Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty

Mark Graham
Oxford Internet Institute, University of Oxford, OX13JS, United Kingdom, (44) 1865 287210
(Mark.Graham@oii.ox.ac.uk)

Bernie Hogan
Oxford Internet Institute, University of Oxford, OX13JS, United Kingdom, (44) 1865 287210
(Bernie.Hogan@oii.ox.ac.uk)

Ralph K. Straumann
Oxford Internet Institute, University of Oxford, OX13JS, United Kingdom, (44) 1865 287210
(Ralph.Straumann@oii.ox.ac.uk)

Ahmed Medhat
Oxford Internet Institute, University of Oxford, OX13JS, United Kingdom, (44) 1865 287210
(amedhatm@gmail.com)

Pre-publication version: Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. 2014 Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty *Annals of the Association of American Geographers* (forthcoming).

Abstract

Geographies of codified knowledge have always been characterized by stark core-periphery patterns: with some parts of the world at the center of global voice and representation, and many others invisible or unheard. However, many have pointed to the potential for radical change as digital divides are bridged and 2.5 billion people are now online.

With a focus on Wikipedia, which is one of the world's most visible, most used, and most powerful repositories of user-generated content, we investigate whether we are now seeing fundamentally different patterns of knowledge production. Even though Wikipedia consists of a massive cloud of geographic information about millions of events and places around the globe put together by millions of hours of human labor, it remains that the encyclopedia is characterized by uneven and clustered geographies: there is simply not a lot of content about much of the world.

The paper then moves to describe the factors that explain these patterns, showing that while just a few conditions can explain much of the variance in geographies of information some parts of the world remain well below their expected values. These findings indicate that better connectivity is only a necessary, but not a sufficient condition for the presence of volunteered geographic information about a place. We conclude by discussing the remaining social, economic, political, regulatory, and infrastructural barriers that continue to disadvantage many of the world's informational peripheries. The paper ultimately shows that, despite many hopes that a democratization of connectivity will spur a concomitant democratization of information production, internet connectivity is not a panacea, and can only ever be one part of a broader strategy to deepen the informational layers of places.

Geographies of Knowledge, Geoweb, Internet Geography, Representation, Wikipedia

Uneven Geographies of Knowledge

Traditional mass media tend to reinforce the already visible and powerful at the expense of minority or oppositional perspectives. Through a concentration of wealth, access to the means of information distribution and aversion to risky counter-narratives, the media tend to produce hegemonic representations of the world that reinforce and legitimate the powerful and dominant (c.f. Gramsci 1971). We have long known that power can be exerted through systems of cultural production and reproduction that can exert cultural hegemony (Laclau and Mouffe 1985). Said differently, culture is a site of conflict and contestation in which struggles for power play out (c.f. Hall 1997), and knowledge and codified information about the social, economic, and political contexts to our lives are always produced under conditions of power (Crampton 2008). Therefore by examining the ways in which the world is represented, we can learn much about global disparities of power.

We have traditionally been able to see some of these conditions of power in voice and representation reflected in stark core-periphery patterns within geographies of knowledge. Almost all mediums of information (e.g. book publishing, newspaper publications and patents) in the early twenty-first century are still characterized by huge geographic inequalities: with the Global North producing, consuming and controlling much of the world's codified knowledge, and the Global South largely left out of these processes (Zhang et al. 2013; Graham, Hale, and Stephens 2011; Thompson and Fox-Kean 2005). Most codified information has been created by, about, and for a small group of people and places in the Global North (Graham 2013). More broadly, these uneven geographies of knowledge have been dubbed a 'New World Information Order' by UNESCO (Mowlana 1997) and have even been described as part of 'a new phase in a long history of the West's attempt to colonize not only the territory and the body but also the mind of the Third World "other"' (Sardar 1996).

These findings parallel work conducted under the banners of 'participatory GIS' and 'critical GIS.' Scholars and practitioners of GIS have long been asking critical questions about the representation of places in digital databases. Pickles (1995), for example, has pointed to the fact that those representations are necessarily implicated in broader networks of power.

Elwood (2006) has similarly pointed to barriers to participation by arguing that "while the financial costs of hardware, software, and data have dropped and the options for acquiring and representing spatial information are greatly expanded for the most advantaged users, at the bottom of the digital divide relatively little has changed." Others have demonstrated that financial and skill barriers can also act as a significant obstacle to the use of digital data and GIS (Craig and Elwood 1998, Sawicki and Craig 1996, Weiner et. al. 1995). Scholars have also pointed to the challenges of including non-Cartesian or unstable forms of knowledge (and by extension the people, processes, and places that they represent) in codified spatial databases (e.g. Rundstrom 1995). Some work has been carried out to indicate that such divides have likely only served to reinforce digital inequalities. Crutcher and Zook (2009), for instance, by focusing on volunteered geographic information (VGI), show that poor parts of the city tend to have far less information created about them than wealthy areas.

Such uneven geographies of knowledge also only increase the importance of information created in the world's cores, and reinforce what Manuel Castells (2010) refers to as the black holes of informational capitalism. Castells argues that a powerful systemic relationship exists between economic and social exclusion and marginalization from practices and processes of information production and consumption. He went so far as to claim that the global diffusion of information and communication technologies is so uneven that “most of Africa is being left in a technological apartheid” (Castells 1999, 3).

Augmented Realities or Alternative Realities

These highly uneven geographies of codified information matter because they shape what is known and what can be known, which in turn influences the myriad ways in which knowledge is produced, reproduced, enacted, and re-enacted. Because of the ways in which the digital and the material (or the virtual and the real) are increasingly woven together to produce augmented spaces, digital geospatial information thus becomes not just useful in the abstract, but amplifies and actively shapes how we understand, and interact with, the spaces and places that we inhabit and move through.

Geospatial information has become embedded in place itself. Dodge and Kitchin (2005), for instance, highlight how code (or software) can help to bring into being ‘code/spaces,’ in which code and geocoded information dominates the ways that space is produced, and ‘coded spaces,’ in which code is embedded in, but incidental to, space. Very few aspects of everyday life are not ultimately reliant on, or produced by, code (Thrift and French 2002, Kitchin and Dodge 2011), and code therefore plays an important role in the ‘reiterated digital practices that create space anew’ (Wilson 2011).

Graham et al. (2012) expand on these themes to highlight the role of not just code, but also geocoded content, in everyday, lived, geographies that are enacted. They use the term *augmented realities* to describe “the indeterminate, unstable, context dependent and multiple realities brought into being through the subjective coming-togethers in time and space of material and virtual experience” (Graham et. al. 2012: 465). In other words, they flag up the need to pay attention to ways that everyday life is experienced in conjunction with, produced by, and mediated by digital and coded geographic information[1] that helps us to understand, enact, re-enact, produce and reproduce place[2].

Places that are left off the map of knowledge thus become absent from our understandings of, and interactions with, the world because of the ways that geographic content, geospatial information, maps, and mappings increasingly form integral parts of our everyday movement, understandings, and interactions. This is especially important in parts of the world characterized by highly uneven power relationships: allowing dominant groups to fix distinct forms of representations onto otherwise contested places while maintaining an outward appearance of rationality and objectivity (c.f., Leuenberger and Schnell 2010). Brunn and Wilson (2013) and Graham and Zook (2013) have already demonstrated the power of geospatial content to reinforce power in highlighting contested parts of the world (a South African township and Jerusalem, respectively). However it remains that there has yet to be any large-scale empirical analysis of the factors that explain information geographies at the global scale.

[1] For instance, the tweets, Wikipedia articles, digital maps, geotagged photographs, reviews, videos, streetviews, descriptions, and map elements that augment our spatial experiences.

[2] It is important to point out that because of the ways that geocoded information is often used in portable, mobile devices, it can also augment the ways in which we bring place into being whilst we are in place, enacting place.

Promises of Changing Connectivity

While mass media appear to reinforce dominant narratives about place, alternative media and a democratization of connectivity could in theory allow for a much greater diversity of voice and geographic representation. Ideas and practices like volunteered geographic information (VGI), user-generated content, peer-production, and participatory GIS all appear to provide unique entryways into leveling the geographically uneven representations of place.

Because Internet penetration rates are increasing rapidly, and because there are now over two and a half billion Internet users (a majority of whom live in the Global South), many commentators now see the potential for a significant global shift in the ways that information and knowledge is made, shared, and used. There are hopes that open platforms like Wikipedia have knocked down many traditional barriers to participation and sharing, and allowed a plethora of new voices from the South to be heard within what Wikipedia's founder refers to as the "sum of all human knowledge" (Slashdot 2004; Graham 2011). Sui and Goodchild (2011, 573) similarly note that VGI "might be one of the most important phenomena to impact our discipline in recent years and one that could dramatically alter the landscape of geographic information production".

