Uneven openness: Barriers to MENA representation on Wikipedia

Mark Graham, Principal Investigator

Bernie Hogan, Principal Investigator

University of Oxford

Final Technical Report

Submitted to IDRC, December 31, 2013

IDRC Project Number: 106228

IDRC Project Title: Who represents the Arab world online? Using the case of Wikipedia to map and measure local knowledge production and representation in the Middle East and North Africa.

Country/ Region : Middle East and North Africa

Full name of research institution:

Oxford Internet Institute, University of Oxford

Address: 1 St. Giles, Oxford, UK

Research Team

Principal Investigators

Mark Graham, Senior Research Fellow, University of Oxford

Bernie Hogan, Research Fellow, University of Oxford

Senior Researchers

Ilhem Illagui, Associate Professor, American University of Sharjah

Ali Frihada, Faculty Member, Ecole Nationale d'Ingénieurs de Tunis

Junior Researchers

Claudio Calvino, University of Oxford

Richard Farmbrough, University of Oxford

Heather Ford, University of Oxford

Frederike Kaltheuner, University of Oxford

Ahmed Medhat, University of Oxford

David Palfrey, University of Oxford

Research Associates

Gavin Baily, Tracemedia

Kalina Bontcheva, University of Sheffield

Taha Yasseri, University of Oxford

Project Manager

Clarence Singleton, Project Manager, University of Oxford

Abstract

In this report we detail the activities and research insights of the funded project "Who represents the Arab world online? Using the case of Wikipedia to map and measure local knowledge production and representation in the Middle East and North Africa." We highlight the persistent information asymmetries between the MENA region and much of the developed world. We explore this through the use of data from Wikipedia and the Wikimedia foundation that has been extensively processed using a variety of tools from computer science. We augment this analysis with a series of qualitative insights discerned from two workshops with active Wikipedians in the MENA region. These workshops also served as forms of capacity building.

In general, we believe that editing Wikipedia is an intensive task that is dominated by the global North and replete with a great deal of barriers to entry. It is nevertheless a key activity for future development of a stable information ecology within the MENA region. This is especially valid given how extensively Wikipedia is used across the web in sites such as Facebook and Google.

We believe that capacity building among key Wikipedians can create greater understanding and offset much of the emotional labour required to sustain activity on the site in the face of intense arguments and ideological biases. However, we also believe that a distinct lack of sources both owning to a lack of legitimacy for MENA journalism and a paucity of open access government documents inhibit further growth in this area. Future work should be dedicated to these issues of support for active existing Wikipedians and knowledge sharing of Canada's best practices in information sharing to facilitate accelerated growth of geographic content in the MENA region. But the ultimate difference will come from increased diffusion of broadband internet technology as demonstrated through several statistical models. We articulate local exceptions to this rule, but remain committed to an overall strategy of capacity building and broadband diffusion.

Table of Contents

1. Introduction

1.1 The Research Problem

"The basic rationale for the project and the research problem or problems that were addressed should be stated. Often, the researchers' understanding of the problems will have evolved since the project was approved. The report should describe this evolution and the reasons behind it."

There are obvious gaps in access to the Internet, particularly the participation gap between those who have their say, and those whose voices are pushed to the side lines. Despite the rapid increase in Internet access, there are indications that people in the Middle East and North Africa (MENA) region remain largely absent from websites and services that represent the region to the larger world.

We explore this phenomenon in detail through one of the MENA region's most visible and most accessed sources of content: Wikipedia. This website currently contains over 9 million articles in 272 languages, far surpassing any other publicly available information repository. It is widely considered the first point of contact for most general topics, thus making it an effective site for framing any subsequent representations. Content from Wikipedia also has begun to form a central part of services offered elsewhere on the Internet.

Wikipedia is therefore an important platform from which we can learn whether the Internet facilitates increased open participation across cultures, or reinforces existing global hierarchies and entrenched power dynamics. Because the underlying political, geographic and social structures of Wikipedia are hidden from users, and because there have not been any large scale studies of the geography of these structures and their relationship to online participation, groups of people may be marginalized without their knowledge.

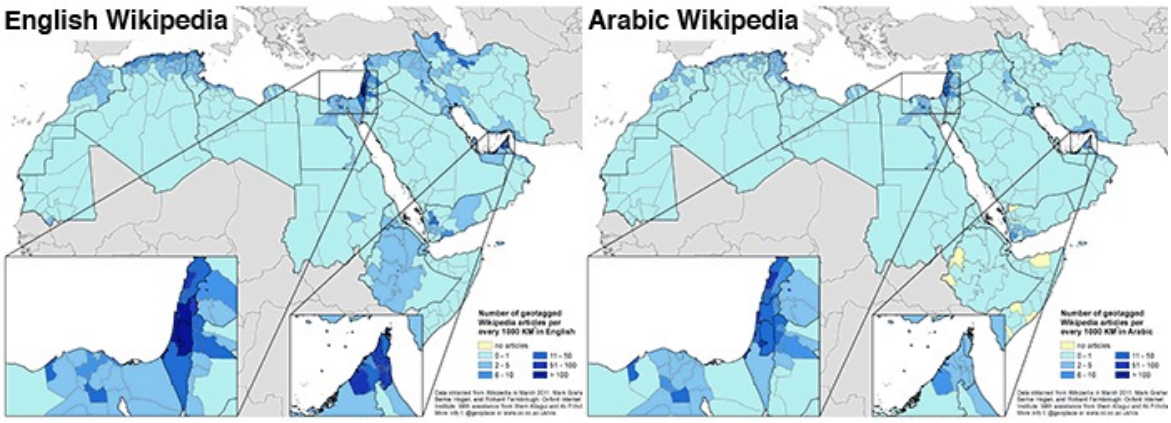

Figure 1.1a. A map detailing differences in content in Wikipedia within the MENA region between English and Arabic.

This relative lack of MENA voice and representation means that the tone and content of this globally useful resource that represents MENA, in many cases, is being determined by outsiders with a potential misunderstanding of the significance of local events, sites of interest and historical figures. Furthermore, in an area that has seen substantial social conflict and political upheaval, participation from local actors enables people to ensure balance in content about contentious

issues. Unfortunately, most research on MENA's Internet presence has been drawn from anecdotal evidence, and no comprehensive studies currently exist.

This project will therefore employ a range of (primarily quantitative) methods to assess the connection between access and representation, using MENA as the first step in an assessment of the inequalities in the global system.[1]

1.2 Research Objectives

"The general and specific objectives of the project specified in the MGC should be restated, with a discussion of whether or not the objectives were met. If the objectives were not met, outline the reasons why and the subsequent impact on the project. Objectives may have also evolved, and the reasons and learning involved should be described. The degree of fulfilment of any new objectives should also be assessed."

In this project, we have two forms of objectives, research objectives and practical objectives. The research objectives focus primarily on a description and analysis of where is represented online and why, with a particular emphasis on describing where in the MENA region is represented on Wikipedia and providing reasons why at a varieties of scale. We justify our choice of Wikipedia in Section 2.1 by highlighting its central role in the organization of online knowledge for most Western nations and indeed most nations in the Global South excluding China.

Within this broad objective are more focused objectives on Wikipedia's administrative structure and the process of creating content. Our practical objectives involve a consolidation of Wikipedians in the Middle East with a focus on how to create more equitable and representative content with the ultimate goal of making Wikipedia a more generative and productive site for reference information about the MENA region (especially in English, French and Arabic).

In detail, our research objectives are as follows:

Descriptive:

- To describe the visibility of the MENA region on Wikipedia
- To describe the visibility and impact of MENA residents and those who directly represent the MENA region to the rest of the world.

Explanatory:

- To understand where place-based content comes from.
- To explain reasons for the relative lack of Wikipedia articles in Arabic and about the MENA region.
- To understand which parts of the MENA region are underrepresented.
- Whether editors from the MENA region have less of a voice than their counterparts from elsewhere.

[1] http://www.oii.ox.ac.uk/research/projects/?id=70

- Whether content from editors from the MENA region is considered more legitimate or less legitimate (as measured through the number of reverts).

To understand the relationship between Wikipedia's administrative structure and the treatment of new editors (especially from the MENA region). In our original document, we also asserted an interest in uncovering potential systematic structures of exclusion that could be a barrier to participation such as:

- Competitive practices such as content deletion
- Indifference to content produced by MENA authors
- Marginalization through bullying and dismissal.

We have sought to answer all of these questions through a variety of research methods. We focus primarily on passive data and trace data, provided by Wikipedia itself. That said, we augment this analysis with a multitude of qualitative data attained from Wikipedia editors in the MENA region. This qualitative data comes from a variety of sources - an online Wiki we hosted, a survey we sent to MENA Wikipedians, a Facebook group that we have created to solicit feedback and most importantly, two workshops in the middle east, one in Cairo in late 2012 and one in Amman in early 2013.

1.3 Practical Objectives

The practical objectives for our project was initially to facilitate the production of Wikipedia articles from local actors and enhance the quality of MENA-specific content.

To further this objective we have sought to work in both direct and indirect ways.

Directly:

- Working with active and interested local content producers through planned comity workshops.
- Making available the resources from these workshops through a project page for future interested content producers. Such resources should emphasize how to justify the inclusion of online content, and how to navigate the administrative and political structure of Wikipedia. We also wanted to highlight how some content may have emerged from institutional actors seeking to 'whitewash' or artificial clean up Wikipedia pages.

Indirectly:

- Creating a database driven website that compares MENA articles to articles around the world and presents actionable information on how to develop these articles more fully.
- Raising the profile of Wikipedia and publicly highlighting such inequalities as a way to motivate content producers themselves to help rectify the stark inequalities in global online knowledge.

1.4 A Reassessment of Objectives

We have met the bulk of the objectives stated in this document through a series of workshops, research reports and in-review academic articles. We have not met all of the objectives, however. We do believe this is not due to a lack of effort. Instead it is because a proper treatment of the

tasks themselves would require substantially more effort, skills and technologies than was practically available. Further, we have uncovered some significant research challenges that will persist in future work. Where these exist, we believe our work will provide steps towards solving these significant challenges in future study.

The greatest challenge has been in concerning Wikipedia's means for managing identity. Unlike Facebook or Twitter, where users have a specific "location" field, Wikipedians can write about themselves in any manner they wish. As a consequence of this use of free-text it has been remarkably difficult to identify the location of registered users. After two years of effort, we have only been able to identify the location of approximately one in ten registered editors on English Wikipedia. Furthermore, we do not believe that future efforts will increase this number substantially simply by analysing the text of the user profiles. If the user does not say where they are from (since doing so is completely optional) there are practical limits to a researcher's ability to infer where the user is coming from. What this means for our research is that without knowing a user's location, we cannot explore whether the user's is speaking on behalf of a specific part of the world.

A second limitation concerns place names in Arabic. In order to assess place names in structured text, one needs to make use of a 'gazeteer'. A gazeteer is a long list of place names arranged in such a way that one can know that cities are nested in provinces and countries along with the associated geographic coordinates. One can then split up a text and look for the words in the text inside the gazetter. Thus, one can split up "I live in London" into four words and look up London. In this case, London is ambiguous (since there is a London, Ontario and a London, England), but we can use supplementary information to clarify which London is the right one. The presence of this supplementary information, and indeed, even the presence of a stable Gazetter in Arabic is not available.

It was our goal to have Prof. Frihida assist us in producing a gazeteer for Arabic, at least for MENA countries. However, Prof. Frihida noted several reasons why the Arabic language resists the use of such a gazeteer. For example, the word for Cairo (like many other places) is actually a different word depending on the context. Such complexities mean that a complete Arabic gazeteer would be a massive undertaking in its own right, and outside the scope of this project.

Thus, our answers to questions about the treatment of MENA individuals on Wikipedia is limited to interviews with workshop participants as well as inferences from English. We strongly believe that the generation of a gazetteer in Arabic with a direct English translation would be a germane project for future development, as gazeteers have been crucial not only in identifying where people come from but producing content in the first place. However, we also believe that such a gazetteer ought to be created in conjunction with national authorities from MENA states in order to ensure accuracy. This task comes with its own complexities in the case of contested borders, especially within the West Bank, which is one of the reasons why the generation of such a list is outside the scope of modest research project.

These two issues notwithstanding, we have sought to accurately and fairly address the existing research objectives. The remainder of this document provides context about online information inequality, our specific research, our project outputs and outcomes and recommendations based on our objectives.

2 Background

2.1 Research Context

To fully understand the scope of this project, below we outline a short history of Wikipedia, detail how articles are created, give an overview of the social history of the inception of Arabic Wikipedia and highlight how Wikipedia is a 'generative' encyclopedia that powers much of the web.

2.1.1 A History of Wikipedia

In 1995, Ward Cunningham created a system that enabled users of a webpage to quickly edit content on that page without going through the complex process of directly editing HTML and uploading it to a server. He labelled this system as a wiki, coming from the Hawaiian word for quick. The first Wiki was released that year as WikiWikiWeb. Over the next 6 years, wikis slowly grew in popularity within academic and technical circles. Sites such as Meatball Wiki and Twiki were used to document source code and categorize online arcana.

In 2001, Jimmy Wales and Larry Sanger used a recently developed wiki package called "MediaWiki" to create an encyclopedia. Wales was not a 'hacker' per se, but a considered himself a philosopher and entrepreneur. Wikipedia was meant to be a not-for-profit website for high quality reference information, while Wales simultaneously set up a for-profit site, Wikia, for advertising-supported topic specific content. Wikipedia currently has extensive articles on statistics and biology, whereas Wikia has thousands of pages on LOST (the television program), The Legend of Zelda and so forth.

Wikipedia has grown steadily in size and scope since 2001. Only five years after its inception, the journal Nature published a remarkable finding that Wikipedia articles were comparable to Encyclopedia Britannica articles in terms of size and quality (Giles 2005). Around this time, Wikipedia began to gain serious legitimacy as a cultural force.

The original Wikipedia was in English, but has spun out into several hundred languages. Stats on these languages can be seen at http://stats.wikipedia.org/ . Wikipedia includes all of the major spoken languages around the world, as well as a large number of local (e.g., Scot and Faroese) and even synthetic languages (e.g., Esperanto and Volapuk). It used to contain Klingon as well (the language was even featured in the Wikipedia logo until 2010), but this has since been moved to its own domain on Wikia. English is undoubtedly the most popular and dominant Wikipedia with 4.4 million articles and almost 14 million views per hour, although many other (typically European) languages have very large Wikipedias. Notably, the Dutch Wikipedia has the second largest number of articles. Other versions are also highly active, such as the Japanese Wikipedia with 1.7 million views per hour and Spanish with 2.4 million views per hour. Arabic has approximately 240k articles and receives (coincidentally) 240k views per hour. Per the number of speakers, Arabic is one of the least represented major world languages on Wikipedia. While Hindi and Malay are also widely underrepresented on Wikipedia, we can still say that currently there is no language with more speakers and fewer articles than Arabic.

Despite its extensive use and popularity, Wikipedia has always been plagued by a reputation problem: If anyone can edit the encyclopedia then why should we trust the information on the site? This idea that Wikipedia is an anarchic site just as easily vandalized by 14 year olds as edited by 40 year olds persists in pop culture. For example, satirist and talk show host, Stephen Colbert, started a campaign in 2006 to alter the content of the article on elephants to suggest that their

population has tripled in the last six months (a physical impossibility and a tragically unlikely one). The reaction of the Wikipedia community was swift and organized while the page on elephants was protected from vandalism. Nevertheless, the notion that anyone could edit the document persisted, with comedians often suggesting someone 'edit that Wikipedia article' to match a gaffe or falsehood.

The tension between Wikipedia's ability to allow anyone with an internet connection to edit and the concern over the quality of content eventually led to the departure of co-founder Larry Sanger. Sanger felt that Wikipedia would forever be plagued by vandalism and the stigma of open access. In 2007 he launched Citizendium as a fork of Wikipedia that required real names and editorial oversight. This project was widely seen as a failure with less than one hundred active editors by 2011. Google Knol was a similar effort to compete with Wikipedia by using trusted editors (who verified using a credit card). This project was also seen as a failure and discontinued as of May, 2012.

Although Wikipedia has survived the demise of many of its closest competitors, this is not to suggest that its road has been an easy one. Presently, Wikipedia still runs as a charitable organization and depends on user donations for regular server maintenance. Through such funding, it has recently introduced a WYSIWYG ("What you see is what you get") editor to help newcomers overcome the tedium of wiki syntax and received near unanimous criticism from its users. Since 2007 and the introduction of significant devolution of administrative powers to volunteers, the site has not been able to retain newcomers. Its editorship is still on a slow and steady decline, even as its content and readership steadily increases year on year (especially in the Global South).

Some say that perhaps Wikipedia is levelling off because there is only so much to write about. What we show in this report is that such a claim is extremely far from the truth. There are still substantial gaps in geographic content in English and overwhelming gaps in other languages. Wikipedia often brands itself as aspiring to contain "the sum of human knowledge", but behind this mantra are policy pitfalls, tedious editor debates and delicate sourcing issues that serve to hamper greater representation. Such challenges are part of Wikipedia's evolution as the *de facto* source for online reference information, but they also serve to entrench particular ways of knowing and ways of validating what is known.

2.1.2 How content is created and updated

To understand much of the subsequent analysis on Wikipedia, it is useful to have a brief primer on Wikipedia's governance and behaviour.

Wikipedia is a form of a wiki. This means that a page's content can be edited by a user and then saved. Wikis tend to save page-specific histories and allow some people to roll back changes (in case a user comes in and vandalizes a page or deletes it). This is called reverting. Wikis are composed using a wiki syntax that uses special characters such as [[ ]] to denote links within an article and between articles. Characters such as ** content ** signify titles, and (content)[http://www.content.ca] signify ways to create URL links respectively. Wikis therefore use a "markup language". This means that what one edits in Wikipedia is not what one sees in the final document.

