied in economic models, and Wikipedia usage. Moreover, and as expected from the illustration in figure 1, Wikipedia clicks are also highly correlated with GDP per capita. This result can be considered as a first indication in favour of the hypothesis that, on a macro level, Wikipedia

[6] `http://data.worldbank.org/`

[7] `http://bit.ly/PageViewsPerCountry`

[8] Thus we include all articles that link to the category "Computer Science" itself (level 1), to the subcategories that link to the "Computer Science" (level 2) and to their subcategories (level 3). This number of layers has been chosen to avoid the collection of edit data that are only very weakly related to computer science.

[9] `http://bit.ly/WikiSQL`

interactions are related to the acquisition of knowledge. However, since usage data are not available on a more narrow level, the potential of this type of analysis is limited.

Table 1: Regression results on three different models on the relation between knowledge indicators and Wikipedia usage / editing - In all three models the relative number of Wikipedia usage in general/ Editors in computer science is related to control variables and indicators of knowledge.

	(1)	(2)	(3)
Dimension	Country-Year	U. S. states	U. S. cities
Dep. Variable (log. values)	Clicks per User[a]	Editors in CS. per 100,000 Population	
Observations	492	51	160
Intercept	-11.72 (1.83)***	0.36 (0.25)	2.01 (0.21)***
% Urban Pop.	0.00 (0.01)	0.22 (0.13)	
GDP per capita[b]	1.56 (0.25)***	-0.00 (0.01)	
Tertiary gross enrollment	0.07 (0.01)***		
% Internet users		1.85 (0.35)***	
% Advanced degree[c]		2.85 (0.79)***	
Pop. Density			0.001 (0.00)**
Med. Age			0.009 (0.01)
Med. Income			0.000 (0.00)
Share of Students			2.952 (0.47)***
CS Department			0.589 (0.11)***
Adj. R^2	0.49	0.30	0.29

a) 127 countries from 2011-2016
b) For model (1) as log of PPP current international USD

More revealing results with respect to the relation between the stock of specific knowledge and Wikipedia interactions can be gained by considering editing data. Figure 2 shows the number of "editors" (individual geolocated IP-addresses) in computer science per state and for the 307 largest cities with more than 100,000 population in the United States. We focus on the U. S. in order to avoid effects of the different language versions of Wikipedia. Moreover, the U. S. are culturally largely homogeneous and many data about the economy and education are available, even on a city level. The darker the contour of the state, the more editors per 100,000 population. The share of editors in computer science on the city level is captured by the size of the circle at the location of the city.[10] The colour represents the presence of an academic computer science department in the city.

Interestingly, the more urban states at the coasts show a higher level of editors than the more rural states in the center of the United States. Particularly revealing is however the relation between cities with and without an academic computer science department. Obviously, there are more large dark circles than large bright circles, indicating a difference in the number of editors between both types of cities. While the cities with computer science department count on average 51.1 editors, the cities without computer science department have on average 18.2 computer science editors. Prominent examples in the map are Cambridge in Massachusetts, as well as Berkley, Pasadena and Seattle.

[10] An interactive version of the map can be accessed via the online dashboard that provides supplementary information to this article: `http://bit.ly/Wiki_Dashboard`