Every day on Wikipedia hundreds of thousands of people write and edit articles, submit images and videos, debate the contours of knowledge, and collaborate on an encyclopedic range of topics. Wikipedia's audience is in many ways engaging in a positive feedback loop: Wikipedia is both one of the world's most accessed websites and almost always[3] appears prominently in search results. This confers a high degree of visibility to the site, which in turn draws in ever more contributors.

Speaking about the possibilities afforded by the Web at the *World Summit on the Information Society*, Harvard Law Professor Lawrence Lessig (2003) asserts that "[f]or the first time in a millennium, we have a technology to equalize the opportunity that people have to access and participate in the construction of knowledge and culture, regardless of their geographic placing." 'Commons-based peer production' (Benkler, 2007), 'produsage' (Bruns, 2008), and even 'citizen journalism' (Deuze et al. 2007) are all similar ideas that point to a potential global democratization of participation and information production. Jenkins (2006) perhaps expressed such sentiments more clearly when he points to the collaboratively created culture that we can now supposedly construct in a democratic process (see also Tapscott and Williams 2006, and Shirky 2011 for similar sentiments). To all of these authors, it is an absence of connectivity and a 'digital divide' that is the central obstacle to a democratization of participation.

Such proclamations are often made in the absence of empirical data and spatially oriented inquiry. This article therefore aims to address this gap through a study of the geography of representation in Wikipedia. Wikipedia is by far the world's biggest and most used encyclopedia, and 1,600 times larger (in terms of number of articles) than the Encyclopedia Britannica. Its size and reach allow it to be thought of as a platform for what Haklay (2013b) defines as distributed intelligence and participatory science. The encyclopedia is now so

[3] Wikipedia appeared on the first page (top ten) of 99 percent of 1,000 Google search results for nouns, in position five or better for 96% of searches and in position one for 56% of searches (Silverwood-Cope 2012)

popular that fifteen percent of all Internet users access it on any given day. It exists in 282 languages; 40 of those language versions have over 100,000 articles, and the English one alone contains over four million (Wikimedia 2013). Furthermore, we see that it is one of the top-twenty websites in ninety-five percent of the world (Alexa 2013), indicating the true global reach that information in the platform has.[4] This is not to say that information within Wikipedia is necessarily used, accessed, and valued in the same ways everywhere.

The prominent role that Wikipedia fills in contemporary 'information ecologies', has resulted in a rich vein of research. Pasley et al. (2008), for instance, looked at the representation of Great Britain in the English Wikipedia and found the coverage to be geographically clustered and correlated with population. Some work has also been conducted about the variable amounts and types of representation of places in different languages. By looking at anonymous edits to Wikipedia Hardy et al. (2012) and Hardy (2013), found that there is generally a decreasing likelihood of edits to geotagged articles with increasing distance between editor and article.

Yet, there are also hints that important outliers to this general trend exist. Ahlers (2013), for instance, compared coverage of Honduras in the English and Spanish Wikipedias and found 20 percent more English articles about the country. Graham et al. (2012) have published preliminary results showing similar patterns at the global scale. They reveal that while most European and East Asian countries have more Wikipedia articles about themselves in their dominant language, we see more English-language articles than local-language articles about much of the Global South. These geographic differences in the coverage of different language versions of Wikipedia matter, because, as Graham and Zook (2013) have demonstrated, fundamentally different narratives can be (and are) created about places and topics in different languages. Others have found that Wikipedia offers an uneven view of the world not just in terms of geography, but also gendered content (Lam et al., 2011), and history (Luyt, 2011).

As such, despite Wikipedia's structural openness, there are fears that some parts of the world will be heavily represented on the platform and others will be largely left out (Hecht and Gergle 2009, Haklay 2013a, Sieber and Rahemtulla 2010, Elwood 2010): a situation that could simply reproduce worldviews and knowledge created in the Global North at the expense of Southern viewpoints (e.g. Ford, 2011). These second-generation digital divides are not merely the divides of *access* as was so clearly considered in the late 1990s, but gaps in *representation* and *participation* (Hargittai and Walejko 2008). Despite these initial results, we know very little about what factors produce such uneven geographies in a world where we have 2.5 billion internet users and mass adoption of Wikipedia as one of the most prominent repositories of knowledge.

It is at the intersection between these debates about geography, technology, representation, and the promises of changing connectivity that this paper positions itself. Specifically, by drawing on work about the geographies of knowledge and augmented realities, the paper is able to empirically explore the difference that the recent, potential, democratization of voice

[4] This figure was derived by looking at the list of five-hundred most visited websites for each of the one hundred and twenty countries and territories for which data are collected. The only countries in which Wikipedia fell outside of the top-twenty most popular sites are: China (126th), Egypt (22nd), Cambodia (29th), Mongolia (35th), The Palestinian Territories (29th), Vietnam (24th).

has made to much older practices of knowledge production. As the first step in an assessment of the inequalities in the global system, this article thoroughly investigates representation in Wikipedia globally, using the country level as the scale of analysis. Besides the examination of raw numbers, we establish a baseline using regression models that both hint at necessary conditions for representation and allow to more specifically focus the analysis on the outliers, i.e. countries that fare considerably better or worse than expected. We place an especially strong focus on the Middle East and North Africa (MENA) and Sub-Saharan Africa (SSA) because of existing broad concerns about voice and representation from and about these regions (cf. e.g. Aoragh 2011; Graham, Hale, and Stephens 2011). Despite a recent rapid increase in Internet access, there are indications people and places in those regions remain largely absent from websites and services that represent the region to the rest of the world.

The core concern in this work is that a relative lack of voice and representation serves to reproduce the power of people and places at the center of geopolitical mass-culture, thereby creating a global dependency on voice, culture, and cultural industries emanating from the Global North (Osborn 2010). That is, the tone and content of Wikipedia as a globally useful resource that represents a country or region, in many cases, is potentially being determined by outsiders with misunderstandings of the significance of local events, sites of interest and historical figures. Furthermore, in areas with substantial social and political conflicts, participation from local actors potentially enables people to ensure that a diversity of perspectives are present in content about contentious issues.

Within these contexts, we pose two questions:

1. *What are the geographies of spatial representation on Wikipedia?*

We observe uneven levels of geographic representation globally. These spatial patterns lead us to pose our second, and more important, question:

2. *What factors explain this geography?*

We should be explicit in pointing out that we do not approach these questions from perspectives of technological or social determinism. We instead begin from a perspective in which technology and society co-construct each other thereby allowing us to focus on how infrastructures, technologies, and social practices are (re)produced in order to create and sustain geographies of information. Following contemporary work in Science and Technology Studies, we consider how technologies of participation enable certain actors to have a voice in novel ways (Callon and Law 1997). This is not to say such technologies are liberatory as new actors come to participate. Rather it is to acknowledge that the insertion of participatory technologies prompts a reassessment of the relationship between who produces knowledge and who is marginalized in such production.

Despite hopes that are often invested in Wikipedia as a platform that can flatten uneven and concentrated information geographies, this paper ultimately demonstrates that there is a vast chasm between the necessary and the sufficient. Open platforms and internet connectivity are almost always necessary for participation in global knowledge production. But in the chasm separating the necessary and the sufficient are a host of pitfalls such as language barriers, connection speeds, technical impediments, cultural differences in what constitutes authority and in attitudes towards knowledge sharing, diffusion of the *notion* of participation, and a host of other factors. Some countries (or groups of editors from countries) have crossed that chasm with gusto, filling in gaps. Other countries are almost entirely being written about by others (or largely not being written about at all), and are thus subject to the voices, attitudes and biases of those outside the country.

Data sources

In this article, we are not concerned with the length or quality of the articles representing places, but their mere existence and thus, the visibility of places. Georeferenced articles are any Wikipedia article about a place or event that happened, is happening, or will happen in a place. These articles can, in theory, be created by anyone with an internet connection about anything notable on our planet.[5] In order to address our two research questions, we employ three primary data sources: geotagged Wikipedia articles, country-level indicators from the World Bank and Wikipedia usage statistics.

Wikipedia articles: We obtained a list of geotagged Wikipedia articles in 44 languages by combining results from a computer script that we devised to extract geotags from two geocoding projects: WikiLocation (2013) and WikiProjekt Georeferenzierung (2013). We took data from the most recent data dumps in November, 2012. Both projects provide data on a geotag (rather than article) basis. We, however, are interested in articles as an intermediate unit of analysis some of which feature multiple coordinates. Our spot-checking revealed three predominant reasons that could cause multiple coordinates to exist in an article. The first type are pages that simply list other geospatial entities, such as hot air balloon sites, every village in Yorkshire or every meteor crash site in America. We excluded these pages as they do not, on their own, facilitate representation of a geospatial entity. The second reason is a place that references other places in content. The third reason is that a place refers to a route, and thus has at least two geotags, the start and end point.

To mitigate these issues, we first filtered articles to those with four or fewer geotags (in order to exclude lists). Then if a geotag occurred more than once, we assigned that geotag to the article. If every geotag is unique, we used the geotag that occurred first, as it tends to be in a privileged location (such as an infobox or the initial description of a place). In total, we obtained 3,924,308 articles in the aggregated and cleaned data from Wikipedia.