The criteria for inclusion in Wikipedia is contested, but tends to be described in one of Wikipedia's policy documents. These policies are denoted on the site by the prefix WP:. So the document describing how to handle a conflict of interest (e.g., users creating articles about themselves) would be having a WP:COI issue. When conflict occurs on Wikipedia, it is often first

introduced in a "Talk Page". Every article on Wikipedia has a supplementary talk page where people can discuss what should or should not be in an article and how to resolve such concerns. When inclusion is contested it is typically because there is not enough external source material about a topic to establish notability. As Ford (2011) notes, notability is often culturally mediated. For example, several years ago, a story in Al Jazeera would not be considered a sufficient criterion of notability. However, since Al Jazeera's central role in reporting on the Arab Spring this has changed dramatically.

Notability creates a sort of feedback loop. If an area of the world is underreported, there are no sources. If there are no sources, then journalists do not always have enough information to report about that part of the world. The use of sourcing trumps personal experience on Wikipedia. So that even if an author is from a place, and is watching a building being destroyed, their edit on Wikipedia will not be accepted by the community unless that event is discussed in another 'official' medium. Often the edit will either be branded with a 'citation needed' tag, eliminated or discussed in the talk page. Particularly aggressive editors and administrators will nominate a page for 'speedy deletion', a practice that makes responses from an author difficult.

Different editors on Wikipedia have different levels of privilege. The most basic level of privilege is the anonymous editor. Such an editor is identified only by an IP address and has no ability to vote on governance or edit certain protected articles (such as the articles on "Prophet Mohammed" or "Barack Obama", both of which are routinely subject to vandalism). A registered editor has log-in credentials. Such an editor gets a history for their comments, has their IP address hidden (even from Wikipedia, who delete this information after 90 days), and gets a user page where they can self-describe. We make extensive use of this user page both for contacting editors for our workshops and for analysis of representation. Administrators have greater powers than editors and can ban individuals as well as delete certain pages without discussion (called "speedy deletion").

In addition to text, Wikipedia articles can include "templates" and "rich content". Templates are pre-rendered kinds of content, such as the info box for a country. In that info box, an editor can include a country's GDP, population, flag and location. Info boxes are topic specific. The info box for actors includes a place to list movies, year of birth and a photo. Info boxes typically show up in the upper right corner of an article (in English, and upper left in right-to-left languages such as Arabic and Hebrew). Other templates, such as the annual rainfall, exist inside the main text.

Rich content refers to pictures, sounds or videos. Such content can be included in Wikipedia but must abide by certain standards. The content has to be uploaded to the Wikimedia Commons, and vetted by staff. This is an onerous procedure, and one that is a far cry from the simplicity of adding photos to Flickr or Facebook. Some Wikipedians are particularly adept at this task. Most photos of Tunisia in French and Arabic come from a handful of editors who have actively been uploading to the Wikimedia Commons for years. When at our Cairo workshop, one of these editors expressed his disappointment that other sites do not make it easy to pipe content from their site to the Commons. Particularly in the Arab World, there is a substantial amount of visual content on Facebook that would be useful for Wikipedia but cannot be imported due to a lack of proper sourcing. One of the Tunisian editors who regularly uploads content, Wael Gabara, has received funding from Wikimedia to travel Tunisia taking photographs of major landmarks in order to have appropriate photos for the respective Wikipedia articles. We consider this a process worth supporting.

In general, we talk about "Wikipedia", when in fact this is incorrect. Each language is considered its own Wikipedia. So there is a French Wikipedia, an Arabic Wikipedia and so forth. Thus we should refer instead to Wikipedias. Each Wikipedia can implement these rules somewhat

differently. The Arabic Wikipedia, for example, requires moderation for virtually all edits to the site from people other than administrators. An oral history of the creation of the Arabic Wikipedia was presented by one of its founders at our Amman workshop. An edited transcript will be published to the Wikipedia community's newsletter following the completion of this report.

2.1.3 A social history of the Arabic Wikipedia

The Arabic version of Wikipedia began in 2003, several years after Wikipedia's founding in English. It was initially started by a handful of academics and graduate students in Germany with Arabic backgrounds. Dr Rami Tarawneh, one of the earlier founders of the Wikipedia noted that he stumbled upon Wikipedia in English and thought that it would be great to have such an encyclopedia in Arabic. He was studying in Germany where he met several of the other early founders. Initially this contact happened through Wikipedia, although later the founders met in person.

One of the first articles in the Arabic Wikipedia was for the country of Syria on July 30, 2004. It was started by an anonymous IP address in Hanover, Germany. It was followed in a few months by Estonia and Carthage. By the beginning of 2004, Geotagged articles started appearing across the world, with a particular emphasis in the Gulf states.[2] As Dr Tarawneh describes it, the early days of the Arabic Wikipedia were anarchic. The editors were indifferent to copyright concerns and started simply copying content from elsewhere - encyclopedias, translations from the German Wikipedia or other places on the web. Early content was often started or included by IP address, and there was little governance structure. This freedom was coupled with a compulsion from early editors to transfer and translate the mass of regulations. He describes early years where he would get only 2-3 hours sleep a night, "go to the university from 8 to 4 or 5, then it is [editing for]10-14 hours, then sleeping. I don't know, there was something inside me telling me that you should do this". But as he sees it, this was not an exercise in completeness as much as an exercise in community building. As he states, "Wikipedia is not the articles, Wikipedia is the community." He noted that by "the end of 2005, we had enough people that made me at least sleep for five or six hours per day".

Dr Tarawneh also noted how in the early 2000s, Arabic visibility on the Internet was severely hampered. "Nobody in the Middle East, at that time, even tried to Google 'Wikipedia'…[it] wasn't the first article to pop up using Google or any search engine. So when I saw that kind of information that it can provide people, I was amazed. For example, my kid had a kind of asthma when we first went to Germany. I could find the whole thing on Wikipedia. That was in 2003, and it was amazing."

As Dr Tarawneh claims, Arabic Wikipedia exists partially as a steam valve for conflicts in other Wikipedias. For example, when authors want to include details about the Israeli-Palestine conflict in English and are chastised for being unbalanced, they vent this in articles in Arabic. This insight is reinforced by our later work on controversy patterns (see Section 6.5). On the other hand, the Arabic Wikipedia maintains a strong sense of Muslim identity and decorum. Articles on pornography, sex and sexual orientation remain biased against sexual expression and Western notions of gender roles. Dr Tarawneh, however, maintains that this is an expression of Arabic values rather than Islamic values explicitly, "It is the way people were raised here, not based on religion, but based on culture…initially when we started back in 2004-2005, four of us were Christians, and they had the same ideas."

[2] We discuss this emergence using a visualization in Section 5.3.

For several years, the state of the articles on the Arabic Wikipedia were in limbo. As we discuss in the next sub section, part of Wikpedia's power is in its ability to be repurposed. However, this is based on a legal license called "Creative Commons". The Arabic Wikipedia did not include a creative commons licence until after the start of this project, putting the ownership and licensing of content on the Arabic Wikipedia in a legal limbo and hampering repurposing efforts.

In all languages, Wikipedia's popularity is steadily growing. In 2008, there were approximately 9 billion page views a month across all Wikipedias. In 2013, this number has doubled to 18 billion. The Arabic Wikipedia has more than kept pace with this growth. It was the 20th most popular Wikipedia in November of 2008. It is currently, the 13th most popular in the world. Arabic's position relative to other Wikipedias increased significantly around October and November 2012, (presumably coincidentally) just after our workshop in Cairo. It moved from an average of 85 million page views per month in October 2011 to an average of 135 million page views a month for the same period a year later. However, to put this in perspective, the English Wikipedia has 18 billion page views a month.[3]

2.1.4 Wikipedia as generative platform

Beyond Wikipedia's success as a direct site for consuming content, there is a case for an interest in Wikipedia as a contemporary form of economic development as well as development that is in the public interest. This comes from Wikipedia's position as a generative platform. In "The Future of the Internet", Jonathan Zittrain lays out a case for many further directions of the Internet as civic space and economic engine (Zittrain 2008). In particular, he focuses on the extent to which certain online technologies can be considered 'generative'. A generative technology is one that enables third parties to engage with the technology along several axes: capacity for leverage, adaptability, ease of mastery and accessibility. Wikipedia is a generative technology *par excellence*. The following are but a few examples of how the presence of content on Wikipedia is multiplied across the web:

Facebook. Currently the world's dominant social network site, Facebook is not merely an online community but a social utility that links offline friends together. Its significance has been touted in the Arab Spring along with numerous prevailing cultural trends. It has a billion regular (at least monthly) users and over half a billion daily users. When a user looks for information about a place on Facebook, the description of that place as well as its latitude and longitude coordinates come from Wikipedia. If one wants to "check in" to a museum in Doha to signify they were there to their friends, the place they check in to was created with Wikipedia's data.

Google. The search engine giant, Google, has moved away from simply indexing pages based on their links to other pages. Instead, Google has made a turn towards semantic data. When one looks up "House of Saud" on Google, they are presented not only with a list of links (where Wikipedia is at the top), but a special 'card' that gives a summary of the House of Saud. The data for this comes from Wikipedia. When one is looking for people or places, Google now has these terms inside of its 'knowledge graph', a network of related concepts with data coming directly from Wikipedia. Similarly, on Google maps, Wikipedia descriptions for landmarks are part of the default information.

[3] See http://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm

Natural language. In 2011, IBM famously created an artificial intelligence called Watson and had it compete successfully on Jeopardy. While Watson erred several times (including guessing Toronto instead of Chicago for a final Jeopardy question), it still easily beat two of Jeopardy's most celebrated winners, the single day champion and the longest running contestant. Much of Watson's knowledge came from parsing Wikipedia data.

Tourist Guides. There is presently an emerging market for online tourist guides. Whereas past guides would involve specific writers who scour place for new and novel information (e.g., *Lonely Planet* or *Frommers* guides), current application guides, such as the free ones provided by Triposo, overlay data from Wikipedia on top of maps curated by OpenStreetMap. Triposo currently has free guides for Morocco, Egypt and the UAE, all of which have extensive landmarks with descriptions drawn from Wikipedia.

This is to say that Wikipedia is not merely a site of reference information, but rapidly becoming the de facto site for representing the world to itself. If there are biases, absences or contestations on Wikipedia, these issues spill over into numerous domains that are in regular and everyday use. If a place is not on Wikipedia, this might have a chilling effect on business and stifle journalism. If a place is represented poorly on Wikipedia this can lead to misunderstandings about the place. Wikipedia is not a legislative body (at least not except for legislating on its own matters). However, in the court of public opinion, Wikipedia is one of the world's strongest forces as it quietly inserts itself into representations of place throughout the world.

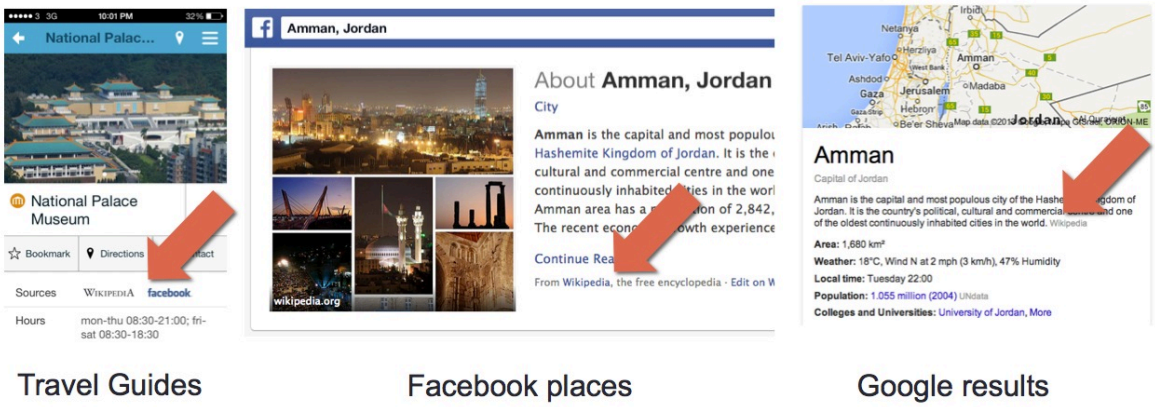

Figure 2.1.4a. Examples of Wikipedia as a generative platform. From left to right: Triposo's Taiwan travel guide, Facebook's places and Google's knowledge graph search results.

3 Methodology

This project has used a variety of methodological approaches for a variety of purposes. This is based on our multiple objectives of creating research knowledge, facilitating greater cohesion and interaction among Wikipedians, and to ensuring our work reaches out to the public in ways that bring to the fore the issue of uneven representation online.

Below we discuss the methods in several subsections:

- Article geolocation methods
- User geolocation methods
- Article and user networks
- Online editor surveys and text input
- Workshop participation
- External data

3.1 Article geolocation methods

Our first research question and the basis of future work is essentially the descriptive task of identifying where is being written about on Wikipedia. To perform this task we focused on the use of geocodes (I.e. latitude and longitude coordinates) on an article page. This was a practical choice. Latitude and longitude coordinates tend to be "well-formed" data, expressed either as decimal numbers or within hours, minutes, seconds notation (e.g., 1.4, 38.2; 16"13'53). This means that we could automatically read the contents of thousands of Wikipedia pages per minute in order to determine whether the article had a geocode.

We had a geocode parser up and running within a week of starting the project. So why would it take over a year to finalize this work? Employing geocodes has several limitations:

- The HH:MM:SS or decimal notation is not sufficient on its own. Some things might be misunderstood as a geocode when it is in fact another representation such as declination for stars and comets.
- The earth is not the only astral body with latitude and longitude coordinates. Thus we had to remove references to craters on the Moon and mountains on Mars.
- We needed a grounded approach to determine a point location for an article when the article has multiple geocodes. We opted to choose either the most frequently used code if there are multiples or the code in a privileged location on the page, such as the code in the upper right hand (or upper left in right to left languages).
- Latitude and longitude coordinates are sometimes suffixed by North, South, East and West. In other languages, it is possible to use the language specific words such as nord, sud, est and ouest. These had to be accounted for.

To assign a geocode to a country typically involves the use of a 'spatial join', or an algorithmic process for determining where a point exists relative to country borders. If a point is on a border, or in the water just outside a country, we had to account for this. While parsing text is relatively

fast, performing spatial joins is a notoriously slow task (I.e. Taking a full day to assign a million such points to their respective countries).

We should not assume that merely because an article has a geocode, that the geocode is correct. While cross-checking all four million identified places across the 44 languages we employ is unrealistic, we were able to determine some codes as being incorrect through poor syntax, inconsistencies across language and instances where the latitude exactly equals the longitude.

Our first parser was written by Richard Farmbrough. In addition to Mr. Farmbrough's masters-level training in maths, he is also notable as one of the most active Wikipedians ever with over one million personal edits (a fact we did not know when we initially hired him). His commitment to Wikipedia helped us learn more about Wikipedia's cultures and norms, although his focus on editing Wikipedia also meant that we had to compete for his time with a substantial commitment to many editing tasks. Also, it is worth noting that Mr. Farmbrough was most comfortable in the scripting language Perl, whereas P.I. Hogan was most familiar with Python. This was assumed to not be an issue at the outset, but over time, it led to a lack of proper code review. While Mr. Farmbrough's Wikipedia experience was initially an asset, over time it was clear that Mr. Farmbrough's skill set had significant limitations for tasks related to the MENA region and the more complex machine learning tasks we required.

Our second parser was written by researcher Ahmed Medhat. Mr. Medhat is a native of Cairo and a fluent speaker of both English and Arabic, with training in Egypt and the UK, including a Master's Degree in computer science from the University of Oxford. Mr. Medhat wrote a parser in Python and cross-checked these results with the previous parser. Through a number of improvements we were able to increase our collection of articles in Arabic and Farsi.

During this process we became aware of two additional efforts to parse all geocodes on Wikipedia: WikiLocation[4] and WikiProjekt Georeferenzierung.[5] We were focused on seven languages relevant to the MENA region: English, Arabic, Egyptian Arabic, French, Hebrew, Farsi and Swahili. By contrast, these other efforts had a combined 44 languages coded. The 44 languages (featured in Appendix V) comprise all Wikipedias with more than 100,000 articles as well as select Wikipedias that represent national interests. These 44 languages also included two synthetic languages, Volapuk and Esperanto, that have unusually large Wikipedias. These have been excluded. Unfortunately, we still omitted several key languages that have experienced substantial growth in the last two years: Waray-Waray, and Cebuano in South Asia and Kazakh and Uzbek in Central Asia. These omissions were because neither language was included in WikiLocation or WikiProjekt Georeferenzierung and we did not have the local expertise in order to verify the quality of a parser for these languages.

There are other limitations of this process, but one in particular stands out as worth considering: the absence of a geocode does not imply that the article is not about a spatial entity. Particularly when referring to countries themselves, a single geocode point does not really capture the shape of the entire country. In fact, if one takes the centroid of a country's boundaries, as in the case of Italy, it might not even fall within the country's territorial borders. We sought to rectify this absence of geocodes following the technique of WikiProjekt Georeferenzierung: since all articles can exist in multiple languages and there are links between the languages, we can look through all other languages to see if a geocode exists there. If that fails, the process of determining an article's

[4] http://wikilocation.org/

[5] http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung

location is remarkably complex and would involve a great deal of human inference. This was out of scope for this project, and indeed considered out of scope for other geographic efforts on Wikipedia.

The geo-identified articles form the basis of a great deal of our work, as well as the majority of our most popularly received public output. Therefore, we believe it is prudent that we spent a significant amount of time ensuring high data quality for this particular effort.