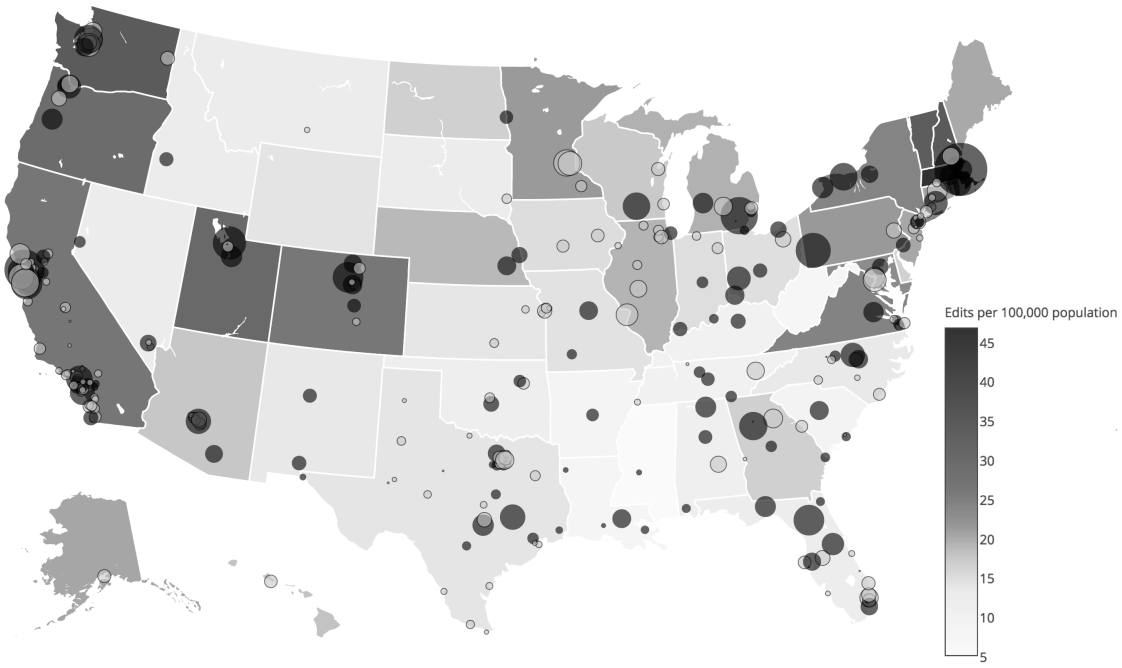

Figure 2: *Computer science edits per 100,000 population on U. S. state and city level - On a regional level, edits in computer science differ substantially across the United States, particularly between the more urban states at east and west coast and the more rural states in the center. Moreover, even larger differences are evident on a city level, as displayed by the circles for the 307 U. S. cities with more than 100,000 population. The color represents the presence of an academic computer science department in the city (dark gray: Department = 1, light gray: No Department = 0). Those cities with an computer science department have on average significantly more computer science edits per 100,000 population than those cities without such an institute:* $\bar{x}_1 = 51.1, \bar{x}_0 = 18.2$.

These descriptive results are confirmed by the inferential models presented in table 1. Model (2) relates the share of editors per 100,000 population on a state level (50 states and the District of Columbia) to the share of urban population, GDP per capita, the share of internet users and the share of individuals with an advanced degree (Master or PhD).[11] Both the portion of internet users and the share of advanced degree holders, indicating the stock of knowledge, are significantly related to the number of editors in computer science.

On an even more granular level, model (3) shows that the number of editors is significantly correlated to the number of students and the presence of a computer science department in the city.[12]

These results confirm the hypothesis, at least for the example of computer science, that Wikipedia data is suitable to capture the stock of knowledge on a very granular geographical level.

[11]The data stems from the U. S. census: `https://www.census.gov/`

[12]Data on the number of students stems from `http://www.stateuniversity.com/`. City-level covariates are collected from `http://www.city-data.com//` and a list of academic computer science departments is available on Wikipedia: `http://bit.ly/CS_Departments`

5. Conclusion and Future Work

The goal of this exploratory analysis is to test the feasibility of Wikipedia as a data source to measure the stock of knowledge. The initial findings confirm, first, that it is feasible to establish a robust relationship between Wikipedia data and traditional metrics of knowledge. Page views correlate to a sizable extent with tertiary education rates on a country level. For specific fields of knowledge this association is shown to be even more pronounced when compared to domain-specific Wikipedia engagement. For the case of Wikipedia editors in computer science, it can be shown on a fine-grained geographical level that they are significantly correlated with real-world entities of knowledge production, namely the presence of an academic computer science department on a city level. This relation is robust on top of the share of student population per city, which is incorporated into the inferential model as control variable. The most prominent cities with respect to computer science edits per capita are the places of top universities in the United States. It is an astonishing that the approach of simply counting the number of IP-addresses in the editing history of specific Wikipedia articles reveals exactly this result.

This study sheds some light on the enormous potential of Wikipedia as a data source to proxy the stock of knowledge. In order to investigate how robust and generalisable the results are, more Wikipedia data from more domains should be analysed. Moreover, it would be interesting to investigate whether Wikipedia edit data can actually be employed to accurately predict the presence of a specific academic department in a city. For this purpose, more sophisticated data science algorithms should be used.

Nonetheless, it can be summarised that the access to data with both a precise geographical and contextual focus opens a vast horizon for exploring domain-specific regional knowledge distribution via Wikipedia.