We should point out that it is very common for articles to have multiple matching geocodes. For example, the article on Cairo lists the coordinates in the upper right corner and again in the infobox. As might be expected, these coordinates are the same. There are however case when the coordinates differ. Of the 3,924,308 articles in our database, 113,153 had duplicate coordinates. A further 13,673 had three coordinates, 2913 had four, and 22,195 had more. It is only the 22,195 articles with five or more geotags that were excluded (or 0.006 percent of the total number of articles). This set of articles primarily comprises of lists of places. The largest is in Danish, as a list of monuments in Thisted, Denmark.[6] The

[5] It could be argued that some places (e.g. Rome or Angkor Wat) are characterized by more inherently describable sites of interest. However, the fact that Wikipedia both allows and encourages mundane and everyday places and processes to be represented would indicate we shouldn't necessarily assume an a priori privilege to sites of archaeological or touristic significance.

[6] http://da.wikipedia.org/wiki/Fredede_fortidsminder_i_Thisted_Kommune

monuments in Thisted do not feature a history or discussion, but merely a list. However, they do all link to Kulturarv.dk (presumably where the list's data originated). Because of the long tail of lists of geographic entities, these 22,195 articles actually indicate 900,297 different geocodes. Our analysis would therefore look substantially different if they were included, despite the fact that these entities would not necessarily have their own page and thus their own clear means for representation.

We further removed numerous articles out of concerns for data quality and accuracy. This includes articles that referred to locations outside earth (e.g., referencing the lunar crater, Tycho) as well as articles whose coordinates appeared suspect (e.g. where latitude equaled longitude which is typically an error). We then joined a dataset of national boundaries with a 5 mile buffer around all bodies of water (accounting for data imprecision and allowing for near-shore articles about e.g. lighthouses to be counted) to the remaining data spatially in a Geographic Information System and thus assigned each geotagged article the country it falls within.

Country-level indicators: In this analysis, we selected four variables that describe differences between countries: Population, GDP per capita, gross enrollment ratio, and total number of broadband connections. Each measurement was taken from World Bank's 2011 national level data.[7] Each one of these can theoretically make a difference in the presence of geotagged content and have been either associated with content or prophesized to bring about more content.

Population could signify at least two different reasons for the level of geotagged content. First, as noted by Pasley et. al. (2008) with regards to Wikipedia in Great Britain, greater numbers of people tend to imply more settlement and theoretically more sites to write about. Second, more people may mean more individuals inclined to write (local) content in the first place. *GDP per capita* can be used as a rough proxy for a host of necessary ingredients for Wikipedia editorship, such as leisure time, access to computing resources, and access to local information sources (like libraries). *Gross Enrollment Ratio* (GER) calculates the number of those enrolled in school relative to the population of 5-17 year olds. It indicates the share of the population that could plausibly create and edit a Wikipedia page. Having an Internet connection is a necessary condition for editing Wikipedia. We therefore employ *broadband Internet connections* as our final country-level metric. It is likely that faster, i.e. broadband, Internet connections create a qualitatively different relationship to user-generated content when compared to non-broadband connections.

Wikipedia usage statistics: In addition to country-level indicators, we test whether the presence of an active editing community is associated with greater geographic representation. This is secondary to our key analysis on population, GDP, education, and broadband. However, it makes a crucial link in our argument: by claiming that national level indicators make a difference, particularly broadband, we are implicitly assuming that it is the locals that create this content. The use of Wikipedia statistics tests this explicitly.

Every three months, Wikipedia provides a country-level location of every 1000th edit to the website.[8] We take the mean of all quarterly measurements from 2007 to 2012 to be an

[7] http://data.worldbank.org/use-our-data

[8] http://stats.wikimedia.org/wikimedia/squids/SquidReportsCountriesLanguagesVisitsEdits.htm

estimate of the how frequently edits come from a country (i.e. the average per quarter). Because of Wikipedia's sampling method, countries with particularly low editing frequencies will sometimes show zero edits per quarter. We correct for this bias by taking the average of the country's share in the global number of edits in the previous and successive quarters. This interpolated percentage is then applied to the global number of edits in the quarter at hand to estimate the missing number of edits.

In general, the quarters correlate very strongly, but for low editing countries there is a great deal of variance between any two quarters. By taking the average of 2007 to 2012 we simultaneously make the distribution much smoother as well as account for the fact that Wikipedia is a cumulative project across many years.

For countries where there are very few edits, our data will be inevitably proximate. Even by measuring six years of Wikipedia data, there are approximately fifty countries with no edits to Wikipedia in the dataset. Most are quite small in terms of population (such as St. Lucia), although several countries appear to have zero edits with large populations.[9]

Geography of representation

With the data described above, we begin by descriptively examining the geography of information in Wikipedia in two respects: total number of geocoded articles in our data set and a per-country map of which language contains the most geocoded articles.

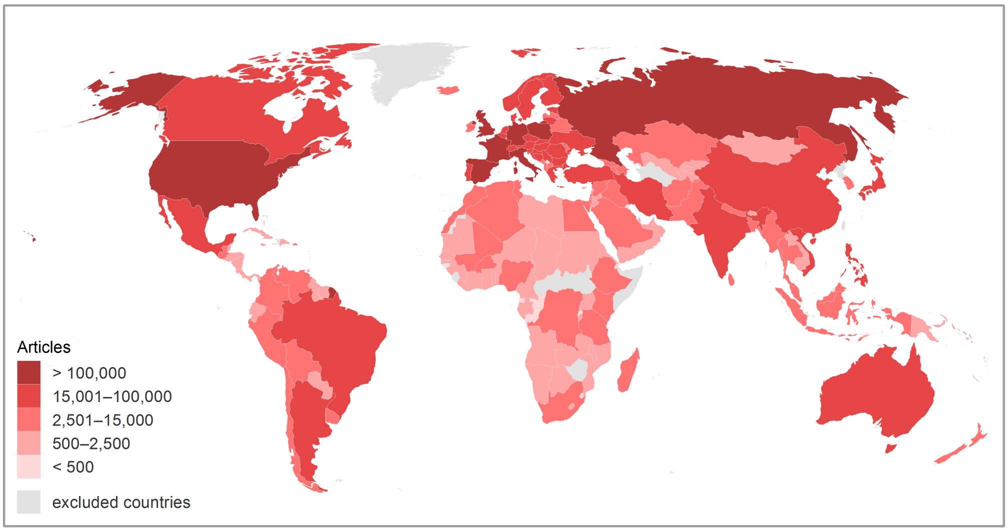


Figure 1: Number of geotagged Wikipedia articles per country

[9] In the 25 most populated countries with no edits, only North Korea, Tajikistan, East Timor, Palestine, Kosovo and Laos are not from Africa. The absence of edits from the Palestinian Territories is due to all Palestinian IP addresses as being recorded as from Israel. As we only have access to aggregated data, we cannot disentangle this and instead must simply note that Israel will be slightly over-estimated and Palestine underestimated in any model dealing with edits.

Figure 1 displays the number of articles across all captured languages per country. There are a staggering number of articles in the United States (564,084 in total in our dataset, 279,287 of which are in English) and tens of thousands in many European countries, Japan, Australia and India. However, there are also far fewer in much of the rest of the world. The most apparent asymmetry is perhaps between the global North and South, although many countries provide a modest exception to this norm.

But, importantly, not only are some parts of the world massively under-represented on Wikipedia, but a lot of the content that does exist tends to be in only a few languages. Specifically, Figure 2 demonstrates that many countries in the Global South have more articles in a non-local language (often the language of a former colonial power) than a commonly spoken local language. In other words, we see a broad pattern of the Global North being represented in local languages while the South is largely being defined and described by others.