The source code for our finalized geo-location parser is available at https://github.com/oxfordinternetinstitute/Wikiproject/tree/master/GeoParser

3.2 User geolocation methods

The Article geolocation methods helped us to understand where is being referenced on Wikipedia, but not the location of who is doing the writing. Without this extra information we only have half of the story about representation.

The identification of locations through geocodes was a technically complex task across a number of languages. Yet, the identification of locations of users is a far more complex by an order of magnitude. This is because of the nature of identity and identification on Wikipedia.

As mentioned in Section 2.1.2, Wikipedia editors can be either logged in or anonymous. If the editors are anonymous, then they are identified by an IP address. If the editors are logged in (or 'registered'), then their IP address is hidden from other users. This IP address is only linked to a user for a period of 90 days in order to assist with law enforcement in cases of unambiguously objectionable content (e.g., slander, child pornography). As part of Wikipedia's terms of use, this information is never used for analysis and thus is not available to us. While this is an unfortunate limitation, it is also an incentive for users to create a registered account as a form of privacy management.

As a consequence of this limitation, the only feasible (and to a great extent, ethical) way to assign locations to users is to make use of self-documented locations. Also mentioned earlier, every registered user on Wikipedia has a user page in each Wikipedia language that they edit. This page is in 'free-text', meaning that the user can write what they want, in any order, including the user of graphics, pictures and videos. Many editors use their page as a place to denote their affiliations on Wikipedia, their achievements (typically awards called "Barnstars"). Also on the user page are a lists of the groups that the user is associated with (such as "Wikipedians in Qatar") and "userboxes" (designed boxes that denote affiliations and interests). A user's page can also have sub-pages for specific things such as a collection of photos or places of interest.

For the most part, objects of interest on Wikipedia can be described by a single point or shape. Carthage is in Tunisia, the Acropolis is in Athens, Greece. People, on the other hand, move around. They may have parents from different countries, and spend time abroad. Indeed, as discussed in Section 2.1.3, Dr Tarawneh, a founder of the Arabic Wikipedia, joined after moving from Jordan to Germany. This complicates our ability to map where people are writing from. Does an editor still have to live in Egypt to write about Cairo? If an editor vacationed in Sharm el-Sheikh would that be enough to consider them as being able to represent Egypt? (In this case we would say no.)

In the process of performing text mining on Wikipedia user pages, we had to come up with a practical schema for associating users with locations, and a way to verify the accuracy of this schema. To do this, we went through several stages of analysis.

The first stage, starting with Mr. Farmbrough, Prof. Frihida and Prof. Bontcheva, was to employ GATE, the 'Generalized Architecture for Text Engineering'. Unfortunately, there were a number of logistical issues that inhibited clear communication between all three parties, as well as significant challenging in putting together a workable schema with GATE. Nevertheless, we learned much from the way GATE processed text and instead created our own bespoke location identification process. This was written by Mr. Medhat, under the supervision of P.I. Hogan. The verification of the analysis was done by researcher Dr. Palfrey using CrowdFlower. The source code for the language identification system is available at https://github.com/oxfordinternetinstitute/Wikiproject/tree/master/UserLocation%20Parser . Below

we describe the process involved first in the free-text gazeteer and then how we sought to validate this with human cross-checking.

3.2.1 Free-Text Gazeteer

Our method consists of three steps: first, we produce a multilevel gazetteer that enables us to disambiguate similar place names (such as London, UK and London, Canada) based on existing work; second, we employ the gazetteer to identify locations present in editor profile pages; third, we discern locations that are considered salient to one's identity (*"I was born in Canada"*) from those that are incidental to one's identity (*"I visited Japan last year"*).

We begin with an English place name gazetteer encompassing about 2.7 million toponyms from the GeoNames database. GeoNames produces the world's largest gazetteer that is based on freely available national gazetteers and datasets (GeoNames 2013), as well as volunteered geographic information (VGI; Goodchild 2007). For all entries the gazetteer contains FIPS codes (a standard developed by the United States federal government for tagging locations) linking place names to various national and subnational entities. *New York City*, for instance, would be linked to *New York* (state) and the *USA*.

We extended this gazetteer with articles in Wikipedia that contain geographic coordinates. We employ an expanded set of all 4.4 million georeferenced articles across the 44 languages in our database. This is because even if an individual is writing in English, they may use a non-English word for their place of work or residence. We did not include all of the articles as places, however, since events and even people can have geocodes but still should not be considered places for a gazeteer. For example, the Persian article on Muhammad contains the geotag for his place of birth. To overcome this we sought additional textual features such as "I work in X" or "I live in X" where X is a place. We used Naive Bayes rather than standard NLP because these features only work some of the time. We wanted to avoid false positives, such as "I work in Python, but you work in C++".

We further use rich content from Wikipedia places to help disambiguate place names that occur in multiple locations. For example, in the GeoNames gazetteer, "London" can refer to places in England, Canada, Australia, the USA and Myanmar. However, if the relation London:Myanmar is not substantiated in the Wikipedia data then we discard it from the gazetteer. Locations are discarded if they neither occurs as a place in geocoded Wikipedia articles or in editor profile pages (e.g. the statement "I live in London, Myanmar"). With this procedure, we were able to improve the resolution of 199,285 place names, by minimizing the number of potential countries that they could be referring to.

Extracting location signifying patterns.

When referencing locations in text, users tend to follow consistent grammatical rules. For example, a place will be preceded by a preposition ("I went *to* Cairo", "I was born *in* Jakarta"). Instead of simply building a list of such phrases based on our best guesses, we leveraged the original gazetteer. That is, we looked for the occurrence of gazetteer words in the text of user descriptions and then examined the words surrounding this text that followed certain patterns. To perform this analysis we employed NLTK (the natural language toolkit) in Python.[6] From this initial run, we

[6] http://nltk.org

were able to identify the most common patterns that emerge surrounding a location. We manually inspected these patterns and retained the ones we considered were most salient to our research question (i.e. determining individuals who *represent* particular locations). In particular, we selected terms that signify "born in/nationality" and "lives/works" in a particular location. For "lives", we had a series of terms such as "reside in" and "based in".

Appendix III lists all of the specific terms we included in our search.

Parsing and geocoding

With an expanded multilingual gazetteer and a list of appropriate n-grams that signify reference to a place we performed a text analysis on all 1,302,808 user profile pages that occurred in our Wikipedia editor data dump.[7] Our goal was to assign each editor to both national and where possible subnational locations according to prevailing FIPS codes. We had four possible cases for dealing with the occurrence of location signifying text and a place name in our gazetteer:

A place is unique and singular. Here we assign it to the appropriate level 1 [subnational] and level 0 [national] code. For example, *"I come from Manhattan"* would be assigned to New York state and USA.

A place is not necessarily unique, but is immediately followed by a larger region that disambiguates that place. For example, *"I come from Birmingham, Alabama"*. We again assign the place the appropriate codes drawn from our gazetteer.

A place is in a larger region that does not disambiguate that place because such a pairing does not exist in our gazetteer. For example, *"I worked in Boatswain Tickle Island, Newfoundland"*. Here we assign an individual a code of the larger region and store that pair even if Boatswait Tickle is not in the gazeteer.

A place name is ambiguous and is not followed by a larger region that disambiguates it. Here we make a best guess in a separate disambiguation algorithm discussed below.

Location disambiguation algorithm

We regularly encountered the problem of names that could refer to multiple places without an immediate means of resolution. The place name Columbus is mapped in the gazetteer to more than ten states in the US, including Ohio, Georgia, Illinois, Kentucky, New York and Nebraska. For example, *"I've lived in Columbus all my life."* Followed by *"member of the Ohio Wikipedians group"*. Since Ohio does not immediately follow Columbus, it is not considered an obvious match. However, there is still information in this text that would tie it to Ohio rather than Illinois since the author indicates such later on.

To resolve such ambiguities, we took a network analytic approach that links towns to regions and regions to countries. A directed edge from each of these places to the nodes representing their respective country locations. If a country is specified only a single edge is created from the specified place to the specified country, as would be the case in the following paragraph:

I was born in Cleveland, Ohio. Currently living and working in Columbus.

[7] We used the Wikipedia editor dump from October 7, 2011 available from http://dumps.wikimedia.org/backup-index.html

Here an edge will be created between Cleveland and Ohio, and an edge between Ohio and USA. As for Columbus, edges will be created from Columbus to the different possible regions or countries that have a Columbus. Building this network of all the geographic references to different regional and national locations, some regions thus become more central. We use PageRank (Page et al. 1999) because of its ability to account for weighted directed networks. We weighted our links between places in proportion to the number of possible matches. So mentioning Columbus creates ten links worth 0.1 each (from Columbus to Ohio, to Illinois, etc…), whereas mentioning Cleveland creates only one link to Ohio worth 1. By doing this we circumvent the need to have a user directly mention a country or region and still permit us to disambiguate the toponyms in the user pages.

Geocoding success of unregistered edits is very high with 99.7% of all anonymous edits being successfully allocated to a country. The number of registered editors that have been successfully geocoded using the methods outlined above is 122,888 (9.4%) out of a total of 1,302,808 . The total number of found locations for these users is 137,365, i.e. on average, a user has been affiliated with 1.118 locations.

Table 1 indicates the success rate of the geocoding approaches regarding edit types. Somewhat counter-intuitively, it is much easier to pin down the location of anonymous editors than registered ones: While anonymous edits could be geocoded almost in their entirety, only a third of the edits by registered editors could be assigned an origin location. Overall, for about half of all edits the editor location could be retrieved.

Metric	Total	Geocoded	Not geocoded
Anonymous edits	12,690,334	99.7% (12.6 mil.)	0.3% (0.04 mil.)
Registered edits	33,991,052	33.6% (11.4 mil.)	66.4% (22.6 mil.)
Total edits	46,681,386	51.6% (24.1 mil.)	48.4% (22.6 mil.)

Table 3.2.1a. Geocoding success of edits by anonymous and registered editors. Percentages are per edit type.

3.2.2 Crowdsourcing

We do not believe that automated coding on its own should be treated as necessarily correct. As we noted about our geographic parser, there were many instances of incorrect data, such as coding errors, geocodes on the moon and Mars, codes that were not actually geocodes, etc…In the case of geocodes we were able to use maps in order to inspect the data and detect inconsistencies. In the case of free text in user pages, however, we needed a different strategy in order to detect anomalies or errors. For this we turned to human coders on the crowdsourcing platform "Crowdflower". Many people are familiar with Crowdflower's competitor, Amazon Mechanical Turk. We opted for Crowdflower for two reasons. The first was that this platform allowed us to select where our coders came from. We believed it was imperative that these coders

were from English-speaking countries. The second reason is administrative: Mechanical Turk required payment from a U.S. Bank account or credit card. We did not have such an account available.

Crowdflower workers were tasked with identifying the origin and current location of users from their Wikipedia user pages. David Palfrey set up this task and uploaded the text from the user pages to Crowdflower for review. It took approximately six weeks to set up this task. Once available, the task was completed by the users in less than six hours. Several thousand workers were able to code every user who had edited a page on the MENA region (approximately 10,000 user pages). We then compared the results of Crowdflower and our automated program in its current version.

Using the identified pages as a benchmark, we were able to arrive at a recall rate of 90 percent for our location identification algorithm. We then personally examined every user who was associated with Egypt and with Germany (as a control). As we did this, we discovered that in some cases our automated algorithm was more accurate and the Crowdflower users had both incorrectly identified a place. We also discovered that our algorithm was overly liberal with identification, since we included mention of ancestry even if a person was neither born or lived in a country. Thus, we had many Americans incorrectly associated with Ireland, England and Germany because of the ancestry tags. We found that in almost all cases we examined, if a user actually lived or was born in a place this information would be available outside an ancestry tag.

Further tuning of our algorithm actually dropped our recall rate to 87 percent. Yet, we believe this latter rate might be the upper limit of legitimate possible recall. The loss of 3 percent precision is actually because of mistakes that both the parser and the human coders made. By elimitating these errors from the code we ironically gain accuracy while losing precision. In the end, our geolocation tagger is meant to be seen a practical approach to the geographic identification of users. We are interested in whether fellow Wikipedians will judge a user based on their location. As such, we believe that the information we have automatically coded is very similar to the sort of information other Wikipedians would use in evaluating each other. Nevertheless, this process demonstrates the importance of continual feedback between algorithm design, human coders and researcher goals rather than potentially misplaced trust in off-the-shelf general purpose NLP tools or human coders exclusively. We believe our iterative technique will be germane in future work. We plan on sending a paper reviewing this technique and our concerns to a peer reviewed journal on big data in January 2013.

3.2.3 Weighted and unweighted user locations

With the locations in hand, we had to make a choice about how to weight locations for users with multiple country associations. We had two options – each user is given a location score of 1 and that score is split between the different countries (either equally or according to some grounded approach), or the location signal is given a score of 1 and a user can have multiple such locations, thus meaning a user could be counted doubly if they are associated with multiple countries.

We refer to the case where we double count users as 'unweighted' and the case where each user is given a score of one to be split between different countries as the 'weighted' version. We use both versions in different cases. In particular, when considering networks, it is important to use the unweighted version so that we can see how many people who signify as from Egypt revert people who signify as from Syria. However, whenever we need to make use of the total number of edits (i.e. "19.7 percent of edits to Israel come from Israel") we needed to use the weighted version so that the totals add up to 100 percent.

3.3 Article and user networks

One of the ways in which we sought to identify community on Wikipedia is through the construction of social networks of the editors on Wikipedia. We did this in two different ways - relationships based on co-editing (I.e. There is a link between two editors if they both edit the same page) and relationships based on reversions (I.e. There is a directed link between two editors if one editor reverts the content of another).

Co-editing graphs

A co-editing graph is a type of "two-mode" network. That is, it is a network where there are links between people and pages. Two-mode graphs are a staple of network analysis going back to Davis et al's graph of southern women. That particular graph highlighted how co-attendance at dinner parties was related to the political power of the women and their husbands. In general, a two-mode graph is 'reduced' or 'projected' into a one mode graph. That is, if two women were at the same party, there is a link between them.

Two mode reductions in our work ended up being extraordinarily complex since there could be upwards of hundreds of editors who edit a single Wikipedia page, thereby leading to thousands of links between these individuals…for only one page. This is further complicated by the fact that pages are of different length and have differing numbers of authors. Fortunately, when one takes a two-mode graph and makes a one mode graph out of it, the edges between the nodes (or the relationships between the editors) are weighted. If two people attended three parties together, the edge between them has a weight of three. We used this weight to filter the networks down to a manageable size. We weighted the number of edits in the same page so that two editors who did a lot of work on a small page will be given a greater weight than two editors who did less work on a larger page. Consequently the value of the weight does not correspond to any tidy number such as 'number of edits' or 'number of articles'. Nevertheless it more fairly represents the joint work of editors.

We then filtered the networks down to a weight of 8 or greater. This number approximately represents co-editing 8 articles together. A value of 8 was chosen based on the distribution of edge weights. Any more any we start to dramatically eliminate edges, any less and we do not eliminate enough edges.

The resulting networks from this analysis were shown at WikiSym, featured as networks on our Wikiproject page, presented at the Sunbelt International Conference on Social Networks and featured in our conference booklet "Representation on Wikipedia in the Middle East and North Africa". We show the networks for editing in Arabic and English below. These were the versions presented at our conferences. At twelve conference members identified themselves in Cairo and ten did in Amman. They were evenly spread across the networks. Names can be cross checked with the Wikipeida names presented in our table of attendees. Additionally, participants were quick to point out other members of the Wikipedia community they know (such as for example the group of Iranian co-editors in the lower right hand corner of the English network).

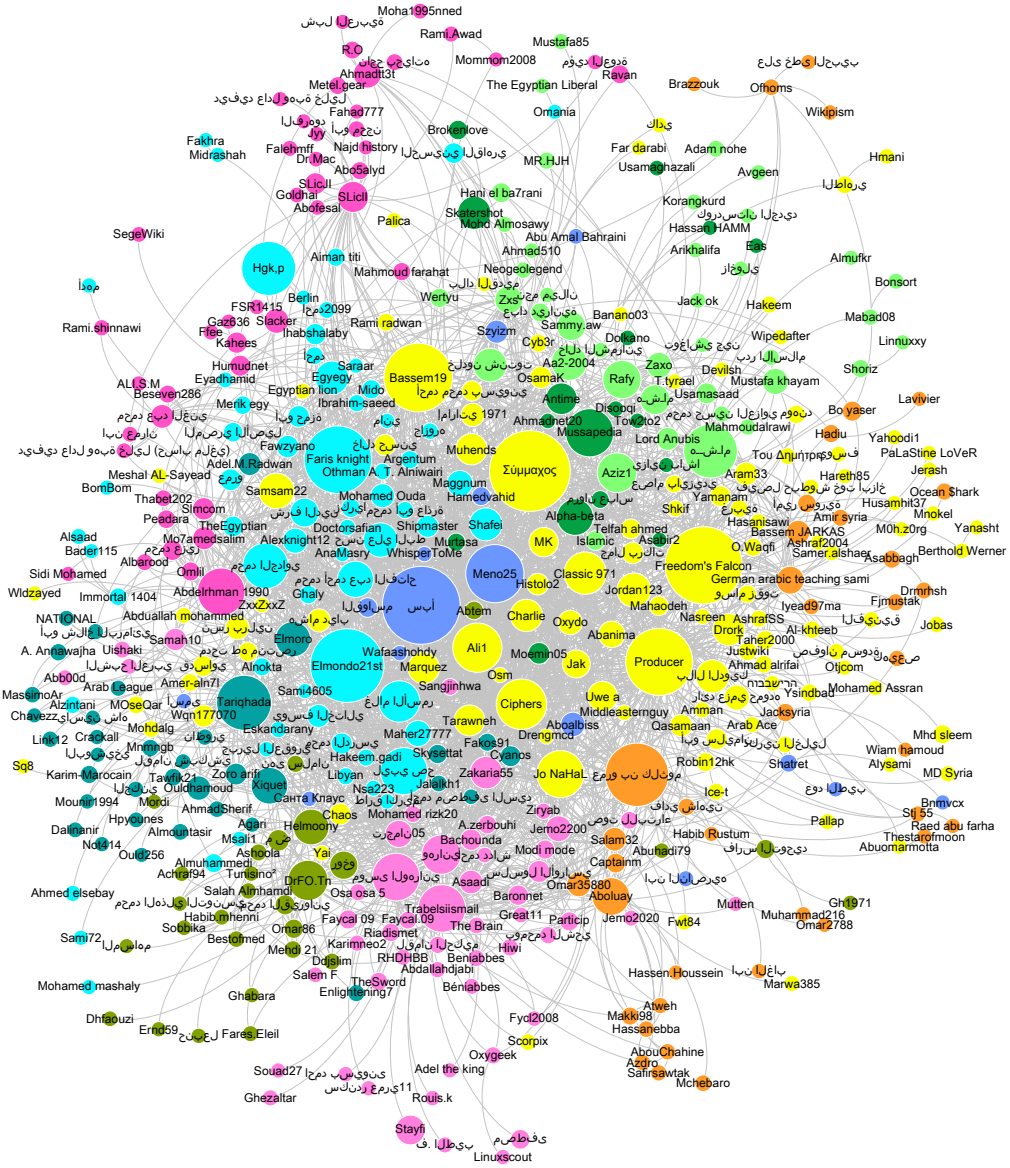

Figure 3.3a. Network graphs of the Wikipedia co-editing structure in Arabic. A link between two authors means they edited many of the same articles (approximately eight or more). The layout automatically arranges authors so that people who co-edit are close to each other. The colors represent automatic clusters of groups who edit many of the same articles. The larger the circle, the more an author has edited MENA articles.