Acknowledgements. The authors are thankful for the feedback this work received on the brown-bag seminar at the Oxford Internet Institute and the SIS Statistics and Data Science conference in Florence, both taken place in June 2017. Particularly helpful comments have been made by Scott Hale, Otto Kässi, and Taha Yasseri.

References

[1] N. Benos, S. Zotou, Education and Economic Growth: A Meta-Regression Analysis, World Development 64 (2014) 669–689. doi:10.1016/j.worlddev.2014.06.034.

[2] V. Mayer-Schönberger, K. Cukier, Learning with Big data: The future of education, Houghton Mifflin Harcourt, 2014.

[3] Wikimedia, Wikipedia (2017).
URL https://en.wikipedia.org/wiki/Wikipedia

[4] C. L. Moy, J. R. Locke, B. P. Coppola, A. J. McNeil, Improving Science Education and Understanding through Editing Wikipedia, Journal of Chemical Education 87 (11) (2010) 1159–1162. doi:10.1021/ed100367v.

[5] M. Ebner, M. Kickmeier-Rust, A. Holzinger, Utilizing Wiki-Systems in higher education classes: a chance for universal access?, Universal Access in the Information Society 7 (4) (2008) 199. doi:10.1007/s10209-008-0115-2.

[6] J. Cain, B. I. Fox, Web 2.0 and Pharmacy Education, American Journal of Pharmaceutical Education 73 (7) (2009) 120. doi:10.5688/aj7307120.

[7] B. Collier, J. Bear, Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions, in: Proceedings of the ACM 2012 conference on computer supported cooperative work, ACM, 2012, pp. 383–392.

[8] P. Gloor, P. De Boer, W. Lo, S. Wagner, K. Nemoto, H. Fuehres, Cultural Anthropology Through the Lens of Wikipedia-A Comparison of Historical Leadership Networks in the English, Chinese, Japanese and German Wikipedia, arXiv preprint arXiv:1502.05256.

[9] Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, D. L. Shepelyansky, Interactions of cultures and top people of Wikipedia from ranking of 24 language editions, PloS one 10 (3) (2015) e0114825.

[10] P. Laufer, C. Wagner, F. Flöck, M. Strohmaier, Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures, in: Proceedings of the ACM Web Science Conference, ACM, 2015, p. 3.

[11] S. Ronen, B. Gonçalves, K. Z. Hu, A. Vespignani, S. Pinker, C. A. Hidalgo, Links that speak: The global language network and its association with global fame, Proceedings of the National Academy of Sciences 111 (52) (2014) E5616–E5622.

[12] T. Yasseri, A. Spoerri, M. Graham, J. Kertész, The most controversial topics in Wikipedia: A multilingual and geographical analysis, arXiv:1305.5566 [physics].

[13] E. Borra, E. Weltevrede, P. Ciuccarelli, A. Kaltenbrunner, D. Laniado, G. Magni, M. Mauri, R. Rogers, T. Venturini, Societal controversies in Wikipedia articles, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM, 2015, pp. 193–196.

[14] M. Graham, R. K. Straumann, B. Hogan, Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia, Annals of the Association of American Geographers 105 (6) (2015) 1158–1178. doi:10.1080/00045608.2015.1072791.

[15] M. Graham, B. Hogan, R. K. Straumann, A. Medhat, Uneven geographies of user-generated information: patterns of increasing informational poverty, Annals of the Association of American Geographers 104 (4) (2014) 746–764.

[16] M. Graham, S. De Sabbata, M. A. Zook, Towards a study of information geographies:(im) mutable augmentations and a mapping of the geographies of information, Geo: Geography and environment 2 (1) (2015) 88–105.

[17] D. Hardy, J. Frew, M. F. Goodchild, Volunteered geographic information production as a spatial process, International Journal of Geographical Information Science 26 (7) (2012) 1191–1212.

[18] M. Graham, S. De Sabbata, Information Geographies at the Oxford Internet Institute (2014).
URL http://geography.oii.ox.ac.uk/

[19] H.-T. Liao, B. Hogan, M. Graham, S. A. Hale, H. Ford, Wikipedia's Networks and Geographies: Representation and Power in Peer-Produced Content (2010).
URL https://www.oii.ox.ac.uk/research/projects/wikipedias-networks-and-geographies/