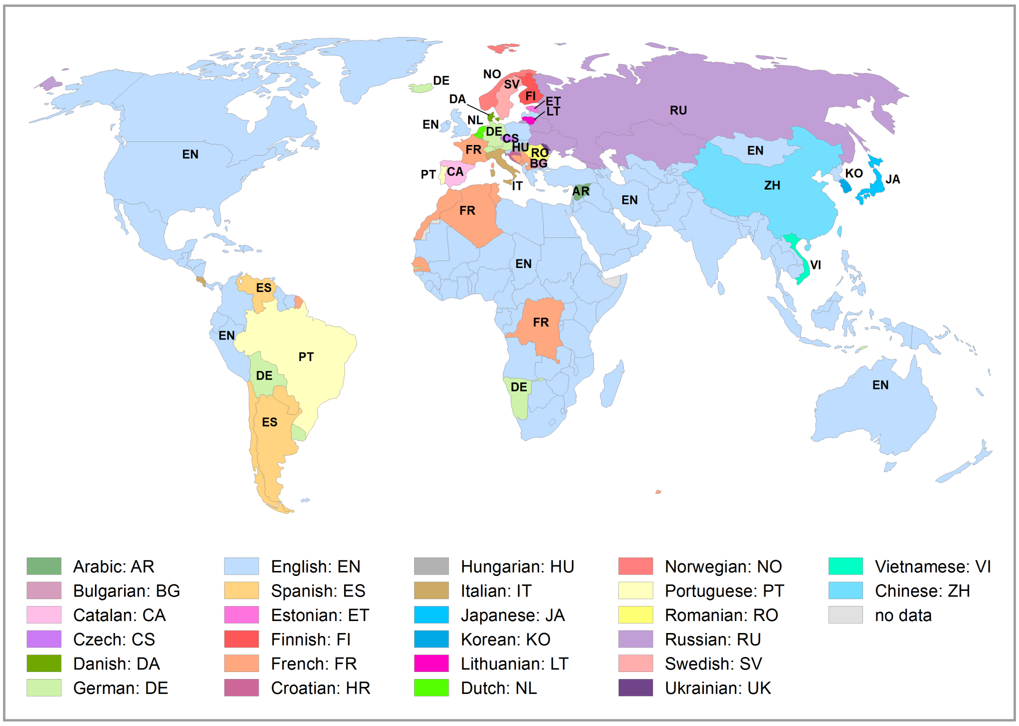


Figure 2: Dominant language of Wikipedia articles (by country).

The map shows that almost every European country has more articles about itself in its dominant language than any other language (e.g. there are more articles in Czech about the Czech Republic than there are English articles about the country). But we do not see that pattern across much of the South. English is dominant in much of Africa, the Middle East, South and East Asia, and even parts of South and Central America. French is dominant in five countries in Africa (although some traditionally Francophone countries like the Ivory Coast still have more content in English). German is also dominant in one former German colony (Namibia) and a few other countries scattered around the world (e.g. Uruguay, East

Timor). In the Middle East, we see only one country (Syria) with more articles in Arabic than any other language[10].

In sum, not only is most of the world's content written about global cores, but even of the relatively small amount of content produced about the rest of the world, much of it exists in just a few languages. The following section is dedicated to uncovering what factors might produce these very uneven geographies.

Explaining the geography of representation

In what follows we explain some of these patterns using a series of multivariate regressions. Figure 3 shows the distribution of geocoded articles per country within all included languages. It has an approximately lognormal distribution. To note, specific valleys in the bell curve are a result of binning what is a very long tail on a non-logarithmic axis. These lognormal distributions will persist throughout most of the analyses.[11]

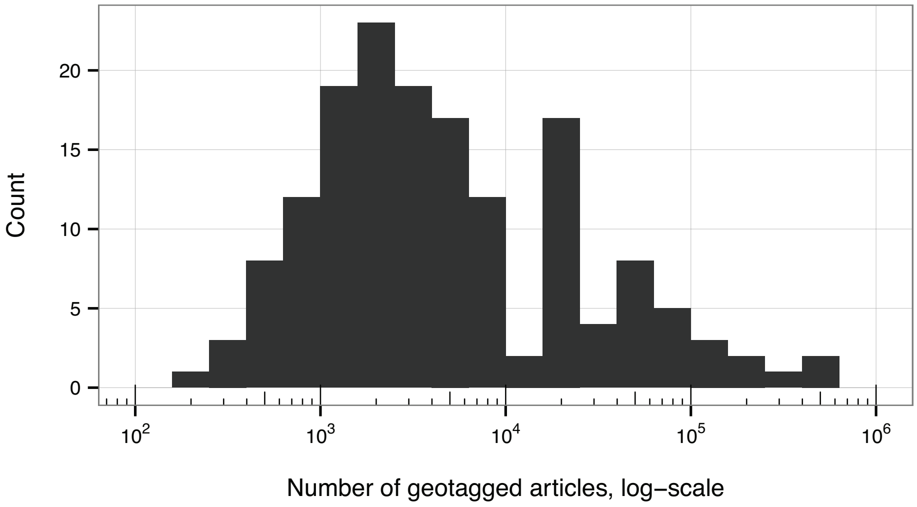


Figure 3: Number of geotagged articles per country.

Employing these data, we wish to make two overarching points: The first is that the vast majority of differences between places can be explained by relatively simple factors relating to access to the Internet. The second claim is that the representation of place is a decidedly local affair. People edit about their local area, albeit not exclusively.

[10] To illustrate the scale of the cross-linguistic representational differences, it is worth pointing out that there are even more English articles about North Korea than there are articles in Arabic about Saudi Arabia, Libya, the UAE and many other countries in the region.

[11] We also offer a word of caution in that the interpretation of such log transformed data masks the vast differences between the countries that are the least and most visible. An increase of 1.2 in a regression model or on a distribution thus means a difference of an order of magnitude of 1.2 on a log10-scale. Thus an increase of 1.2 from 10 articles means a 60% increase ($10^{1.2}$ =15.8), not a 20 percent increase.

To explore our claims, we first draw upon national level statistics and, second, on pre-processed edits data, as explained in the section on *data sources*. In our analyses, the number of geocoded articles per country is the response (or dependent) variable. We hypothesize that all explanatory (or independent) variables will predict positively with the number of Wikipedia articles. The explanatory variables that we use are: *population, GDP per capita, gross enrolment ratio, broadband internet connections,* and number of edits to Wikipedia.

<Table 1 about here>

Table 1 shows basic statistics of all variables in non-transformed state. Note the often high (positive) skewness reflecting the uneven distributions of most of the variables which led us to $\log_{10}$-transform them for the regression analyses. All variables except GER are $\log_{10}$ transformed, and thereafter tend to show roughly normal distributions. With the transformation, we are able to mute the leverage of countries scoring especially high on any of these variables (such as China in population and France in the number of Wikipedia articles). This approach helps to stabilize our models and increase robustness.

We hypothesize that all independent variables will be positively related to the presence of geotagged content. We further hypothesize that there is a relationship between the immediacy of the indicator to the editing process and the strength of the predictor. For example, the number of individuals connected by broadband is more immediately related to the capacity to edit than the total number of people. Thus, we would predict broadband to have a stronger effect than total population. The number of edits would have a stronger effect still.

<Table 2 about here>

Table 2 shows the bivariate correlations of all variables under investigation. Many of the correlations are particularly high, with two relationships as high as 0.8. These correlations are not the result of specific outliers but broad trends across the entire distribution. We feel confident in saying that across the entire world, there is a strong relationship between edits from a country and the number of geocoded articles in that country. While the presence of this relationship is not a surprise, the strength of the association is.

Examining the mutual correlations between the independent variables points out potential challenges in employing them in the regression framework: GER, GDP p.c. and Broadband show strong mutual correlations. Thus, we need to take particular care with respect to multicollinearity. Some of the challenges in this area are mitigated already, since we include GDP in its *per capita* form to operationalize the levels of development.

Simply using bivariate correlations, it is difficult to disentangle the relative influence of the variables in Table 2. We employ a forward selection OLS linear regression framework to predict the total number of Wikipedia articles and to assess the importance of the variables for the prediction.

<Table 3 about here>

Model 1

Model 1 can be characterized as follows:

$$log(A) = \beta_0 + \beta_1 log(Population) + \beta_2 GER \ + \ \beta_3 log\left(GDP_{p.c.}\right) + \ \beta_4 log(Broadband) + e$$

This model includes our basic "conditions of possibility" variables: population, GDP p.c., Broadband and Gross Enrollment Ratio (see Table 3). This model explains a substantial 71 percent of variance in the data (ie. adjusted R^2 = 0.71). In this model, Broadband and population emerge as significant predictors, GDP and GER do not.

Unfortunately, this model suffers from a key problem given the assumptions of ordinary least squares: multicollinearity. Conventionally, a VIF score (variance inflation factor) of 5–10 spells caution, and above 10 indicates serious problems. We do not have scores that high. However, since there is such a high correlation between GDP and broadband and both have very high VIF scores, we believe we ought to exclude one of these variables. Thus, we employ a forward selection procedure at $p < 0.05$ in order to ensure that the model stays simple, and thus more stable and robust.