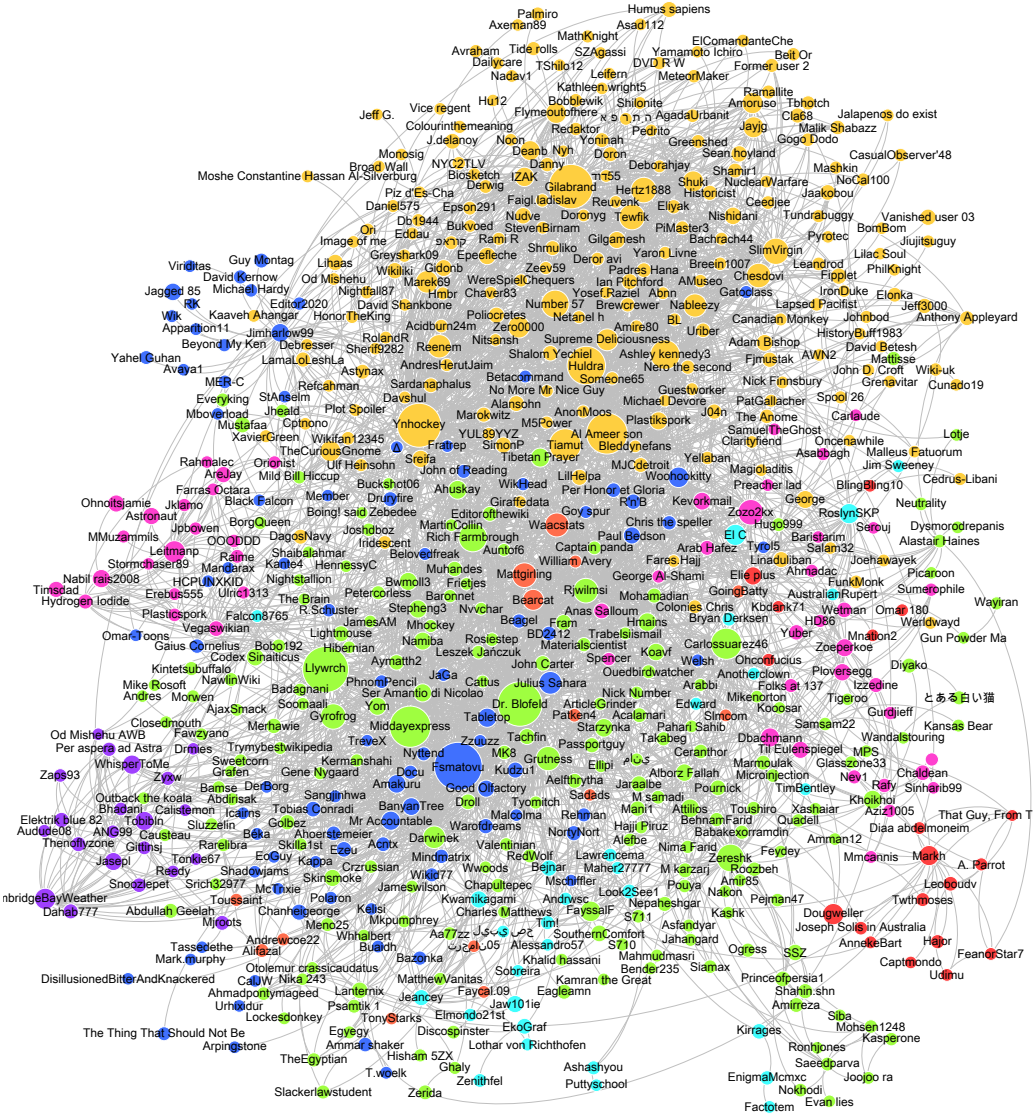

Figure 3.3b. Network graphs of the Wikipedia co-editing structure in English. A link between two authors means they edited many of the same articles (approximately eight or more). The layout automatically arranges authors so that people who co-edit are close to each other. The colors represent automatic clusters of groups who edit many of the same articles. The larger the circle, the more an author has edited MENA articles.

Reversion network graphs

We were only able to produce reversion network graphs very late in the project, as they depended on our identification of authors. We could have presented reversion networks of every editor on Wikipedia, but would not be able to make much sense of this graph. Instead, by showing the networks of reversions between different countries (where an edge represents a user from one country who reverts a user from another country), we are able to show how different nationalities in essence "police" Wikipedia. Thus, if there are many reversions from Canada to many editors from Egypt, it suggests that Canadian Wikipedians are seeking to regulate the content on this page.

For the most part, such reversions are done in good faith. Thus, having a high number of reversions means a country tends to produce poor content or vandalism and doing reversions suggest that a country is being protective or taking an active interest in the article's quality. From the analysis in Section 6.5 we can see that the United Kingdom tends to do the most reverting, whereas the USA tends to be reverted the most on geographic articles in the MENA region. We can further see a stark variation in countries in the MENA region on reversion. Israel tends to revert other countries, thereby being very protective of its articles. Egypt also does this to a limited extent. However, content about smaller countries such as Tunisia, Qatar and Yemen appears to be almost entirely policed by the West (especially the UK). These editors tend to revert content coming primarily from anonymous editors in the U.S.

3.4 Online editor surveys and text input

Prior to our second workshop we conducted a survey of participants and potential participants to assess their impressions of Wikipedia, editing, identity issues and conflicts. This survey received 19 responses from attendees to the Amman workshop. The results of this survey inform Section 6.6 where we discuss qualitative responses to representation on Wikipedia. It is not meant to be generalizable to Wikipedians, but rather to ask about the sort of issues related to this project. Our goal was to be able to review this survey before the workshop in order to better understand the motivations and goals of our participants.

We sought additional input from people who were not available at the Wikipedia workshop through a wiki page hosted at the Oxford Internet Institute. We opted for a wiki since it was assumed that all editors would be familiar with the wiki format. Not only did editors abide by our set up, but actually cleaned up and arranged each others answers in keeping with the spirit of editing in good faith. This wiki was limited to OII IP addresses several months after the completion of the Amman workshop as interest had waned and we started to see vandalism creep.

3.5 Workshop participation

The workshops were not merely a form of capacity-building, even though they worked very effectively in this regard. They were also an opportunity to hear from Arabic and other MENA Wikipedians about the issues, concerns and limitations of Wikipedia. The detailed schedule for the workshops is covered in Section 4.2. In general, we alternated between three different activities: one on one interviews, conducted by Ilhem Allagui, Heather Ford, and Mark Graham, smaller group sessions organized by Bernie Hogan and Mark Graham and whole group sessions run by all organizers. Different insights emerged both from different techniques and from the different workshops. In particular, because the Cairo workshop exclusively featured primarily Arabic participants (and one Iranian who also edited in Arabic), there was a great deal of discussion about

the Arabic language and Arabic cultures. On the other hand, the Amman workshops included two Israelis, and tended to be less critical of Arabic culture. Instead, the focus was on conflict resolution, neutral point of view and ways to bridge between cultures. In both workshops there was a great deal of dissuasion both on and off the record about political intervention by governments on Wikipedia. Much of this discussion is summarized in Section 6.6.

These sessions in both English and Arabic have been transcribed and focused coded. They have been paired with text input from the sessions done using the collaborative software *Etherpad*, one of the technologies behind Google Docs, but open source and openly accessible without a Google account. Etherpad allows for multiple participants to navigate to a webpage and edit concurrently. Like the wiki document, our participants had a knack for extensive documenting and cleaning up each others' sections. It also supported text input in both Arabic and Latin languages. We strongly recommend the continued use of Etherpad for collaborative working.

3.6 External data

We paired our geocoded data with other datasets made available from Wikipedia and other relevant sources. In particular, we made use of Wikipedia's sampled geolocation data.[8] Since 2009 Wikipedia have been releasing quarterly snapshots of sampled views and edits. We have downloaded and made extensive use of this data. It has both advantages and disadvantages.

The greatest advantage of this data is that it covers all Wikipedias, not merely the 44 'principal' languages that we use or the 7 key Wikipedias of interest. Thus, when we compare between countries, having data for all edits is extremely valuable.

The greatest disadvantage of this data is the resolution. It is very low resolution data. Wikimedia create this data by querying the IP address of every 1000th view or edit and then multiplying the resulting count by 1000 to get a close estimate of Wikipedia's relative activity in multiple countries. What this means is that countries with very small editing populations (say less than 1000 edits in a month) run the risk of never being discovered in this data. Fortunately, we have data for every quarter from 2007 to 2012. What this means is that we can correct for this sampling bias by inferring whether the number ought to be higher than it is. For example, if a country has 3000 edits one month (i.e. an IP address from that country was picked up three times by Wikimedia) and then next month it has 0, followed by 4000, we can infer that the intervening time was probably closer to 3000 but was simply not picked up by the 1 in a 1000 parser. For views, most countries have very large scores so this is not a problem. For edits, however, such a correction is particularly important for estimating activity in low activity places.

The second external data we make use of is the World Bank's global data. The world bank provide annual statistics on population, broadband use, GDP and other indicators of economic and social development. We exported these tables and merged them with our own work for several analyses in this report.

[8] http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryOverview2009Q3.htm

4. Project Activities

4.1 Data Analysis

A substantial dimension of this project was focused on data analysis, cartography, network analysis and natural language processing. The results of these analysis are found in "Project outputs" (Section 5) and in the lengthy Project Outcomes section. Given the extensive coverage of these activities in these sections and the detail on methods provided in Section 3, we wish here only to note that PI's Graham and Hogan were actively involved in managing a significant amount of activities in this domain with colleagues from Egypt, Tunisia, United Arab Emirates, Switzerland, United Kingdom, and Italy. This activity was conducted primarily from Oxford with coordination via Google Docs and Google Drive.

4.2 Meetings

Below we detail the three main meetings facilitated by this project.

Place	Date	Objectives
Oxford	27th-28th April, 2011	Consolidate research team and align goals with the research objectives.
Cairo	21st-22nd October 2012	Explore the barriers to participation on Arabic Wikipedia and to MENA related articles on English and other Wikipedias.
Amman	26th-27th January 2013	Investigate representation of MENA on Wikipedia: who represents the Arab world online, who are the people that should represent it and do they get to have their say?

Both workshops, including the interviews of individuals that took place at them, were reviewed and approved by the Oxford Internet Institute's Departmental Research Ethics Committee, a sub-committee of the University of Oxford's Central Research Ethics Committee.

4.2.1 Cairo

4.2.1.1 Planning

The overall objective was to hold discussion-led workshops for Wikipedia editors contributing on Arabic Wikipedia or to MENA-related topics on English Wikipedia. The aim of these workshops was to disseminate results of the research undertaken so far, and to further explore reasons for lack of representation of the MENA region on Wikipedia by engaging with Wikipedians themselves. The initial plan had been to hold a workshop in Beirut and Cairo in each year of the project, but various obstacles to arranging these meant that the decision was taken to run two larger workshops; one in Cairo in 2012 and one in Amman, Jordan in 2013.

The team met with difficulties when initially attempting to recruit participants for the workshops. The plan was that the majority of Wikipedians invited would be local to the city in which the

workshop was being held, with separate travel 'scholarships' offered for those active Wikipedia editors coming from further afield. Invitees were chosen by compiling a list of editors who had the most number of edits in Arabic Wikipedia over the past five years, and then systematically attempting to contact them through their user pages on Wikipedia, either by leaving a message on their talk page, or by email if their email address had been made available. There were two clear problems with this method. Firstly, several of these editors, despite having made a large number of edits over the past five years, had not been editing recently, or had left Wikipedia altogether, so were unreceptive to our communications. Secondly, there was a certain lack of trust among Wikipedians who saw our communications with their community as, at best spam, and at worst, a scam. Despite the efforts made to provide relevant links to the project and University websites as well as contact details for the lead researchers and the promise of a catered event, as we later found out, many were sceptical that anyone would be offering them a 'free lunch.' There was also concern that we had not sent out the invitation through the Wikimedia Foundation.

After the initial difficulties arranging the workshops, a project manager was recruited to take over their organisation so that the researchers could concentrate on developing the workshop schedule and other aspects of the overall project. The process for recruiting participants was changed, and it was decided that to attract a larger and more diverse range of editors funds would be used for providing travel and accommodation for participants over the course of two workshops rather than four. Due to the volatile situation in Syria that was threatening to affect Lebanon, the decision was taken to hold one of the workshops in Amman rather than in Beirut. There had been trouble in Cairo earlier in 2012, but at the time of planning events had calmed down so the first workshop took place there at a hotel outside the centre of the city.

4.2.1.2 Participants

We created a project page on Wikipedia as a reference point for those interested in the event.[9] Editors were again invited through their talk pages and using any other available contact information, but this time invites were sent to those Wikipedians that had recently edited articles on relevant subjects on both Arabic and English Wikipedia. This process was a little more time consuming, but ensured that each message could be tailored towards each editor to minimize suspicion. For example, an invitation would begin 'we noticed that you have recently edited these particular articles (for example about a middle-eastern conflict or on an aspect of Arab culture) and thought that you might therefore be interested in participating in our project…' Including this kind of personalised message sent to recent editors, along with the project website, links to researcher profiles and contact details increased the proportion of invitees that responded. This tactic meant that we had no previous knowledge of the nationality or location of the invitees, but due to the decision to fund participants' travel and accommodation this time we could ensure that most of those willing to come were able to. This also meant that the workshops involved a diverse range of Wikipedians providing a large range of perspectives for the research. We set up a Facebook group and sent the link to respondents so that they could interact with the other attendees in advance of the event.[10] This too helped reassure participants that the workshops were legitimate and gave them the opportunity to discuss ideas in advance and see which other editors would be attending. Table 4.2.1.2a provides details of the participants that attended the workshop at the Sheraton Dreamland Hotel in Cairo, 21st-22nd October 2012.

[9] http://en.wikipedia.org/wiki/Wikipedia:Meetup/Cairo

[10] https://www.facebook.com/groups/menawiki/

Table 4.2.1.2a. Cairo workshop participants.

Attendees	Wikipedia username	Country
Anas Eljamal	aboluay	UAE
Mohamed Mostafa Ouda	mohamed ouda	Egypt
Dr Nidal Yousef		Jordan
Nasser (Farsi Khaled)	Nasser	Algeria
Farah Jack Mustaklem	fjmustak	Palestine
Omid Mojabi	princeofpersia1	UK
Mohamed Amarochan	Mohamed Amarochan	Morocco
Faris El-Gwely	Faris_knight	Egypt
Moemen Metwally	عبد المؤمن	UK
Maysara abudlhaq	uwe_a	Egypt
Espada the Dark	Espada the Dark	Egypt
Salah Almhamdi	Salah Almhamdi	Tunisia
Wael Ghabara	Ghabara	Tunisia
Ahmed Koraiem	Koraiem	Egypt
Mohammed Farag	Meno25	Egypt
Aya Mahfouz		Egypt
Doaa seif	Doaasaifeldine	Egypt
Ahmed Yousif	ashashyou	Egypt
Yara Kassem		Egypt
Mohammed Bachounda	Bachounda	Algeria
Kamal Osama Elgazzar	Kamal Osama Elgazzar	Egypt
Kareman Baknam	Kareman Bknam	Egypt
Walaa Abd El-Monaem	مستخدمة:ولاء عبد المنعم	Egypt
Maram Gamal	M.Gamal	Egypt

4.2.1.3 Objectives

The objective of the Cairo workshop was to explore the perceived barriers to participation on Wikipedia when editing or contributing articles on MENA related topics and places. We asked Wikipedians how articles could be made better, and explored what could be done to expand and improve Arabic Wikipedia. The workshop was held over two days at the Sheraton Dreamland Hotel on the outskirts of Cairo, and was fully catered with accommodation provided for those who

came from outside the city. All sessions were recorded, transcribed, and then coded; the analysis of these sessions can be found in Section 6.6.

The results of research already undertaken were shared, both in a booklet handed out to each participant, and presented as a talking point for discussion. The booklet was written in both English and Arabic, with the translation work done by project members to ensure that the intent of various maps and descriptions were fully represented. Initially participants were split into groups to discuss the maps showing the disparities in number of Wikipedia articles between the MENA region and other areas of the world, focusing on the reasons the MENA region should be represented on Wikipedia, and why this representation is so relatively small. The second session focused on those who *are* contributing to Arabic Wikipedia, including organisations, projects and individuals. The group were shown the network map of Wikipedia editors on Arabic Wikipedia, in which several of them could find themselves and their connections to one another. Participants were asked to talk about what kinds of initiatives encouraged participation in Wikipedia, which had worked in the past and which had not. The final session of the day included a talk via Skype by Mina Nagy from Taghreedat, an initiative that aims to create and grow an Arabic digital content creation community to increase the amount of quality Arabic content online. The second day of the workshop involved 'open space' discussions, where participants had the opportunity to choose an aspect of Wikipedia they wanted to discuss or learn about. These topics were then presented to the whole group and participants chose which of them they wanted to join a discussion on.

Table 4.2.1.3a. Workshop Schedule for Cairo meeting.

Sunday 21 October

Time	Activity
9:00-9:30	Registration
9:30-10:30	Introductions and goal setting
10:30-11:30	Facilitated brainstorm: Issues/barriers/worries/problems related to the representation of place
11:30-12:00	Tea
12:00-13:00	Facilitated brainstorm: Why is it important to have Arabic Wikipedia content and/or content about the MENA region on Wikipedia?
13:00-14:30	Lunch
14:30-15:30	Facilitated brainstorm: Who is participating in Middle East/Arabic topics? What are they talking about? What can we learn from this?
15:30-16:00	Tea
16:00-17:00	Institutional stakeholders: talks by WMF, Taghreedat, then recapping the key themes from the day's discussions.
20:30:00	Dinner at Sequoia Restaurant

Monday 22 October

9:00-10:00	Recap of yesterday, summary of wiki discussions, scheduling of open space
10:00-10:30	Open Space
10:30-11:00	Open Space
11:00-11:30	Tea
11:30-12:00	Open Space
12:00-12:30	Open Space
12:30-14:00	Lunch
14:00-14:30	Sharing the results of open space
14:30-15:30	Facilitated brainstorm: The future of Arabic open content
15:30-16:00	Tea
16:00-17:00	Closing session

4.2.2 Amman

4.2.2.1 Planning

The second workshop took place at the Four Seasons hotel in Amman, Jordan, between 26th and 27th January 2013. As there would be several editors requiring accommodation, it was decided that, as with the Cairo workshop, it would be simplest, most effective and economical to combine all aspects of the workshop including rooms, meals and conference facilities in one package deal with a single hotel. Recruitment for this workshop was easier due to the success of the initial event in Cairo, which had led to independent discussions between Wikipedians on social media, confirming the legitimate nature of our endeavour as well as confirmation that the previous workshop had been stimulating and worthwhile. Members of the Wikimedia foundation that the team were already in contact with also facilitated the communication with Wikipedians by sending out an email encouraging members of their network to apply. We were then approached by several editors with applications to attend that included their interests and details of their Wikipedia edit histories. These applications were assessed by the researchers and decisions were taken as a team as to whether they should be formally invited. Criteria for invitation included having edited about MENA regions, cultures and related events. As well as selecting from applications, we targeted editors who had recently contributed to MENA related articles, including those on places in the MENA region and controversial articles on events such as 'Operation Pillar of Defense.'

Participants were sent an online survey to complete in advance of the workshop in order to provide the research team with information with which to structure the time and to feed into ideas for discussion based on what the participants felt was important. The results from this survey can be found in Section 6.6.

4.2.2.2 Participants

A large proportion of the Cairo participants applied to take part in the Amman workshop, which was a testament to its success. A few were invited to take part in Jordan due to their relative

proximity to the workshop venue. It was decided, however, that priority should be given to new applicants with relevant experience in order to encourage novel discussion. We nevertheless felt that having a small number of repeat attendees would help to build continuity and facilitate post-workshop discussions among editors. Among the editors we invited (due to their extensive editing of MENA related articles) were two Israelis, who both agreed to come. We realised that this may have caused a problem among some of the Arab editors, so took the decision to make everyone aware that the Israelis were coming with the aim of diffusing any potential conflict at the event in advance. Several invitees were upset by this news and responded very negatively, with the ultimatum that they would be unwilling to attend if the Israeli participants were there. There was subsequently a great deal of activity on the Facebook group, with a limited number of invited editors and miscellaneous group members contributing to anti-Israeli threads and calling for a boycott of the workshop. Luckily this activity did not reflect the opinion of the Arab Wikipedians we had invited, the majority of whom emailed the project manager to say that in the spirit of collaboration they were happy to meet with the Israeli editors and discuss issues about the representation of the MENA region on Wikipedia.

We addressed the issue on the Facebook group by posting a detailed explanation of the selection process, the reasons the Israelis were invited and the hope that a mutually beneficial discussion could arise from their inclusion. After this post, further negative and anti-Israeli comments on the Facebook group ceased and positive responses were received by email from those editors who were planning to attend. Once the workshop was underway, all participants engaged with one another with no problems, and we managed to have productive discussions and discover shared general goals and hopes related to knowledge sharing. As a result of the workshop, some of the participants later formed a group called 'Wikipedians without borders' (https://www.facebook.com/pages/Wikipedians-without-borders/585663901453048). The stated goal of the group is: "Wikipedians from around the world reaching out, individually and collectively, to share their thoughts, ideas and projects for increasing human knowledge and cultural understanding." The creation of this group demonstrated a positive outcome from the two-day workshop that was due to, and not in spite of, the diversity of its participants.

The table below provides details of the participants that attended the two day workshop at the Four Seasons Hotel in Amman, Jordan, from the 26th-27th January 2013.

Table 4.2.2.2a. List of Amman Workshop Participants.

Attendees	Wikipedia username	Country
Ravan Jaafar	Ravan	Iraq
Mohammad Nabil Rais	nabilrais2008	Pakistan
Mohamed Ouda	Mohamed Ouda	Egypt
Habib Mhenni	Dyolf	Tunisia
Abbad Diraneyyah	عباد ديرانية	Jordan
Abdullah Hussain Mohamad Ahmad	Abdullah Ahmad	British
Zakaria Oudrhiri	Zakaria	Morocco
Fareh Abdelhak	Fareh_abdelhak	Algeria
Abbas Salamat	Elph	Iran

Elias Ziade	Elie plus	Lebanon
Ziyad Alsufyani	xziadx	Saudi-Arabia
Helmi Hamdi	Helmoony	Tunisia
Farah Mustaklem	fjmustak	USA/Palestine
Gila Brand	gilabrand	Israel
Dror Kamir	drork	Israel
Mervat Salman	Mervat Salman	Jordan
Rami Tarawneh	Tarawneh	Jordan

4.2.2.3 Objectives

The objectives of the Amman workshop were different to that of the Cairo workshop. Rather than concentrating on the barriers to editing Wikipedia and how these might be overcome, the focus was on who represents the MENA region on Wikipedia and the conflicts between opposing points of view about how MENA-related places, peoples and historical or current events should be represented online. The results of the survey were used to generate discussion about conflict on Wikipedia and 'edit wars.' Participants were encouraged to discuss issues around bias in articles, and the problems that arise when editors try to edit articles about places they have no direct experience of. There was also discussion about article deletion and the issue of notability, for example when a person, place or meme is notable in the Arab world, but may be deleted from English Wikipedia due to its lack of notability in the English-speaking world. Participants had been asked to bring laptops or tablets if possible, and wi-fi was freely available for all, ensuring that everyone could look up articles, find examples to contribute to the discussions and share them with the rest of the group.

Throughout the workshop, all participants were encouraged to contribute to *Etherpad*, a web-based real-time collaborative editor that allowed attendees to comment and make notes about the discussions as they took place. This was both for the benefit of the researchers, as a written record of the discussion that included comments and points of view that had perhaps not been expressed verbally, and for the participants, as it enabled them to add any thoughts that they had not felt comfortable volunteering verbally to the group, to share photos and contact information, and also as a way of solidifying their collaborative efforts in a medium they were used to as editors of online content.

The group were first asked to split into smaller groups to consider if and why content about the MENA region on Wikipedia (and online generally) is important. There was also a brief discussion about barriers to participation in order to gather the opinions of a fresh cohort of Wikipedians to supplement the information gathered on this topic at the Cairo workshop. Groups were asked to consider the way various places are represented on Wikipedia, and who had 'voice' about these places. They were asked to find examples of articles on places they felt strongly about, and discuss how those articles demonstrated difficulties in reaching consensus about places in the region, from names (e.g. is it the Persian Gulf or the Arab Gulf?) to political status (e.g the disputed territory of Western Sahara).

One of the exercises the attendees were asked to perform was to get into groups and write a Wikipedia article in English Wikipedia about a MENA-related Internet meme. This was part of the

exploration of how the notability guideline in Wikipedia might hinder the acceptance of articles on subjects that had gained huge notoriety in the Arab world but were less well known elsewhere. Almost a year on from this exercise, two of the four articles that were created at the workshop remain on Wikipedia as articles in their own right: http://en.wikipedia.org/wiki/Myriam_Klink and http://en.wikipedia.org/wiki/Jojo_Khalastra, with one now linking to another article and another having been deleted.

As with the Cairo workshop, open space sessions took place on the second day, this time in the afternoon, with the morning spent talking about editing non-local articles and the results of the pre-workshop survey. The workshop ended with a positive discussion about the future of Arabic open content and what could be done going forward to enhance both the quality and quantity of MENA related content online.

Table 4.2.2.3a. Workshop schedule Amman:

	Saturday January 26
9:00-9:30	Registration
9:30-10:30	Introductions and goal setting
10:30-11:30	Facilitated brainstorm 1: Is it important to you to have Arabic Wikipedia content and/or content about the MENA region on Wikipedia?
11:30-12:00	Tea/Coffee
12:00-13:00	Facilitated brainstorm 2: Who represents place in the Middle East? Have you experienced any of barriers to participation or particularly pronounced conflict when writing about MENA related articles?
13:00-14:30	Lunch
14:30-15:30	Facilitated brainstorm 3: Place and identity in participation about MENA related topics.
15:30-16:00	Tea/Coffee
16:00-17:00	Editing non-local articles
20:00:00	Dinner at Tannoureen
	Sunday January 27
9:00-10:00	Recap of yesterday, summary of wiki discussions, scheduling of open space
10:00-11:00	Editing non-local articles
11:00-11:30	Tea/Coffee
11:30-12:30	Discussion of key outcomes of survey
12:30-13:30	Lunch
13:30-14:00	Open Space group making
14:00-15:00	Open Space
15:30-16:00	Tea/Coffee

16:00 - 16:30	Open Space wrap up
16:30-17:00	Closing session and facilitated brainstorm: The future of Arabic open content

4.3 Online activities and the diffusion of knowledge

As this project involved user-generated content online, it is unsurprising that we were able to make some of our outputs as direct offerings to the web, and hopefully to users and content creators of MENA content. Our two main offerings were the Wikiproject page and blog posts about MENA and global Wikipedia efforts.

4.3.1 Wikiproject Page

We created a subdomain at the Oxford Internet Institute's domain.[11] From this page we were able to host pointers to four key elements to our research: an interactive map, interactive network diagrams, a wiki with questions about MENA content for invited contributors and a page that points to the source code for our geographic parser and other source code use during this project.

The purpose of this page was primarily to demonstrate to potential participants of our workshops that this was a legitimate, publicly-funded project about content creation in the Middle East. A supplementary goal for this page was to enable access to the core discussion questions for those who could not attend the workshops.

4.3.1.1 Interactive Map

The interactive map shown on Wikiproject was a collaboration between Gavin Bailey of TraceMedia and the project team. This map superimposes dots on a scalable world map. We use Google Maps' standard map as our background and "OpenLayers" in order to place the dots on a map. The map allows highlighting of specific countries and regions. The dots can be colour coded by a number of features, such as different language, word count of the articles and number of authors for the articles. We have determined that comparing languages on the same map reveals in great detail many of the features of Wikipedia that are discussed herein. The map has been a rousing success. We discuss this more in project outputs under "5.3 visualizations"

4.3.1.2 Networks

Part of this work has been a focus on the social structure of editors on Wikipedia and whether this structure enables or inhibits content from individuals in MENA countries. Part of our work on this was the creation of multiple co-editor networks. These are networks were individuals are dots, and there is a line of the individuals have made 'significant contributions' to the same articles. The term 'significant contribution' refers to making multiple edits to the same articles. Rather than use 3 edits or 20 edits, we create a hybrid metric that accounts for differences in article size and contribution length. In this way, two authors who make 5 contributions each to the same small article will be weighted more heavily than two authors who make 5 edits to a very large and active article.

What is notable about these networks is that while they look like ambiguous hairballs to the outside observer, we have been able to confirm that these networks look coherent to those who are in the diagram. We discovered this by showing the diagrams to Wikipedia editors at Wikisym

[11] http://wikiproject.oii.ox.ac.uk/

in 2011 as well as to the invitees to our workshops in Jordan and Egypt. The workshop participants were actually delighted and easily able to find themselves by noticing the other editors around them. PI Hogan asked the workshop attendees to autograph next to their name. His personal copy is available upon request but not publicly shared to protect the privacy of user signatures.

4.3.1.3 Local Wiki

In the run up to the Egypt and Jordan workshops we wanted feedback from participants (as well as invitees who could not come) about the core questions on the project. We understood that because the workshops would have primarily experienced Wikipedia editors that they would be comfortable with the wiki style. We prompted the invitees with the following questions:

- "Why are there such extreme differences in participation to Wikipedia between different countries in the Middle East and North Africa?"
- "Why do we see such big differences in what people write about in Wikipedia? (some parts of the world having much more written about them than others)"
- "Why is there more content in English about almost everywhere in the Arab world than in Arabic?"
- "Do you face any specific challenges in writing about Certain topics? If so, what are they? For instance, having notability challenged, articles marked for deletion, or demands for citations (where none may exist)."

As can be seen from Figure 4.3.1.3a We created custom templates that allowed us to include both English and Arabic on the same page (mainly by using simple two column tables). Invitees edited in both languages on the same page, replied to each other and built up a conversation.

Figure 4.3.1.3a. Detail from the local wiki requesting opinions on Wikipeida in the MENA region, presented simultaneously in English and Arabic.

We required users to have a registered email address before editing. Unfortunately that did not prevent vandalism to the site. Such vandalism was rare, but it started to accumulate over time. It occurred long after the workshops, but consequently the site now requires one to be on the OII network to access. We plan on compiling the wiki responses for release.

4.3.1.4 Source Code

This project involved a hybrid of quantitative and qualitative skills. Much of the quantitative work involved a great deal of cutting edge computer science code. Where possible we publish such code on our source code webpage: http://github.com/oxfordinternetinstitute/wikiproject . Here one can find content related to how we created networks based on co-editing, how to determine latitude and longitude coordinates and how to determine an individual's location from free text using a custom made gazetteer (I.e. A list of names and their spatial coordinates). All the code is freely downloadable and comes with a Creative Commons share-alike licence.

5 Project outputs

Although this project is primarily academic, it is also has a strong emphasis on both capacity building and public understand of research issues. Academic publications are discussed in Section 5.1. Our open access datasets are discussed in section 5.2. Our publicly accessible visualizations (primarily via blogs) are listed in Section 5.4. Capacity building is featured in Section 5.5.

5.1 Publications

Below are seven papers in various stages of completion that emerged directly from work in this project.

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

Authors: Mark Graham, Bernie Hogan, Ralph Straumann, Ahmed Medhat, David Palfrey.

Venue: *Annals of the American Geographer*.

Status: Second Revision.

Summary: In this paper, we employ ordinary least squares regression to model the variation in representation globally. We indicate how the MENA region falls below expected values, controlling for differences in broadband diffusion, GDP and population size. Data comes from geolocation work and the World Bank. This paper mirrors much of the analysis in Section 6.1.

Title: " Wikipedia Arabe et la Construction Collective du Savoir"

Authors: Ilhem Allagui, Mark Graham, Bernie Hogan.

Venue: Book "Wikipédia, objet scientifique non identifié". Presses Universitaires de Paris Ouest.

Status: Accepted for publication.

Summary: In this paper we qualitatively examine the factors that inhibit editing Wikipedia within the Arab world. Data comes primarily from the Cairo and Amman workshops. This paper mirrors much of the analysis in Section 6.6.

Title: Exclusion on Wikipedia and the many cultures of editing

Authors: Ilhem Allagui, Mark Graham, Bernie Hogan.

Venue: TBD

Status: Draft

Summary: In this paper we expand on the previous paper with additional qualitative coding of interviews. Data comes primarily from the Cairo and Amman workshops as well as our Wiki page.

Title: "Interactive Mapping of Wikipedia's Geographies: Visualizing Variation in Participation and Representation."

Authors: Gavin Baily, Bernie Hogan, Mark Graham, Ahmed Medhat.

Venue: *WikiSym '12*. Linz Austria

Status: Published.

Summary: In this paper we detail the interactive map application "Mapping Wikipedia" mentioned in Section 4.4.1.1 and Section 5.3. We indicate how this map can be used to highlight inequalities and eccentricities of Wikipedia.

Title: "Informational Magnetism: Mapping Participation in Wikipedia"

Authors: Mark Graham, Ralph Straumann, Bernie Hogan, Ahmed Medhat.

Venue: TBD.

Status: Ready for submission

Summary: In this paper we model the variation in participation on Wikipedia, both globally and in the MENA region. This paper mirrors much of the analysis in Section 6.1.3 and Section 6.4.