Model 2

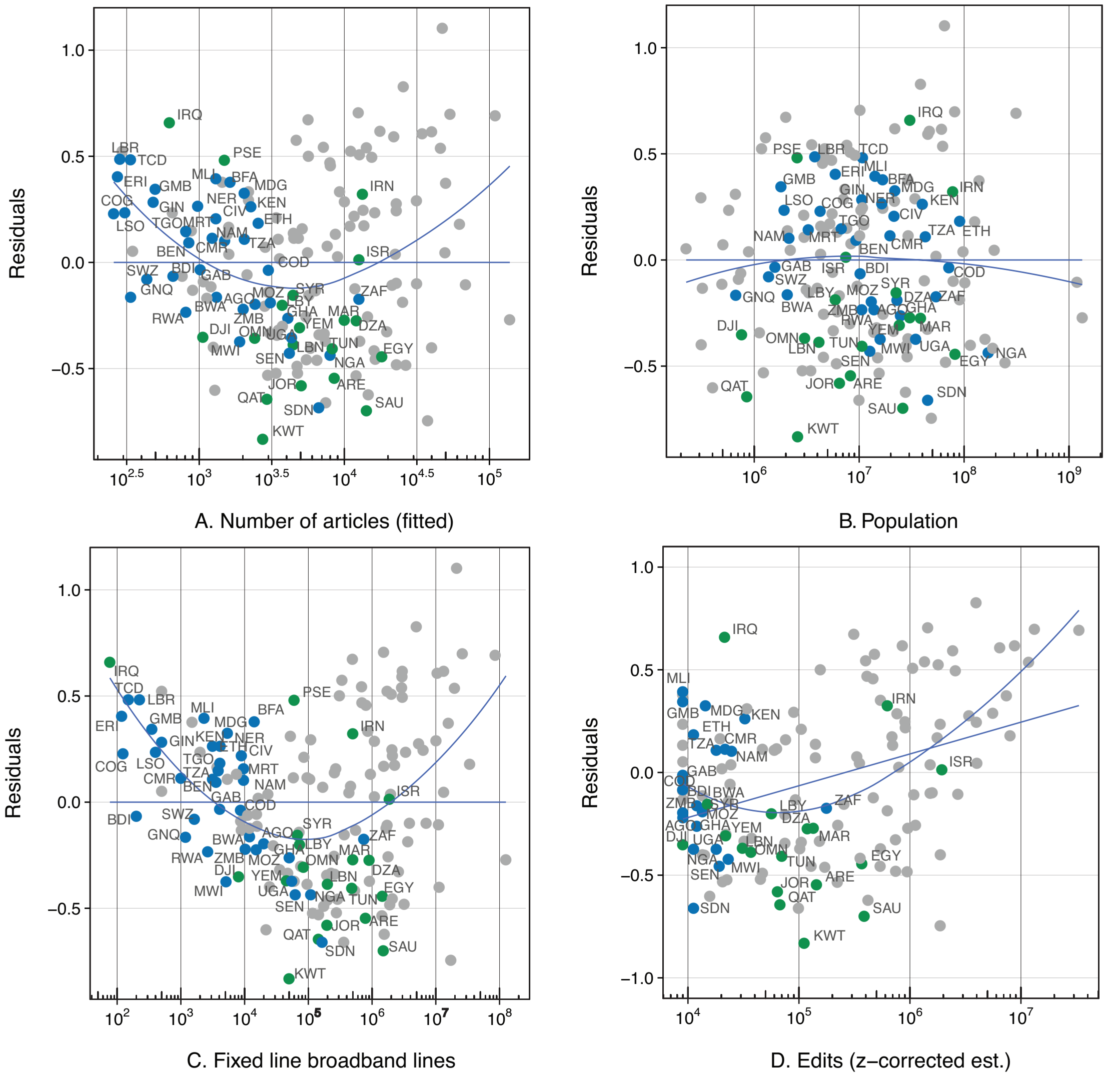
In Model 2 we include edits, the only independent variable that is endogenous to Wikipedia, alongside population and broadband (see Table 3). This model can be characterized as follows:

$$log(A) = \beta_0 + \beta_1 log(Population) + \ \beta_2 log\left(Broadband_{p.c.}\right) \ + \ \beta_3 log(Edits) + e$$

The inclusion of edits may raise concerns about endogeneity. To note, we do not perceive this endogeneity as theoretically problematic since articles are only created once, but there are many edits from any given country. Moreover, such edits can be sent to non-geographic articles as well as geotagged articles.

This model fits almost as well as the previous model (adjusted R^2 = 70.3 percent), but has substantially lower VIF scores all well below critical thresholds. The contribution of edits is considered significant at a level of $p < 0.03$.

This model ultimately points out that from all our considered independent variables, *population*, *access to broadband Internet* and the *number of edits to all Wikipedias* originating from a country explain a large part of the variance in the number of geotagged Wikipedia articles.



A. Number of articles (fitted)

B. Population

C. Fixed line broadband lines

D. Edits (z−corrected est.)

Figure 4: Scatterplot of residuals versus observed: A. Number of articles, B. Population, C. Fixed line broadband connections, D. Z-corrected edits to Wikipedia.

Linear models, however, produce linear expectations across distributions. That is, we assume that an increase in broadband among those with little broadband is going to have the same effect as a proportionate increase among those with much broadband. However, we do not find this to be the case with this model. In the residuals versus fit plot for all the independent variables (see Figure 4), we observe a notable curve in the distributions. We show this distribution using a LOWESS curve to plot a line of best fit through the distribution of countries. This curve appears in the full model (Figure 4A), where we plot predicted versus observed values for the full model. This curve does not appear in the population versus predicted, nor the number of edits per predicted article. However, the shape of this curve appears in the distribution of predicted articles versus broadband connections (Figure 4C). In particular, those countries with the least and most broadband have more articles than expected, whereas those countries in the middle of the distribution have fewer articles than expected. We believe this is a key insight related to narratives of the digital divide and user-generated content online. We discuss this below. However, we also wanted to validate the significance this intuition. So prior to the discussion we introduce a third model. It is the same as the previous model except with a curved parameter for broadband.

Model 3

We characterize model three as follows:

$$log(A) = \beta_0 + \beta_1 log(Population) + \ \beta_2 log(Broadband) + \ \beta_3 (log(Broadband))^2 \\ + \ \beta_3 log(Edits) + e$$

The addition of the curved parameter clearly increases model fit (Adjusted $R^2 = 0.76$) and homoscedasticity (i.e. randomly distributed residuals). Since Broadband is included both as such and as squared term, VIF scores are clearly elevated, but this is an expected result of such a model and is not cause for concern. All explanatory variables including the curved broadband parameter (and for the first time, the intercept) are significant. This validates our intuition that the presence of broadband connections is key to the emergence of user-generated content but is very unevenly felt around the world; more broadband does not necessarily or immediately lead to more representation even if broadband is a necessary condition for such representation.

Discussion

A country-level analysis of Wikipedia has revealed that only a small number of variables are needed to explain the bulk of variation in the presence of geotagged articles across the world. In our final model, over three quarters of the variation in geotagged articles was explained by the population of the country, the number of fixed broadband connections and the number of edits emanating from that country. We omitted two other variables that were previously tested for a combination of non-significance and lack of robustness: the country's Gross Enrollment Ratio and the per-capita GDP.

The inclusion of population is telling: it is, as we suggested, a dualistic variable insofar as it pertains both to content worthy to cover in Wikipedia and, possibly, to presence of people to edit articles. The former fact, we can hypothesize, is linked to settlements as areas affording Wikipedia coverage and, maybe less so, a pool of people about whom Wikipedia articles can be written and geotagged. In terms of editors from outside a country, population size may also play an important role for making a country visible enough to inspire and motivate such outside editing activities. That is, potentially people edit more about more populated countries, which may play an important part in world politics, be a partner in trade or a destination for tourism. With the current analysis it is difficult to further disentangle these multi-faceted meanings of population as a factor. Further studies may want to clarify e.g. the question if (to what degree) built-up areas attract more articles and if this is something that we can observe globally.

Broadband has the most influence in the final model, but this influence is not linear. In Model 3, we examined the residuals versus the fitted model in order to show that an effect (that we later attributed to broadband) led to a U-shaped distribution in the expected number of articles. We revisit that distribution below with an eye to two regions of the Global South that have characteristically been represented by the Global North: Sub-Saharan Africa (SSA) and the Middle East and North African region (MENA).

We find in Figure 5 that many of the countries with the lowest (negative) residuals, that is, those countries which have considerably fewer articles than predicted by Model 3, tend to be in the MENA region. Recall that this work is on a log scale, so being near -1 means that a country (such as Kuwait) has nearly a full order of magnitude fewer articles than would be expected given its population size, broadband access, and number of Wikipedia edits.