Title: "Constructing meaning through big data: Reflexive triangulation and the problem of ground truth in user-generated content"

Authors: Bernie Hogan, Mark Graham, Ahmed Medhat, David Palfrey.

Venue: *Big Data & Society*.

Status: Invited submission to be sent by January 31, 2014.

Summary: In this paper we model the geographic patterns of policing on Wikipedia through reversion networks. This paper mirrors much of the analysis in Section 6.5.

Title: "The most controversial topics in Wikipedia: A multilingual and geographical analysis"

Authors: Taha Yasseri, Anselm Spoerri, Mark Graham, János Kertész

Venue: Fichman P., Hara N., editors, "Global Wikipedia: International and cross-cultural issues in online collaboration". Scarecrow Press (2014).

Summary. In this paper we model the most controversial topics on Wikipedia in multiple languages including English, Arabic and Farsi, among others, noting which topics are particularly controversial across languages (such as U.S. presidents and Israel / Palestine conflicts).

5.2 Datasets

In the course of this work we have created a number of data sets based on Wikipedia data that we have either released publicly or are in the process of releasing. The initial geocoding of the seven

languages of interest is embedded in Tracemedia's "Mapping Wikipedia" tool and is freely available upon request. This data includes not only all geocoded articles to the end of 2011, but also a great deal of metadata such as the number of editors, clean-up tags, links and edits per article.

We are also in the process of releasing the geolocation of authors database, however, this requires significant sensitivity since we are inferring author locations. As such, we are planning on releasing this dataset alongside a means for individuals to correct specific locations and to add their own. In this regard the dataset will not be a dead dataset but a living means of signifying local representation.

Our data has already been put to use with a number of collaborators and we hope this will continue. Data from Wikimedia is given freely in good faith using a Creative Commons license, and we hope to share alike in the same regard. Data will be available at http://wikiproject.oii.ox.ac.uk/data . In the interim we have accommodated all requests for access where possible.

5.3 Visualizations

As has been noted in some of the earlier sections (and is evident from the lengthy list of figures, especially in Section 6), there is a great deal of visual analysis to our work. Such visualizations are not merely static and buried in reports. Our work has led to several visualizations that have been useful both as a form of public interest and as research tools in our own work (as well as the work of Wikipedians). Some of these visualizations are static, as we wish to ensure that a very specific message or research insight was presented effectively. Others are dynamic and encourage users to find insights within the diagrams. In general, our visualizations are largely maps, but also include some network analysis graphs as well.

Numerous maps have been presented to the public via the blogs "Floating Sheep"[12] and "Zerogeography".[13] We currently count 6 posts on Floating Sheep attributed to this project (where IDRC funding is always acknowledged), and 26 on Zerogeography. The full list is available in Appendix IV.

For a dynamic map, we worked in conjunction with Gavin Baily at Tracemedia to produce a large scale map of articles on Wikipedia. This map was meant to push the boundaries of an emerging web technology called "Open Layers", while simultaneously highlighting variations in representation across the world. Our project "Mapping Wikipedia" was picked up by the UK's Guardian newspaper who embedded this visualization on their data page, prompting over 30,000 people to try our maps within a single day.

The data for this map came from the parsing done earlier. In addition to simply identifying the geographic coordinates of articles, we also parsed and coded the number of authors, the word count, the number of tags and a few other metrics. These can be used as features for creating a novel map of a country or the globe.

[12] http://www.floatingsheep.org/

[13] http://www.zerogeography.net/

An extended version of this tool with a timeline is featured on Tracemedia's website.[14] With the timeline enabled (for the date an article was created), we can see in stark relief the relevance of local gazetteers for the creation of articles. Around 2007, one user created a series of articles automatically (using a "bot") about a variety of states in the U.S., as well as several provinces in Spain. The articles mark the boundaries of a state and appear in quick succession.[15] From either version, one can observe the variations in emphasis for the different language versions. It is not merely that there is Hebrew in Israel and more Farsi in Iran, but in many other parts of the world, there are substantial variations in what gets included.

We demonstrated this tool at Wikisym 2012 and have committed to hosting it for the foreseeable future.

Most of our static networks have been in the form of conference presentations and are available from Bernie Hogan upon request. However, two maps were included in our booklet given to conference participants. The dynamic networks were made with the cutting edge Sigma.js networks toolkit and featured on our wikiproject website.[16] These networks are the first of a long line of interactive networks produced or facilitated by the Oxford Internet Institute.

5.4 News coverage

As a core part of this project is highlighting the asymmetries of representation on Wikipedia. This is not merely an academic exercise. Consequently, we consider it fortunate that our work highlighting such asymmetries have been well-received by the news media at large. In particular we have received a great deal of traction from our earliest maps highlighting differences in content across different language groups, the follow up of this in the interactive networks, and our later work with Taha Yaseri on controversies in various languages on Wikipedia.

Our work has now been featured in such international publications as the Guardian, The Atlantic, The Economist, The Huffington Post, Wired and The Toronto Star. A particularly clever feature on the Guardian from April 2012 even embedded our interactive feature inside the article.[17]

Comments on these stories range in unsurprising ways from the trivial, "Wow there are Farsi articles in antarctica", to the profound, "It's fascinating how much of these maps corresponds with the historical expansions of the countries that spread these languages so well".[18] Many people were bemused in our earlier maps by the predominance of Swahili articles in Turkey. However, this actually has a simple explanation. A lone Wikipedian, User:Muddyb_Blast_Producer. As he notes on his user page: "Also, I've created some articles about the cities of Turkey (I did almost all the articles there are on the Swahili Wikipedia about Turkish cities, provinces, districts, and I've

[14] http://www.tracemedia.co.uk/mapping_wikipedia_timeline/

[15] Incidentally, the creator of this bot was at our Amman workshop.

[16] http://wikiproject.oii.ox.ac.uk/networks/

[17] http://www.theguardian.com/news/datablog/interactive/2012/apr/04/wikipedia-world-language-map?newsfeed=true

[18] http://discussion.theguardian.com/comment-permalink/13239837

written about a few villages). I did this because I promised someone from Turkey that I would write a lot about the cities of Turkey because we're friends".[19]

Overall, we count 25 stories in national news press. These are listed in Appendix I.

One of the ironies of this work is that it may in fact deepen this asymmetry as much of this work has been picked up in traditionally Western news media, media that is not as likely to be read by those in the Global South. If readers in the West see this as a call to action they may be prompted to write about Wikipedia when they were not otherwise. However, we also believe an alternative scenario is equally likely - the heightened visibility of these asymmetries help Wikipedia editors to be more sensitive to these topics in their own actions on the site.

Some of the key recommendations regarding these outputs is that blogging activities are to be considered valuable efforts in research exercises. Blogs allowed us to send out our work on a far faster schedule than our academic publications. Much of our peer reviewed work is still in revision, or pre-print status. By waiting until this work is published we would essentially inhibit ourselves from contributing to an immediate and important public discussion. Blogs are not a substitute for scholarly work, but should sit alongside such research when possible.

5.5 Organization

We initially believed that our workshops would be a form of capacity building for Wikipedia editors. While this is true to some extent, we now believe that the workshops are to be considered a specific event in a longer term process that is sustained online. Our earliest attempts to generate enough interest for a workshop were in vain. This was quite unexpected. Our researcher, Richard Farmbrough is one of the most well known names on Wikipedia with over a million personal edits. He shows up prominently in the network map of co-editors on MENA topics and is highly visible in a number of Wikipedia organizations. However, when Mr. Farmbrough messaged people on Wikipedia, typically by posting on their user page, we received remarkably few responses. Consequently, we postponed our workshop from 2011 to 2012, and adjusted our strategy.

As might be obvious in hindsight, especially following the role of social media in the Arab Spring, what really helped us were interpersonal networks facilitated by Facebook. We started a Facebook group, MENAwiki[20], and invited some key contributors. These contributors then invited their friends and co-editors. We coupled this initiative with advertisements in Wikipedia-specific venues, such as the Village Pump, stating explicitly that we were looking for editors from the MENA or editors with extensive experience in the MENA region (which ended up including editors from Pakistan and editors from the UK and Canada of Arabic origin). Most all of these individuals joined our Facebook group, which to this day is active with regular stories about Wikipedia and Arabic content.

The workshop thus served the vital function of building trust through shared co-presence, but it was part of a larger endeavour that existed virtually. After the workshop people were able to put a face to a name, so to speak, which helps with coordination, conflict resolution and information diffusion. Wikipedia understand this and use this idea to promote Wikimania, the annual

[19] http://en.wikipedia.org/wiki/User:Muddyb_Blast_Producer

[20] https://www.facebook.com/groups/menawiki/

conference for editors of Wikipedia (rather than Wikisym, the annual conference for academic studies on Wikipedia and related platforms). Wikimania has been held in both Egypt (in Alexandria) and in Israel. However, it does not provide much funding for local editors, and no organizational support. In our case, we were able to provide both. Organizational support was essential for several individuals who required visas in order to travel to our locations. Thus, editors from Iraq and Pakistan who previously could not attend any such meeting were delighted to be able to discuss Wikipedia issues with their fellow editors. As one user wrote to us:

> *"It was a wonderful experience of sharing our thoughts and experience of editing article (In fact most controversial articles) on Wikipedia, i have given an opportunity to share my problems, while editing articles on Wikipedia with you people and participants of workshop.*
>
> *Well !!! i have never ever thought when i did my first edit on Wikipedia that i will get this lifetime opportunity to attend this fruitful workshop in my life ! so i really thank you people for organizing such a wonderful workshop, and in my opinion Wikipedia Foundation should also try to arrange such workshops in order to fill the gaps in editing on Wikipedia specifically on such regions that are neglected on Wikipedia."*

Thus, in future we believe that such activities will be very useful. However, we believe that simply giving a blank cheque for funding for bonding among high profile content creators is not as useful as funding organizational efforts, such as Wikimedia or other associated NGOs who will be able to write letters of invitation, sort out visas, maintain and moderate online groups and provide a "light touch" where necessary.

The latter point about a 'light touch' is particularly relevant for work in this area. In particular, it is well known that there are significant tensions between Israel and the rest of the MENA region. So when we received invitations to our workshops from two Israelis, we were faced with a conundrum: exclude them to create a safe space for Arabic speakers or include them to signify that no voice should be marginalized. To guide our decision we privately emailed all other workshop participants to let them know the Israelis would be there. Sadly, two participants flatly rejected being at such a workshop if this was the case, and subsequently did not attend. However, the vast majority of participants did not have a problem and effectively said that the actions of a government should not be taken out on citizens. Beyond this, we also had to contend with some members in the Facebook group who posted explicitly anti-Zionist (and arguably anti-Semitic) messages. We reluctantly removed these messages as we believed that by leaving up this content we were creating a hostile environment.

We consider this tactic of democratic good faith to have been successful. Not only did the Israelis make friends with the other participants, but they highlighted many cases of how to cope with content that is politicized (such as the article on Golan Heights) and eventually founded a group called "Wikipedians without Borders" to help resolve such issues in the future.

6 Project outcomes

6.1 Wikipedia: A story of uneven geographies

There are now two and a half billion Internet-users on our planet, a number that represents over a third of humanity. Those billions of people connect to the Internet in order to communicate, access information, and share knowledge. Within the cacophony of competing narratives, ideas, stories, and voices that constitute the Internet, there are a handful of platforms that stand out as objects of particular attention. One of those is Wikipedia. The website is the world's largest collaborative reference work (or encyclopaedia) and contains over thirty one million articles written in just shy of three-hundred languages. It is estimated that over a hundred million hours of labour have been put into Wikipedia (Geiger and Halfaker, 2013): an effort that puts Wikipedia on par with some of humanity's most ambitious feats of engineering (Shirky 2010; Graham 2011a, 2011b).

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 Encyclopædia Britannica. 15 percent of all Internet users access it on any given day. It exists, as mentioned, in 287 languages, while 40 of those language versions have over 100,000 articles, and the English one alone contains over four million articles (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.[21]

Wikipedia is often seen to be both an enabler and an equalizer. Every day 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. This structural openness combined with the visibility of its content (Wikipedia is both one of the world's most accessed websites and almost always appears prominently in search results) has led highly visible commentators to now argue that Wikipedia's information is theoretically accessible to all, and anyone can have their say. For example, 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.*"

Indeed, despite the rhetoric that one can contribute regardless of place, one's location is intimately bound with where people actually decide to edit. By looking at anonymous edits to Wikipedia Hardy, Frew, and Goodchild (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, Hogan, and Medhat (2012; 2013)[22] have published preliminary results showing similar patterns.

---

[21] 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).

[22] See also http://www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html

Whereas most European and East Asian countries have more Wikipedia articles about a country in the dominant language of that country, 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) and Graham (2014) have demonstrated, fundamentally different narratives can be (and are) created about places and topics in different languages.

Despite Wikipedia's structural openness, there have been many fears that the platform simply reproduces worldviews and knowledge created in the Global North at the expense of Southern viewpoints (e.g. Graham 2011b; Ford 2011). There are indications that global coverage in the encyclopaedia is far from even: with some parts of the world heavily represented on the platform, and others largely left out (cf. Hecht and Gergle 2009; Graham 2011b, 2013b, 2014; and, in a wider context, Elwood 2010; Sieber and Rahemtulla 2010; Haklay 2013). 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).

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.

In our research, we address three facets of the geographies and networks of Wikipedia: articles, languages, and edits, to deepen three fundamental issues: representation, potential reach and access, and participation.

We deal with these three issues first at the global level in order to contextualise the MENA region, and then subsequently focus on the local geographies of Wikipedia in the MENA.

6.1.1 Articles

We begin this global analysis by descriptively examining the geography of information in Wikipedia with respect to the total number of geocoded articles in our data set. As such, it is important, to map and measure what Wikipedia actually reveals about our world.

Figure 6.1.1a displays the number of geotagged articles across all captured languages per country. The first thing we can observe is simply the incredible human effort that has gone into describing some aspects of a place.

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 large counts in many European countries. In France, for instance, there are 632,038 articles: which is more than the USA, Japan, Australia and India.

In total there are 3,924,308 articles present in our database on a total of 30,165,105 Wikipedia articles. Most geo-coded Wikipedia articles are located in the countries where the language is listed as an official one. In fact, over 70% of articles written in languages spoken primarily in a single country exist on in the Wikipedia for that language. This means, for instance, that there might be articles about thousands of Czech villages written in Czech, without translated articles occurring in English, French, German, or Japanese.

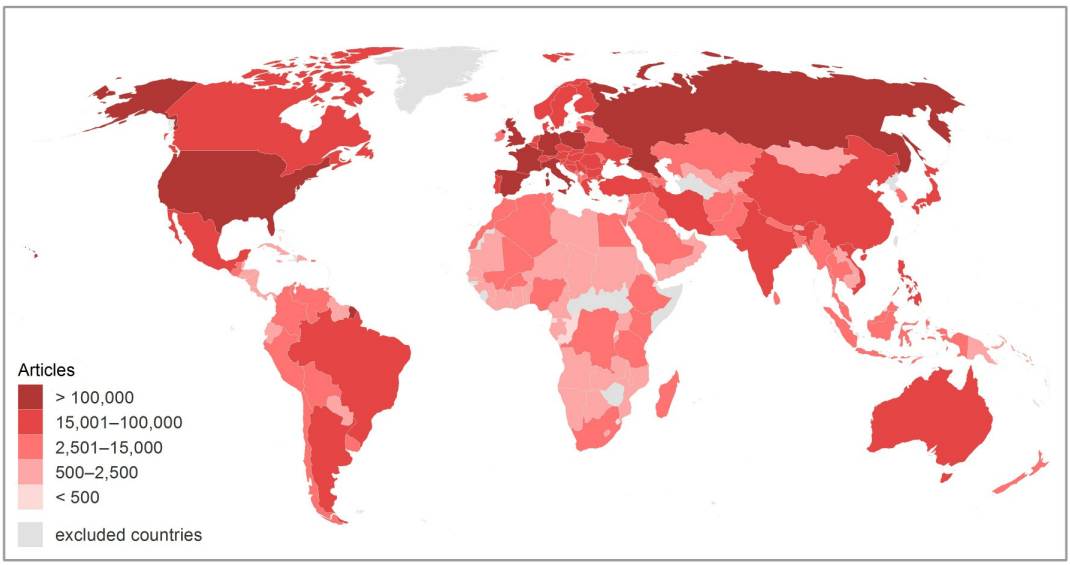

Figure 6.1.1a. Total number of geotagged Wikipedia articles across all 44 surveyed languages.

There is a clear and highly uneven geography of information in Wikipedia. Europe and North America are home to 84% of all articles. Yet, Equatorial Guinea has a paltry 230 across our principle languages, and Kuwait has only 404 (Fifty-four of which are in Arabic and 107 in English). Most small island nations and city-states have fewer than 100 articles. However, it is not just microstates that are characterized by extremely low levels of Wikipedia representation. Almost all of Africa is poorly represented in the encyclopaedia. Remarkably, there are more Wikipedia articles (14,959) written about Antarctica than any country in Africa. And there are more geotagged articles related to Japan (94,022) then for all the MENA region (88,342).

Even China, which is home to the world's biggest population of Internet users, contains fewer than 1% of all geotagged articles. See Figure 6.1.1b to put these asymmetries in perspective.

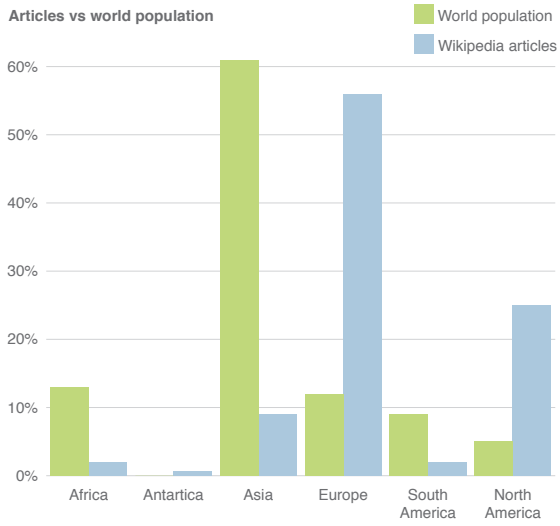

Figure 6.1.1b. Comparison of world population and number of geotagged articles.

Because of the high visibility of Wikipedia in online information ecosystems, countless decisions are made and countless opinions are formed based on information available in the encyclopedia. It is thus important to point to this broad digital *terra incognita* that covers much of the world and potentially reproduces existing representational asymmetries.

6.1.2 Languages

Measuring the total quantity of articles about a place only tells one part of the story. With Figure 6.1.2a we can explore the geolinguistic contours of digital information.

The first map illustrates the language that most Wikipedia articles exist in per country. The broad pattern that we see is once in which some countries are able to define themselves in their own languages and others appear to be largely define from the outside.

For instance, almost every European country has more articles about itself in its dominant language than any other language. In other words, there are more articles in Czech about the Czech Republic than there are English or German or French articles about the country. There are similarly more German articles about Germany there are English or French ones.

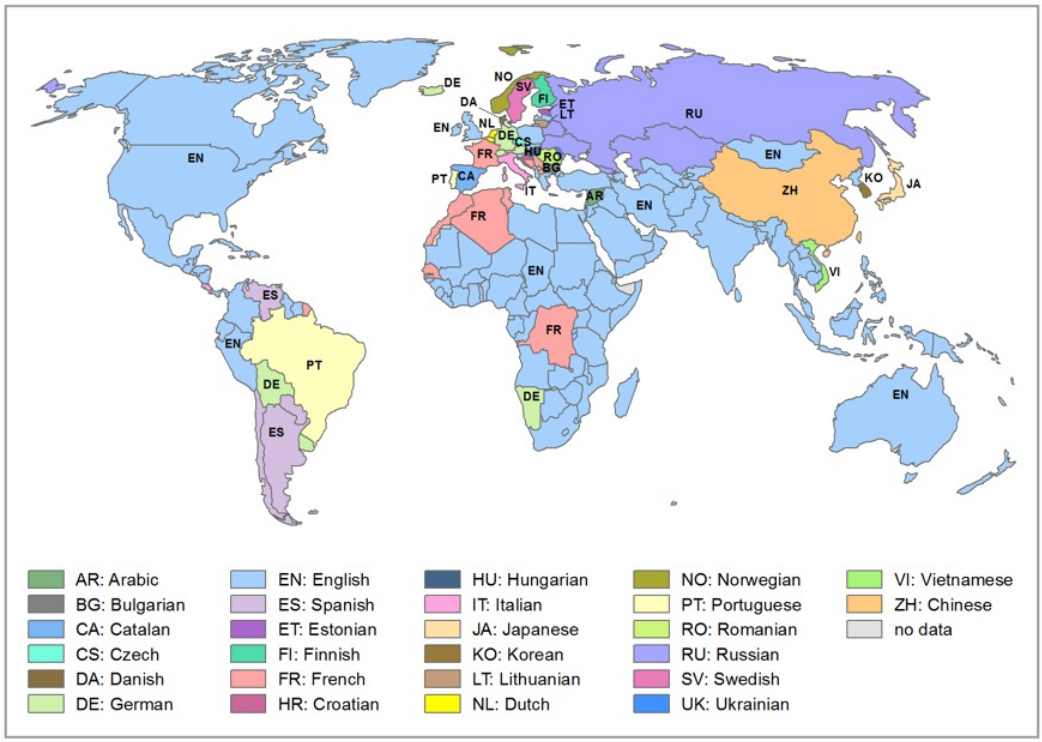

Figure 6.1.2a. Language with the most geocoded articles by country (across 44 top languages on Wikipedia).

But then we do not see that pattern across much of the South. English (which is the blue shade) is dominant in much of Africa, the Middle East, South and East Asia, and even parts of South and Central America. We then see French (the pink shade) in five countries in Africa (other traditionally Francophone countries like the Ivory Coast still have more content in English). German (the green colour) is dominant in one former German colony (Namibia) and a few other countries scattered around the world (e.g. Uruguay, East Timor).

And the scale of these differences results in some almost implausible comparisons. So, for instance, not only are there more Wikipedia articles in English than Arabic about almost every Arabic speaking country in the Middle East (Syria being the only exception as of 2012), but there are even more English articles about North Korea than there are Arabic articles about Saudi Arabia, Libya, the UAE and many other countries in the region.

Thus, not only do we see most of the world's content written about global cores, but the content that is written about the rest of the world ends up being primarily in a few languages of these global cores.

Below we can also explore some of the specific linguistic geographies in more detail. Figure 6.1.2b shows the total number of geotagged Wikipedia articles in English per country. The sheer density of this layer of information over some parts of the world is astounding (928,542 articles about places exist in English).

There is clearly a lot of unevenness in the amount of content about places, and large parts of our planet are still invisible from these digital augmentations, but it is still hard not to be awed by this cloud of information about hundreds of thousands of events and places around the globe.

Nonetheless, in this layer of English content, it remains that only 3.23% of the (geotagged) articles are about Africa and 1.67% are about the MENA region.

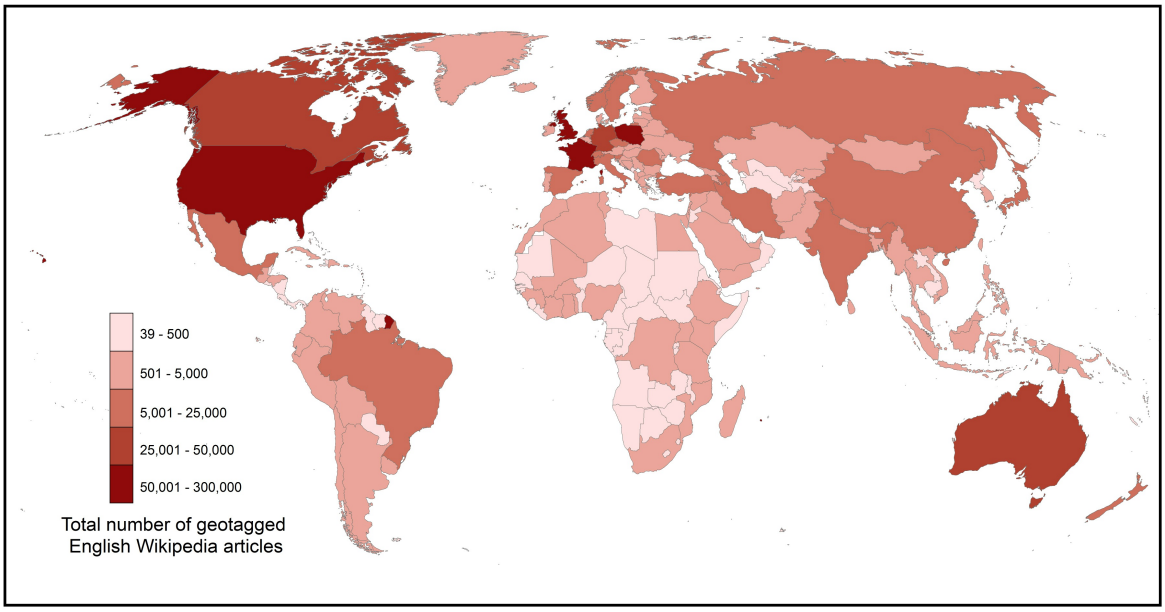

Figure 6.1.2b. Total number of geotagged articles in the English Wikipedia by country.

We see somewhat different patterns when looking at the global geography of all 22,548 articles contained within the Arabic Wikipedia (Figure 6.1.2c). Algeria and Syria are both defined by a relatively high number of articles in Arabic. The US, Italy, Spain, Russia and Greece are defined by a similar number. Yet these information densities are substantially greater than the layers of information that we see recorded over many other MENA countries in which Arabic is an official language (such as Egypt, Morocco, or Saudi Arabia). These geographies of Arabic information are even more surprising once we realise that whilst the populations of Italy and Spain are smaller than the population of Egypt, there are nonetheless between four and six times more Arabic articles related to Spain (1,988) and Italy (2,428) than Egypt (433).

Although we had entered this project anticipating that MENA region would be under-represented in English, we had not anticipated the degree to which it would be under-represented in Arabic.

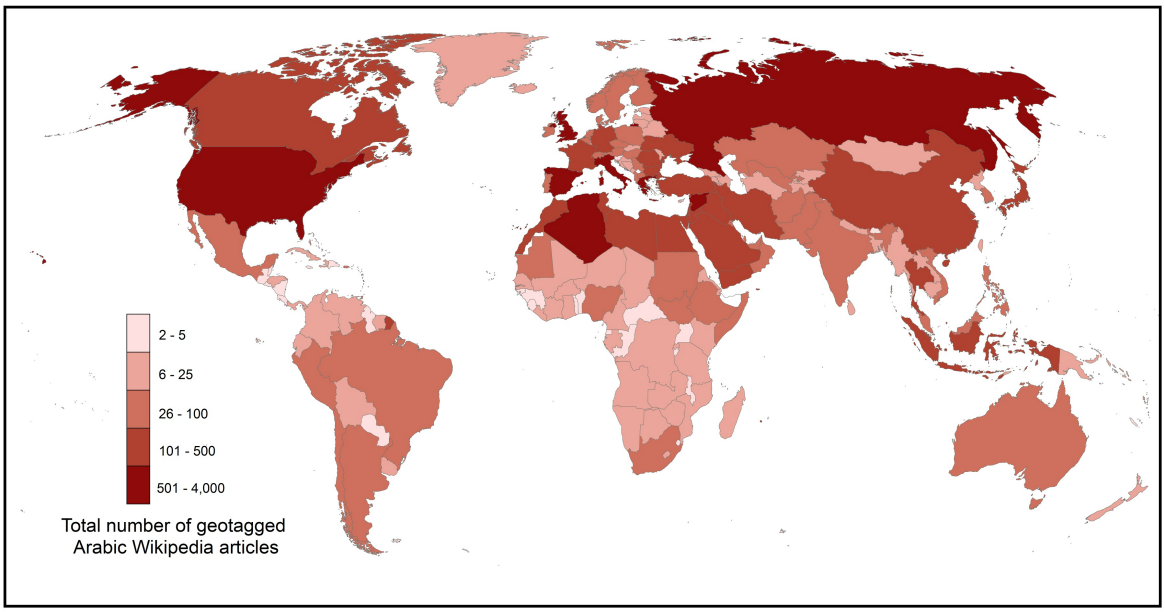

Figure 6.1.2c. Total number of geotagged articles in the Arabic Wikipedia by country.

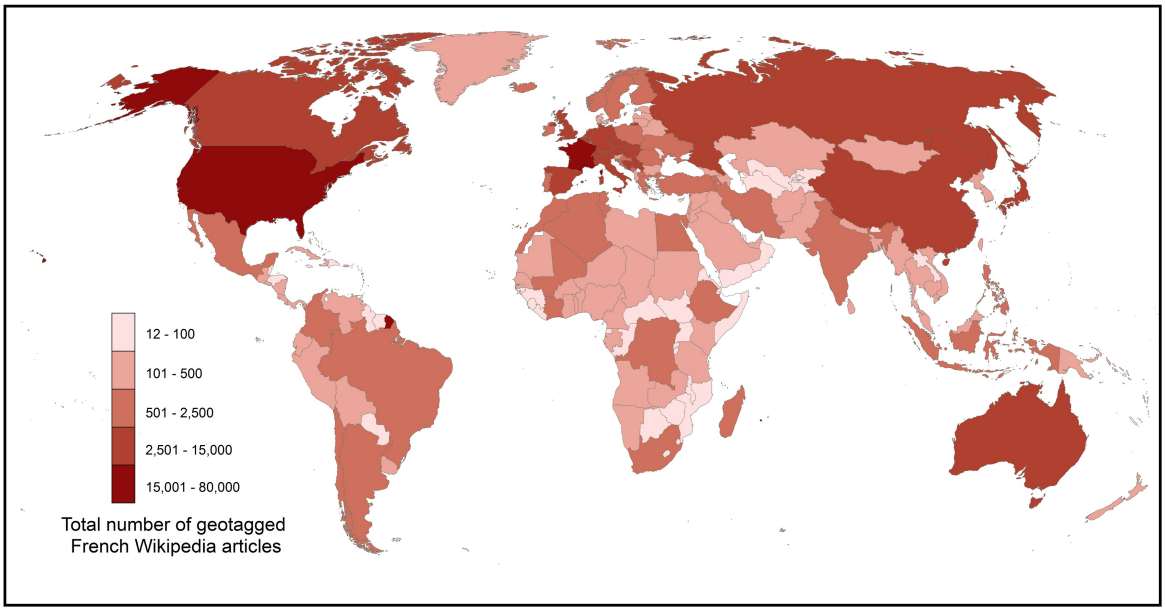

Figure 6.1.2d. Total number of geotagged articles in the French Wikipedia by country.

The situation doesn't change radically if we look at the spread of activity produced in the French Wikipedia (here we mapped 253,840 articles). Here the US and France alone occupy 37% of all articles about places. One tier below, we see Canada, Russia, China, Australia, and much of Europe: all places defined by dense layers of content in French.

It is important to point out that non-negligible values are recorded in Francophone MENA countries suggesting a remnant of cultural and linguistic ties with a former colonial power. Whilst

the MENA region occupies 2.25% of all Wikipedias, it occupies 1.67% of the English Wikipedia, 2% of the French Wikipedia and 20.9% of the Arabic Wikipedia.

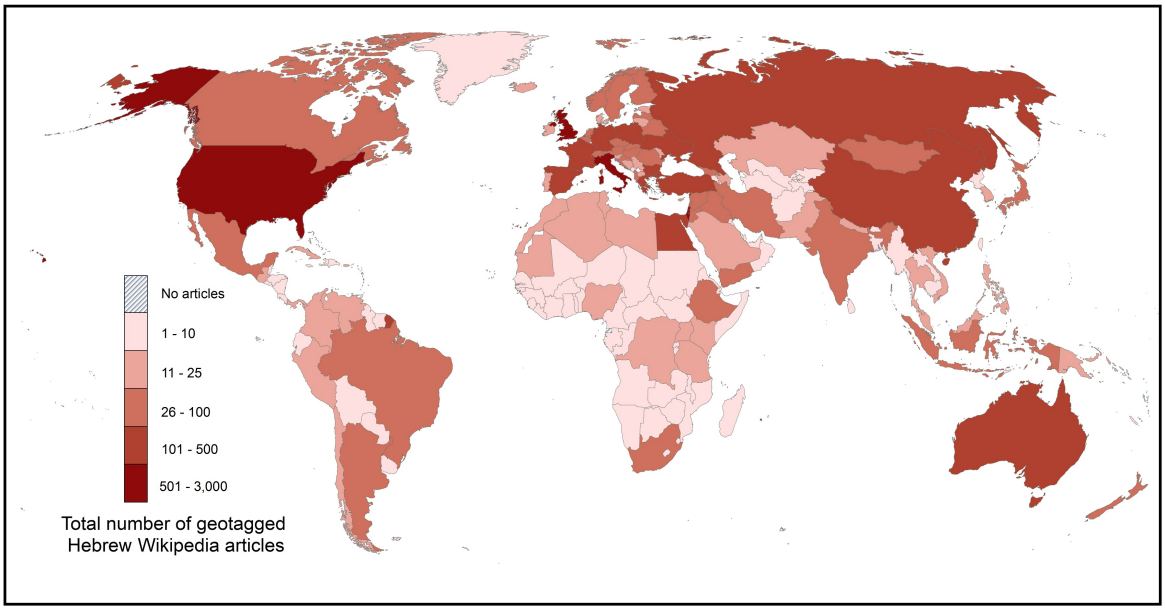

Figure 6.1.2e. Total number of geotagged articles in the Hebrew Wikipedia by country.

Mapping the Hebrew Wikipedia tells a somewhat different story, although the general pattern is similar to the one that we have seen with the Arabic and French platforms. Again, we see very high counts in the UK, the US, Italy, Russia, China, and Australia. Representation of the MENA region is relatively sparse apart from Israel and the Palestinian Territories (which respectively contain 2,517 and 715 articles which together are 33% of the entire corpus of geotagged Hebrew Wikipedia articles. The Hebrew Wikipedia contains 9,707 articles). Said differently, the Hebrew Wikipedia operates with a dual focus: the nation of Israel and the Palestinian Territories as well as some of the world's more traditional informational cores.

Finally, Figure 6.1.2f maps the number of geotagged Wikipedia articles in Farsi.

We could say similar things about Wikipedia articles in Farsi to that of Hebrew. Here we see that Iran and the United States occupy the a large amount of attention (23.7% of articles about places in the Farsi Wikipedia are about Iran and 16.7% are about the US). Some of Iran's regional neighbours do occupy a relatively large amount of attention in Farsi. For instance, 172 articles have been created about Iraq and 139 about Saudi Arabia. This compares to the 649 in the UK and 746 in France.

Otherwise, very few articles exist about Africa, Asia, and South America. So we again see a focus on the 'home' region and a large focus on the already highlight visible informational cores of Europe and North America.

In sum, by mapping the geography of Wikipedia articles in both global and regional languages, we have been able to understand the layers of representation that augment the world that we live in. North America and Europe do not merely attract the vast majority of information in English and French. In other languages, as shown by Arabic, Hebrew and Farsi, these areas continue to be well

represented alongside the language's most populous country. The MENA region, by contrast, tends to be massively underrepresented. There are notable exceptions (for instance Iran in Farsi or Israel in Hebrew), but on the whole we see that much is left unsaid about that part of the world. In the following section, we move away from a focus on content and move our attention towards mapping the practices of production.