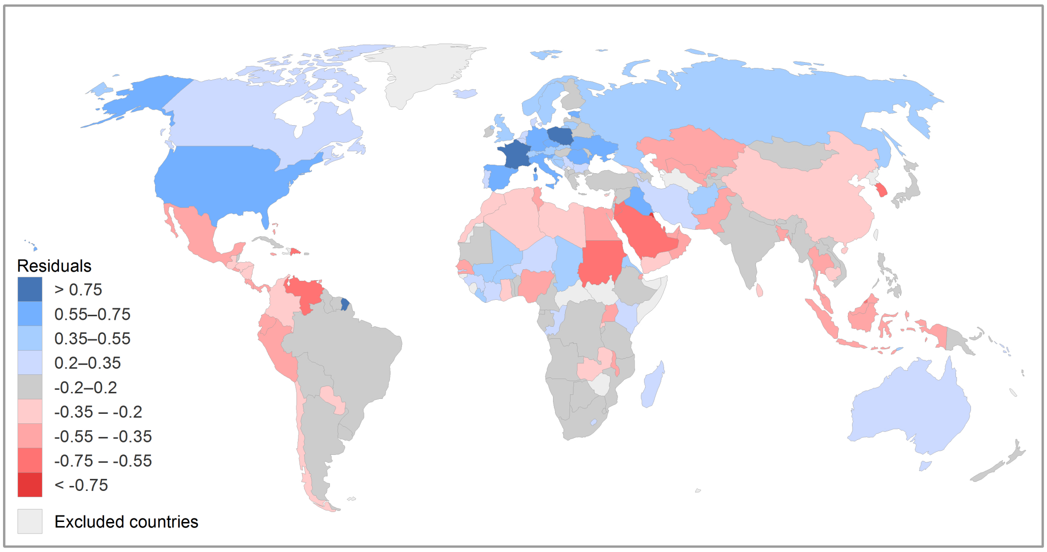


Figure 5: Residuals of Model 3 (Population, Broadband p.c., Edits). Countries with relatively small residuals (approximately up to +/-1 standard deviation) are colored in grey, while countries that were excluded from regression due to missing values are colored in light grey.

However, while the overwhelming majority of MENA countries have negative residuals, this is not the case for all countries. Iraq, Iran, Palestine, and Israel exhibit positive residuals, i.e. they feature more geotagged articles than our model would predict. In the case of Iraq we suspect that in the wake of recent conflicts the region receives considerable attention and thus Wikipedia contributions from abroad. In Israel and Palestine similar factors may be at work, especially regarding edits from abroad. Additionally, the data for Palestine is affected by the way edits are measured by the Wikimedia Foundation as described earlier in this article. In the case of Iran, the circumstance that many people in Iran use the Internet via proxy servers (and therefore would not appear in the Wikipedia edits dataset) may have affected the model predictions and the residuals by artificially lowering the number of edits we see from this country.

Similarly underrepresented are some countries in SSA, for example, Sudan, Nigeria, Senegal, Uganda, Ghana and Malawi. However, looking at SSA as a whole, most countries are found in the leftmost third of Figure 4a. Thus, they are (very) scarcely represented on Wikipedia in general, but not necessarily underrepresented per the expectations of our model. Indeed, approximately half of all SSA countries have positive residuals, i.e. they have more geotagged Wikipedia content than the little content one would expect given their population size, broadband connectivity and number of edits. This may be the result of external individuals seeking to write about these areas for the sake of completeness.

The contrast between SSA countries and MENA countries point to a challenging dilemma. Simply examining SSA countries we might argue that the most immediate route to increased online representation would be through an increase in internet connectivity (i.e. the number of broadband lines). This has precisely been the strategy for many countries in the MENA region, and particularly Gulf States such as Qatar and the United Arab Emirates which seek to grow their ‘knowledge economies’. Yet, these countries do not appear to have kept pace with the expected level of geotagged articles on Wikipedia.

The significant curvilinear effect of broadband suggests a notable hurdle faced by many countries. With little or no broadband, many countries will have external authors write articles (albeit short ones) on major cities and landmarks. However, more people in front of a broadband-connected computer does not immediately translate into an proportionate increase in local articles. We illustrate these stark asymmetries in Figure 6 showing countries with extensive and modest broadband connections colored by their residuals in Model 3. This is further compounded by the slightly curved effect of edits – as more individuals edit, there is a disproportionate increase in the number of new geotagged articles that appear.

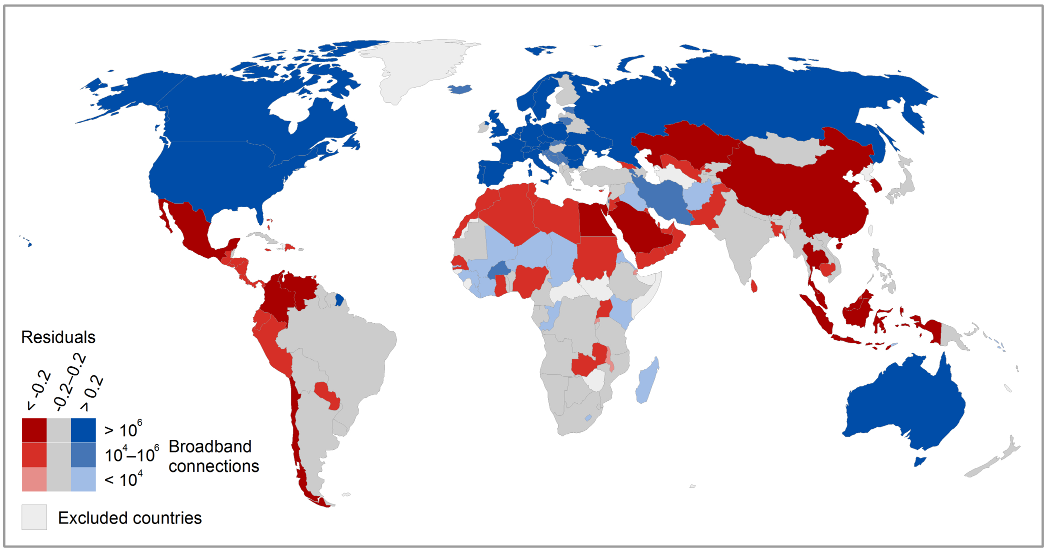


Figure 6: Residuals of Model 3 (Population, Broadband p.c., Edits) on a qualitative scale. Countries are additionally grouped with respect to the degree of broadband availability, where the cut-off between low and high availability has been defined as 10-2 broadband connections per inhabitant. Countries with relatively small residuals (approximately up to +/-1 standard deviation) are colored in dark and medium grey, while countries that were excluded from regression due to missing values are colored in light grey.

In interpreting this final model, we have come to a stark conclusion: representation is not occurring in a linear fashion, but one that is accelerating in a virtuous cycle for those with strong editing cultures in local languages thereby accelerating differences in the volume of online representation. For example, Britain, Sweden, Japan and Germany are extensively georeferenced on Wikipedia, whereas much of the MENA region has not yet kept pace given their levels of connectivity, population and editors. Thus, while some countries are experiencing the virtuous cycle of more edits and broadband begetting more georeferenced content, those on the periphery of these information geographies might fail to reach a critical mass of editors or even dismiss Wikipedia as a legitimate site for user-generated geographic content. Such a situation may lead potential editors in the Global South to consider Wikipedia as primarily the project of the Global North – a far cry from the edict of the “Sum of all Human Knowledge”.

Conclusions

The geographies of codified knowledge have always been uneven, and have always afforded some people and places greater voice and visibility than others. But, the rise of the geosocial Web with the opportunity for anyone with an Internet connection to annotate any part of the Earth’s surface promised a reconfiguration of information geographies. Furthermore, not only would platforms like Wikipedia allow for thicker augmentations of some of the world’s margins: but would also open up the processes and practices of authorship, allowing a greater diversity of voices, opinions, and narratives about any place.

Unfortunately, none of these promises have been realized. Even though Wikipedia consists of a massive cloud of geographic information about millions of events and places around the globe put together by millions of hours of human labor, it remains that the encyclopedia is characterized by uneven and clustered geographies: there is simply not a lot of content about much of the world.

There is a lot of geographic information created about North America, large parts of Europe, and most populated parts of Asia. This is not only a result of self-focus: Some of these regions manage to attract considerable amounts of allochthonous content, that is content in languages that are spoken only by comparably few in their territory and is thus likely contributed by editors in other countries. Few Wikipedia articles exist about places in Sub-Saharan Africa, the Middle East and North Africa and most countries in Latin America and Central Asia. Furthermore, when mapping some of the smaller Wikipedias like Arabic, Hebrew, and Persian, we don't see a similarly large cloud of information over much of the world. They present even more selective representations of the world than the English version of Wikipedia. What is perhaps most interesting about some of the smaller language-editions of Wikipedia is that it is not the Global North that vanishes from the map. It is rather other parts of the South that become absent: an observation that seems to imply a reproduction of the visibility of the already highly visible.

We know that Wikipedia is important to the construction of geographical imaginations of place, and content in the encyclopedia has immense power to augment our spatial understandings and interactions (Graham, Zook, and Boulton 2013). In other words, the presences and absences in Wikipedia that we reveal matter. As such, if a person’s primary free source of information about the world is the Persian or Arabic or Hebrew Wikipedia, then the world inevitably looks fundamentally different from the world presented through the lens of the English Wikipedia.

Seeking to better understand the patterns of Wikipedia content, we found that with just three factors we could explain a large part of the variation we see. Population, availability of broadband Internet and the number of edits originating from a country explain almost equal amounts of the variance in the layers of user-generated geographic information in the encyclopedia. That the number of edits is a significant independent variable despite its covarying with, primarily, broadband internet can be seen as attesting to the notion that local editors matter and, by extension, that much editing is done to local content. Interestingly, the three most important conditions for the existence of content about a place are, for the most part, smaller subsets of each other, but are all, in their own right, conditions for content generation.

Also of note here is the fact that while these variables help to explain the sparse amount of content written about much of Sub-Saharan Africa, most of the Middle East and North Africa still have quantities of geographic information below their expected values[12].

How do we explain the significant inequalities in the geography of user-generated information that remain after adjusting for differing conditions using our regression model? For example, despite high levels of wealth and connectivity Qatar and the United Arab Emirates have far fewer articles than expected. That said, although gulf state countries do now have increasingly high levels of connectivity, many such countries "leapfrogged" over dial-up and low-speed DSL lines.

The work certainly flags up the need for more sustained multi-methods inquiry into editing practices in Wikipedia (particularly in regions like the Middle East that are such outliers). However, from the statistical data alone, we get a sense of the conditions and affordances which are necessary, but not sufficient for the creation of digital content about a place. In other words, although we see the presence of a significant number of people, connected people, and Wikipedia editors as essential factors in the generation of autochthonous content about a place (i.e. content originating in the place that it is created about), each one of those conditions is characterized by its own enabling and constraining factors.

Constraints on the ways in which the populations of a place might relate to the amount of content about that place include a combination of the total amount of human sites, activities, processes, and practices of interest (which would be expected to increase with the population of a place), the nature of the broader information ecosystem (i.e. a broader canvas of written and visual content to use as source material[13]), and the total potential audience that any content might reach (editors may be more attracted to places that have a large audience). Societal attitudes towards learning and the sharing of information as well as towards Wikipedia as a platform also likely factor into the propensity of people in some places to contribute content to Wikipedia.

Constraints on the potential of broadband users to create content include the willingness of a diverse and connected proportion of the population to contribute as well as consume (age,

[12] It is again important to point out that these lower values are not because Internet users in those regions don't use Wikipedia: as the encyclopedia is one of the top-20 most accessed websites in almost every country on Earth (http://www.alexa.com/topsites/category).

[13] This could be both the availability of open government data which can act as seed content for new articles and a corpus of books, articles, and other media that can be employed as citations.

class, and gender are limiting factors in some places). These factors are situated within broader infrastructural and architectural constraints that might limit the efficacy of broadband Internet: such as power cuts, Internet usage through tablet or mobile devices, or the spaces and social settings in which the Internet tends to be used (e.g. at home, school, work, or public spaces).

Finally, constraints on the potentials of Wikipedia editors to create local content could consist of a lack of local Wikimedia chapters or groups (that encourage edit-a-thons and generally increase or sustain the motivation of editors), attractiveness of writing content about other places (e.g. for a larger audience or incrementally building on existing good quality content) or particularly contentious disputes in local editing communities (that divert time into edit wars rather than content generation). A related constraint, that can limit the amount of content that Wikipedia editors produce, is the structural inability of the platform itself to incorporate fundamental epistemological diversity. This is not a new observation: Elwood (2006: 695), for instance, has pointed to the difficulties of "including non-Cartesian, contradictory, or shifting forms of knowledge in a GIS." This fundamental barrier to the digitization of some forms of knowledge means that some groups of people are less likely to inscribe geographic information in codified formats: thus presenting Wikipedia editors with a lack of raw informational material from which to begin to create further (sourced) user-generated content.

These three constrains both independently matter and, through their coalescing, can work to reinforce each other and likely result in the below-expected amounts of geographic information that we see in some regions. What we might also be seeing in addition to, and compounding, the various enabling and constraining factors that we listed, is a *principle of increasing informational poverty*. Not only is a broader base of traditional source material (i.e. books, maps, and images) needed for the generation of any Wikipedia article, but it is likely that the very presence of content itself is a generative factor behind the production of further content. Said differently, although we recognize that digital content might be valued in very different ways by different social groups, the massive inequalities in content that we see could potentially reinforce some of those differences that we see. This consequently renders information produced about information-sparse regions most useful for people in informational cores (accustomed to integrating digital information into everyday practices) rather than people in informational peripheries.

Various practices and procedures of Wikipedia editing likely further amplify the effect. There are strict guidelines on how knowledge can be created and represented in Wikipedia: e.g. the ban on original research and the need to source key assertions. Editing incentives and constraints also likely encourage work focused around existing content rather than entirely new material. Editors creating new content need to consider article templates, formatting requirements, and the need to establish notability, among other concerns. Editing existing articles, on the other hand, is far more straightforward as the framework and general structure of the article is already established. In other words, the very policies and norms of the encyclopedia that govern its structure and its quality make it difficult to populate whitespace and terra incognita with geographic content.

In sum, this paper has revealed broad patterns in the unevenness of user-generated geographic information in Wikipedia and described the necessary conditions for content generation. Places without those conditions have been, and are, covered by relatively thin layers of informational representations. However, it appears that geographies of broadband access, instead of flattening the unevenness of information geographies, have amplified those pre-existing processes.

If we are to attempt to mitigate, circumvent, and reverse the enactment of patterns of increasing informational poverty, then we need to recognize that no one of the three conditions can ever be an entirely sufficient condition for the generation of geographic knowledge. For example, the building of broadband infrastructure could only ever be one part of a broader strategy to deepen the informational layers of places.

As the digital layers of places increasingly matter, it will be important to not only maintain focus on the geographic absences and presences in user-generated content, but also ask what the factors are that encourage or limit that content and its production processes. This paper has offered a starting point, demonstrating both the uneven geographies and the central factors that explain that variance at a national scale. It will now take much more sustained quantitative and qualitative inquiry into locally contingent challenges, barriers, inequalities, and deliberate exclusions for us to understand how to work towards more inclusive, more just, and more equitable representations and digital layers of our planet.

Acknowledgements

We are grateful to the help of numerous people who assisted with our multi-year efforts to map and measure the geographies of Wikipedia. In particular we would like to thank Ilhem Allagui, Claudio Calvino, Richard Farmborough, Heather Ford, Clarence Singleton, David Palfrey, Gavin Bailey, and Taha Yasseri, Adham Tamer, Matthew Smith, and Adel El Zaim. This work was also made possible by research grants from the International Development Research Centre and the University of Oxford's Fell Fund.

Bibliography

Ahlers D. 2013. Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles. In *Advances in Information Retrieval: 35th European Conference on IR Research, Moscow, 2013*, 668–671. Berlin: Springer.

Alexa. 2013. Wikipedia.org Site Info. http://www.alexa.com/siteinfo/wikipedia.org (last accessed 17 April 2013)

Almeida R. B., B. Mozafari, and J. Cho. 2007. On the Evolution of Wikipedia. In *Proceedings of the International Conference on Weblogs and Social Media, Boulder, 2007*.

Aoragh 2011 *Palestine Online: Transnationalism, the Internet, and the Construction of Identity*. London: I.B. Tauris.

Benkler, Y. 2007. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Brunn S. D., and M. W. Wilson. 2013. Cape Town's million plus black township of Khayelitsha: Terrae incognitae and the geographies and cartographies of silence, *Habitat International*. 39 284-294.

Bruns, A. 2008. *Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage*. New York: Peter Lang.

Callon, M., & Law, J. (1997). After the individual in society: Lessons on collectivity from science, technology and society. *Canadian Journal of Sociology-Cahiers Canadiens De Sociologie*, *22*(2), 165–182.

Castells M. 1999. *Information Technology, Globalization and Social Development.* Geneva: United Nations Research Institute for Social Development.

Castells, M. (2010) *End of Millennium (2nd Ed).* Oxford: Blackwell.

Craig, W. J., & Elwood, S. A. 1998. How and why community groups use maps and geographic information. *Cartography and Geographic Information Systems*, *25*(2), 95-104.

Crampton, J. 2008. Will Peasants Map? Hyperlinks, Map Mashups and the Future of Information. In *The Hyperlinked Society: Questioning Connections in a Digital Age*, ed. J. Turow and L. Tsui, 206–226. Michigan: University of Michigan Press.

Crutcher, M., & Zook, M. (2009). Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. *Geoforum*, *40*(4), 523-534.

Deuze, M., Bruns, A., & Neuberger, C. 2007. Preparing for an Age of Participatory News. Journalism Practice, 1(3), 322–338.

Dodge M and R. Kitchin. 2005. Code and the Transduction of Space. *Annals of the Association of American Geographers* 95 162–80

Elwood, S. 2006 Critical Issues in Participatory GIS: Deconstructions, Reconstructions, and New Research Directions. *Transactions in GIS* 10(5) 693-708

Elwood S. 2010. Geographic information science: emerging research on the societal implications of the geospatial web. *Progress in Human Geography* 34(3): 349–357.

Ford, H. 2011. The Missing Wikipedians. In *Critical Point of View: A Wikipedia Reader*, ed. G. Lovink and N. Tkacz, 258-268. Amsterdam: Institute of Network Cultures.

Graham, M. 2011. Cloud Collaboration: Peer-Production and the Engineering of the Internet. In *Engineering Earth*. ed. Brunn, S. New York: Springer, 67-83.

Graham, M. 2013. The Knowledge Based Economy and Digital Divisions of Labour. In *Companion to Development Studies, 3rd Ed.* Eds. Desai, V. and Potter, R. (in press).

Graham M., S. A. Hale, and M. Stephens. 2011. *Geographies of the World's Knowledge*. London: Convoco! Edition.

Graham M., B. Hogan, and A. Medhat. 2012. Dominant Wikipedia Language by Country. http://www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html (last accessed 17 April 2013).

Graham M., and M. Zook. 2013. Augmented Realities and Uneven Geographies: Exploring the Geolinguistic Contours of the Web. *Environment and Planning A* 45(1): 77–99.

Graham M, M. Zook, and A. Boulton. 2013. Augmented Reality in the Urban Environment: Contested Content and the Duplicity of Code. *Transactions of the Institute of British Geographers.* 38(3) 464-479.

Gramsci, A. 1971. *Prison notebooks*. New York: International Publishers. (Gramsci, A. (1971). Selections form the prison notebooks. *Edited and translated by Q. Hoare & GN Smith.) New York: International Publishers*.)

Haklay M. 2013a. Neogeography and the delusion of democratisation. *Environment and Planning A*, 45, 55–69Halavais A. and D. Lackaff .2008. An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication* 13(2): 429–440.

Haklay, M., 2013b, Citizen Science and Volunteered Geographic Information – overview and typology of participation in Sui, D.Z., Elwood, S. and M.F. Goodchild (eds.), 2013. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice . Berlin: Springer. pp 105-122

Hall S (Ed). 1997. *Representation: Cultural Representations and Signifying Practices*. London: Sage.

Hardy D. 2013. The Geographic Nature of Wikipedia Authorship. In *Crowdsourcing Geographic Knowledge*, ed. D. Sui, S. Elwood, and M. Goodchild 175–200. Dordrecht: Springer.

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

Hargittai, E. and G. Walejko. 2008. The Participation Divide: Content Creation and Sharing in the Digital Age. *Information, Communication and Society* 11(2): 239–256.

Hecht B., and D. Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the 4th International Conference on Communities and Technologies, Penn State University, 2009*, 11–20. New York: ACM.

Jenkins, H. 2006. *Convergence Culture: Where Old and New Media Collide*, New York University Press.

Kitchin R and M Dodge. 2011. *Code/Space: Software and Everyday Life*. Boston: MIT Press.

Laclau, E., and Mouffe, C. 1985 *Hegemony and Socialist Strategy*. London: Verso.

Lam, S., K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. 2011. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration.* 1-10.

Lessig, L. 2003. An Information Society: Free or Feudal (talk given at the *World Summit on the Information Society, Geneva, 2003*). http://www.itu.int/wsis/docs/pc2/visionaries/lessig.pdf (last accessed 17 April 2013).

Leuenberger, C. and I. Schnell. 2010 The politics of maps: Constructing national territories in Israel. *Social Studies of Science* 40: 803–842.

Luyt, B. 2011. The nature of historical representation on Wikipedia: Dominant or alterative historiography? *Journal of the American Society for Information Science and Technology*, 62(6), 1058–1065.

Mowlana, H. 1997: Global information and world communication. Sage Publications.

Osborn, D. 2010. *African Languages in a Digital Age*. Cape Town: HSRC Press.

Pasley R., P. Clough, R. S. Purves, and F. A. Twaroch. 2008. Mapping geographic coverage of the web. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, 2008*, Article 19. New York: ACM.

Pickles J (ed). 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. New York: Guilford

Rundstrom R. 1995. GIS, indigenous peoples, and epistemological diversity. *Cartography and Geographic Information Systems* 22: 45–57

Sardar, Z. 1996. alt.civilizations.faq: Cyberspace as the Darker Side of the West. In Ziauddin Sardar, and Jerome Ravetz (eds.) *Cyberfutures: Culture and Politics on the Information Superhighway.* New York: New York University Press.

Sawicki D and W. Craig. 1996. The democratization of data: Bridging the gap for community groups. *Journal of the American Planning Association* 62: 512–23.

Shirky, C. 2011. *Cognitive Surplus: Creativity and Generativity in a Connected Age*. London: Penguin.

Sieber R. E., and H. Rahemtulla. 2010. Model of public participation on the Geoweb. In *Proceedings of GIScience*, Zurich, 2010.

Silverwood-Cope S. 2012. Wikipedia: Page One of Google UK for 99% of Searches. http://www.intelligentpositioning.com/blog/2012/02/wikipedia-page-one-of-google-uk-for-99-of-searches (last accessed 2 May 2013).

Slashdot. 2004. Wikipedia Founder Jimmy Wales Responds. http://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds (last accessed 17 April 2013).

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737-1748.

Tapscott, D. and A. D. Williams. 2006. *Wikinomics: How Mass Collaboration Changes Everything*. New York: Penguin USA.

Thompson P., and M. Fox-Kean. 2005. Patent citations and the geography of knowledge spillovers: A reassessment. *The American Economic Review* 95(1): 450–460.

Thrift, N., & French, S. (2002). The automatic production of space.*Transactions of the Institute of British Geographers*, *27*(3), 309-335.

Weiner D, Warner T, Harris T, and R. Levin. 1995. Apartheid representations in a digital landscape: GIS, remote sensing, and local knowledge in Kiepersol, South Africa. *Cartography and Geographic Information Systems* 22: 30–44

WikiLocation. 2013. The Definitive Geolocation API for Wikipedia. http://wikilocation.org (last accessed 17 April 2013).

Wikimedia. 2013. Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm (last accessed 2 May 2013).

WikiProjekt Georeferenzierung. 2013. Wikiprojekt Georeferenzierung Hauptseite. http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en (last accessed 17 April 2013).

Wilson, M. W. 2011. 'Training the Eye': Formation of the Geocoding Subject. *Social & Cultural Geography* 12 (4): 357–376.

Zhang Q., N. Perra, B. Gonçalves, F. Ciulia, and A. Vespignani. 2013. Characterizing scientific production and consumption in Physics. *Scientific Reports* 3: 1640.

Correspondence: Oxford Internet Institute, University of Oxford, e-mail: mark.graham@oii.ox.ac.uk (Graham); Oxford Internet Institute, University of Oxford, e-mail: bernie.hogan@oii.ox.ac.uk (Hogan); Oxford Internet Institute, University of Oxford, e-mail: ralph.straumann@oii.ox.ac.uk (Straumann); Oxford Internet Institute, University of Oxford, e-mail: amedhatm@gmail.com (Medhat).

Table 1. Statistics of variables in the regression analyses detailing choice to log transform

Variable	Mean	Std. Dev.	Median	Skewness	Subsequently logged
Number of Articles	23780.0	71,874	3,454.0	6.03	Yes
Population (Millions)	43.7	146	10.2	7.47	Yes
GDP p.c.	14809.0	25,141	4,743.0	3.83	Yes
Gross Enrolment Ratio [GER]	76.8	27	86.1	-0.54	No
Broadband connections	3289169.0	12831210	184864.0	7.38	Yes
Quarterly Wikipedia edits (Thousands)	120.4	370	8.4	6.59	Yes

Table 2. Bivariate Correlations (Pearson product-moment) among variables. Values of 0.7 and higher are highlighted in bold.

	Articles	Population	GDP p.c.	GER	Broadband	Edits
Articles*		0.55	0.48	0.53	**0.81**	**0.7**
Population*	0.55		-0.16	-0.04	0.46	0.36
GDP p.c.*	0.48	-0.16		**0.76**	0.64	0.58
GER	0.53	-0.04	**0.76**		0.68	0.64
Broadband*	**0.81**	0.46	0.64	0.68		**0.8**
Edits*	**0.7**	0.36	0.58	0.64	**0.8**	

Table 3. Forward selection Ordinary Least Squares linear regression models predicting log(number of Wikipedia articles) per country

	Model 1			Model 2			Model 3		
	Est.	Std. err.	VIF	Est.	Std. err.	VIF	Est.	Std. err.	VIF
(Intercept)	-0.887†	0.524	–	0.295	0.306	–	2.905***	0.493	–
Population (log)	0.355***	0.068	2.53	0.223***	0.049	1.27	0.112*	0.047	1.48
GER	0.003	0.002	2.95						
GDP p.c. (log)	0.158†	0.085	3.54						
Broadband (log)	0.251***	0.049	4.80	0.303***	0.04	3.06	-0.540***	0.137	45.92
(Broadband (log))^2							0.087***	0.014	45.90
Edits (log)				0.059*	0.027	2.76	0.067**	0.024	2.76
Adjusted R^2	0.704***			0.703***			0.755***		
N	158			158			158		