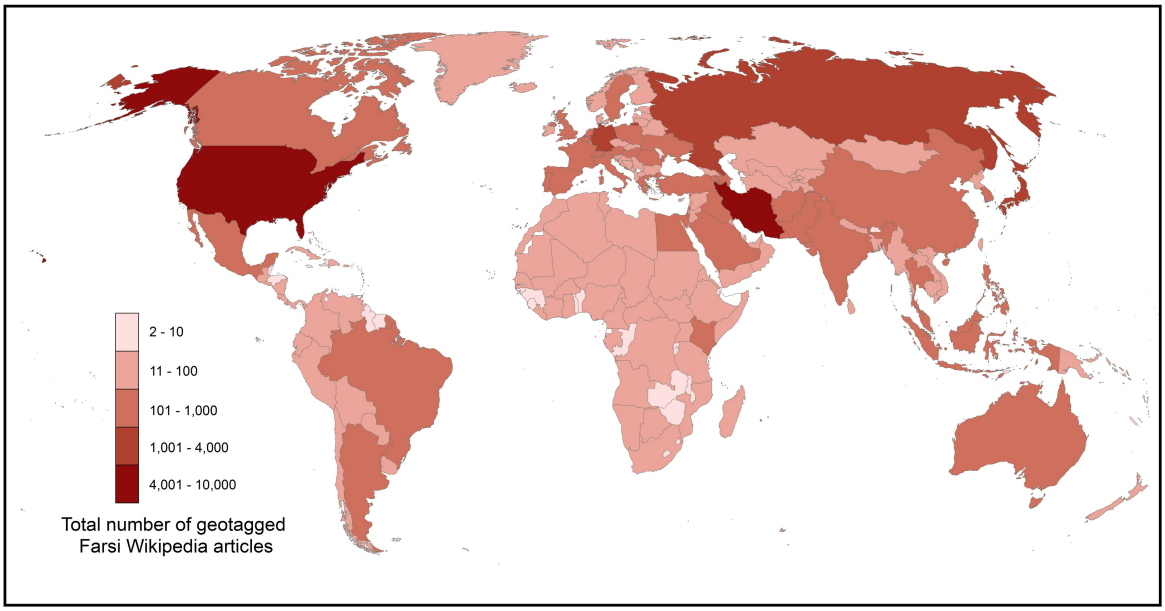

Figure 6.1.2f. Total number of geotagged articles in the Farsi Wikipedia by country.

6.1.3 Edits

One of the distinguishing features of Wikipedia is that anyone with access to the site can, in theory, contribute to its content. This feature combined with its multi-lingual nature means that Wikipedia potentially offers a global democratisation of voice and participation. Because the geographies of published knowledge have traditionally been so spatially clustered and uneven (Graham et al. 2011), the potentials of contemporary media and digital practices to circumvent conventional narratives are more important than ever.

As already highlighted in the methodology paragraph, Wikipedia editors can choose to either contribute to the encyclopedia anonymously or by associating their edits with a registered profile. In this section we offer a global glance on the distribution of Wikipedia edits and editors emphasising the importance and the centrality of the participation issue.

By looking at the geography of edits and editors in Wikipedia, we are able to explore two central lines of inquiry into the geographies of participation in Wikipedia. First, we seek to map and explain the uneven geographies of participation. We do this by mapping edits and editors at the country-level and employing a multivariate regression analysis. Doing so allows us to elucidate the key socio-economic factors that are able to predict the level of participation that we see from any place.

Second, we seek to understand the geographies of local voice within the global networks and practices of participation that constitute Wikipedia. To this end we examine the ratio of *autochthonous* (locally-sourced) to *allochthonous* (non-locally-sourced) contributions to articles

within every country on the globe. This analysis allows us to reveal the parts of the world that have very little voice in the ways that they can get to represent their own environments themselves. We also explore the spatial directionalities of patterns of participation. By mapping both where editors reside, and the locations of articles that they write about, we are able to see distinct geographies of attention and focus.

Table 6.1.3a. Outline of research on participation

Section	Title
6.1.3.1	Geography of Participation
6.1.3.2	Factors of Participation
6.1.3.3	Locality of Participation
6.1.3.4	Autochthonous Vs Allochthonous content
6.1.3.5	Trajectories of attention and focus: Informational Magnetism

6.1.3.1 The Geography of Participation

Having established the ways in which various metrics of participation interact with one another, we examine the geographic distribution of edits committed to all Wikipedias from different countries (using the sampled Wikimedia data pooled from 2007 to 2012). We find especially large numbers of such edits in the USA (33 mil.), Germany (13 mil.), the UK (12 mil.), France (11 mil.), Italy (8.8 mil.), Japan (8 mil.), Russia (7.4 mil.), Spain (6.4 mil.), Canada (5.6 mil.) and Brazil (4 mil.). With the exception of Brazil which ranks 10th, these are all countries in North America, Europe and Asia. From the 20 countries with most edits, only Brazil (10th), Israel (17th) and Mexico (18th) are not in either Europe, Asia, North America or Oceania. Of the 50 countries with most edits all but eight are in Europe, Asia, North America or Oceania. Six (Brazil, Mexico, Argentina, Chile, Colombia and Peru) are in Latin America & Caribbean while two (Israel and Iran) are in the MENA region.

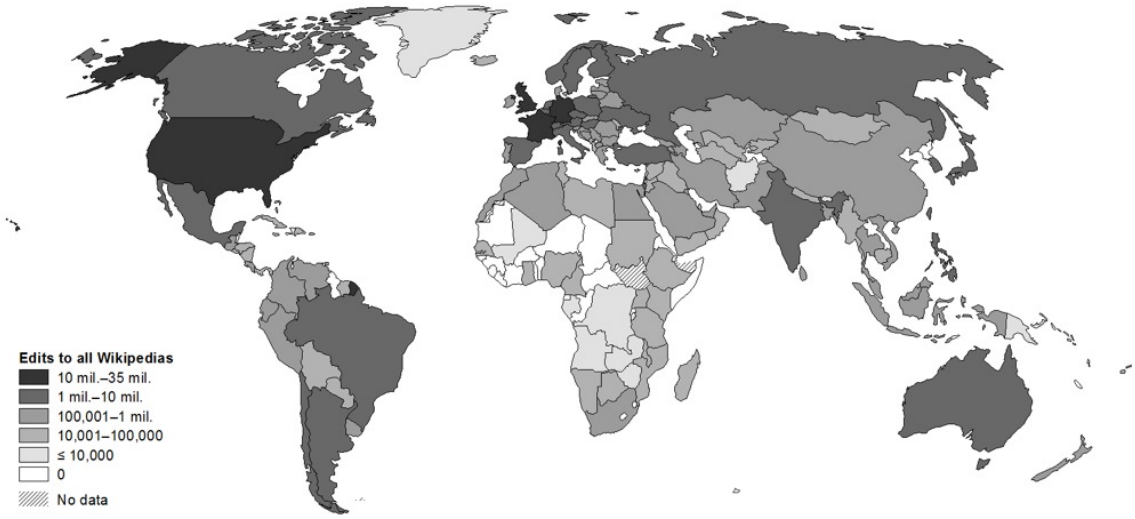

Figure 6.1.3.1a: Distribution of edits to Wikipedia.

Among the 50 countries with fewest edits, there are 22 (44 percent) in Sub-Saharan Africa, 9 (18 percent) in Latin America & Caribbean, 7 (14 percent) in Asia, 6 (12 percent) in Middle East & North Africa and a total of 6 in Oceania, North America and Europe.

The absolute number of edits originating from a country represents the editing activity. Yet, it is also instructive to examine the propensity of people in any country to commit edits to any Wikipedia. We therefore also plot the number of edits normalised by the population of each country (see Figure 6.1.3.1b below). As in the map above, we apply logarithmic intervals in order to account for the strongly skewed distribution of this metric. The map clearly conveys the stark spatial differences in participation to Wikipedia: North America, Europe and much of Oceania stands out strongly against regions with medium participation levels (much of Latin America & Caribbean, Russia, parts of Middle East & North Africa and Asia) and regions with (very) low levels found mostly in Sub-Saharan Africa. These data demonstrate the notion that the Global North is over-represented when it comes to voice and participation in Wikipedia.

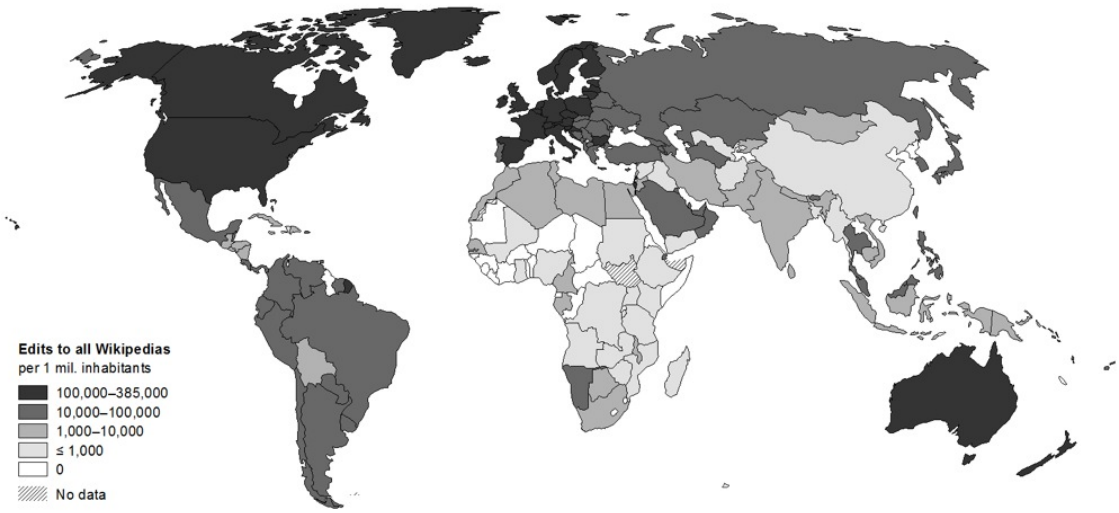

Figure 6.1.3.1b: Distribution of edits to Wikipedia per 1 mil. inhabitants.

The countries in **Table 6.1.3.1a below** have the highest densities of committed edits. They are overwhelmingly in Europe and North America, the only exceptions being Israel, New Zealand, Australia, and Greenland. Sub-Saharan Africa, Latin America & Caribbean and Asia are not listed at all. Recall that this is the edit count across all Wikipedias, not only English.

Table 6.1.3.1a. Countries with high edit activity levels in terms of edits per 1 mil. capita.

Country	Region	Edits p.m.c.	Country	Region	Edits p.m.c.
Norway	EUR	383,644	Germany	EUR	162,461
Estonia	EUR	374,134	Latvia	EUR	159,205
Finland	EUR	335,409	Australia	OCEA	156,432
Iceland	EUR	289,335	Greenland	NOAM	156,060
Sweden	EUR	282,878	Slovenia	EUR	155,993
Luxembourg	EUR	280,150	Ireland	EUR	149,648
Israel	MENA	259,332	Italy	EUR	143,715
Netherlands	EUR	235,769	Czech Republic	EUR	142,195
UK	EUR	186,751	Spain	EUR	137,697
New Zealand	OCEA	185,300	Austria	EUR	128,875
Belgium	EUR	179,744	Lithuania	EUR	112,854
Switzerland	EUR	177,095	Bulgaria	EUR	107,843
France	EUR	165,062	USA	NOAM	106,311
Denmark	EUR	164,379	Croatia	EUR	103,707

Canada	NOAM	163,794	Poland	EUR	102,701

With significant numbers of edits, Japan has the highest editing activity level among countries in Asia (7,992,000 edits total, 66,935 edits per 1 mil. people), Uruguay edits most assiduously in Latin America & Caribbean (255,000 edits total, 77,073 edits per 1 mil. inhabitants) and Namibia in Sub-Saharan Africa (24,750 edits total, 11,524 edits per 1 mil. people).[23] With the exception of Japan, these numbers (neither in their absolute form nor relatively to the population) are no match to the editing activities that are undertaken from within the global cores.

6.1.3.2 Explanatory factors

The findings above lead us to ask why we see such uneven geographies of participation in Wikipedia. By using a series of multivariate regressions we are able to uncover the factors that co-vary with the geography of participation in Wikipedia. Subsequent analysis of the outliers of these models (i.e. countries that do not fit the general patterns that we observe) offers us further insight into the geographies of participation.

In what follows we rely on a set of national-level indicators that we postulate are predictors of participation in Wikipedia. We use the Wikipedia edit data about all language-versions that has been obtained from the Wikimedia Foundation as the dependent variable. The independent (explanatory) variables encompass the following[24]:

- **Population**: Population is a baseline variable that is related to the pool of people that could participate in editing Wikipedia. *Population* has a medium ($r_{Pearson}$ = 0.45) correlation with the number of *edits*. We cannot exclude that *population* may predict positively with the number of *edits* per country.
- **GDP**: Gross Domestic Product (a common measure of the wealth of a nation) can be used as a loose proxy for a range of necessary ingredients for Wikipedia editorship such as leisure time, access to technology, education and informational resources such as libraries. The correlation with the number of *edits* is high with $r_{Pearson}$ = 0.85.
- **GER**: Gross Enrolment Ratio (*GER*) calculates the number of those enrolled in school in relation to a country's total population of 5–17 year olds. GER may thus serve as a proxy for a baseline amount of literacy and schooling necessary to engage with a text-based resource like Wikipedia. The correlation of *GER* with the number of *edits* is $r_{Pearson}$ = 0.59.
- **Broadband Internet connections**: Editing Wikipedia requires an Internet connection. Faster, i.e. broadband, Internet connections likely affect the relationship to user-generated content positively when compared to non-broadband connections. Thus, we suppose that our regression model may benefit strongly from including *Broadband Internet connections* as independent variable, especially due to the strong correlation with number of *edits* ($r_{Pearson}$ = 0.90).

[23] Putting the the lower limit for non-spurious editing activity to 100,000 edits in SSA, South Africa is the most assiduously editing country with a total of 177,000 edits amounting to an average of 3,612 edits per 1 mil. people.

[24] We use log10-values of all variables except GER to account for their skewed distributions.

Table 6.1.3.2a: Correlation between number of edits to all Wikipedias per country and the independent variables (all variables except GER have been logarithmised).

	Edits	Population	GDP	GER	Broadband
Edits	–	0.45	**0.85**	0.59	**0.90**
Population	0.45	–	0.68	-0.12	0.46
GDP	**0.85**	0.68	–	0.47	**0.84**
GER	0.59	-0.12	0.47	–	0.60
Broadband	**0.90**	0.46	**0.84**	0.60	–

Table 6.1.3.2a summarises the Pearson correlations between all independent variables and the dependent variable. Strong correlation between independent variables is a reason for caution due to multicollinearity effects. As such, we will take particular measures to assess model multicollinearity.

In an exploratory analysis we found that in all regressions China is a prominent outlier with considerable leverage on the overall shape of the regression. In China, there is a range of alternative encyclopaedia that successfully compete with Wikipedia.[25] China (together with Panama) is also exceptional in our dataset insofar as it has Wikipedia as lower than the 35th most-visited site (according to Alexa). In what follows we thus decided to exclude China from the modelling attempts as these independent variables are likely to operate in a different manner in China in particular.

We start the analysis by regressing the number of *edits* on the physical variable *population* and the development-related parameters *GDP* and *GER*. This model 1 achieves an adjusted R^2 of 0.77 meaning that the model explains Seventy-seven percent of variation in the number of edits per country. VIF[26] values are between 2.3 and 4.1 meaning that the correlation between the variables is modest but not critical. The independent variables *GDP* and *GER* are significant ($p < 0.001$), but *population* is not ($p = 0.67$).

In model 2, we drop GER and population, while keeping *GDP*. Although *GER* was significant in the first model, when we include *broadband Internet connections* this variable does little explain the dependent variable. Adjusted R^2 increases to 0.86. Despite the high correlation between the two, the VIF values for *broadband Internet connections* and *GDP* are below critical values.

To interpret the relationship between the independent and dependent variables, we plot the predicted values from the model against the independent variables as well as plotting the residuals versus the predicted scores (an RVF plot). In RVF plots, what we want to see is whether the model predicts equally well across the entire range of values, or whether it predicts some regions better than others. If the marginals (i.e. the difference between the predicted value and the actual number

[25] e.g. Hudong Wiki and Baidu Baike, c.f. http://en.wikipedia.org/wiki/Chinese_Wikipedia#Competitors

[26] Variance Inflation Factors (O'Brien 1997). We use VIFs to measure how much the standard error of the regression coefficient estimates are inflated by multicollinearity among the independent variables. Values above 4–5 (sometimes 10) are usually considered critical.

of edits made) are very high for some countries, such as those with few edits and very low for countries with many edits, we would say that the model works better for very active countries.

Fig6.1.3.2a (top) plots residuals versus fitted values of model 2 along with a LOESS curve indicating (curvi)linearity. From the plot we can observe that the model's residuals follow a curved (decreasing-increasing) shape. The plots against *broadband Internet connections* and *GDP* (**Fig6.1.3.2b** and **Fig6.1.3.2c** center and bottom) both exhibit a curvilinear shape. The shape is more curved for *broadband Internet connections*, however, the increase tapers off at the upper end of the distribution. Hence, countries with very few and very many broadband Internet connections commit more edits to Wikipedias than one would expect given this model. On average, countries with medium numbers of broadband Internet connections commit fewer edits than predicted in the model. These countries are mostly located in MENA, SSA, and LACA. In fact, almost all countries in MENA and most countries in SSA commit below-expectation numbers of edits.

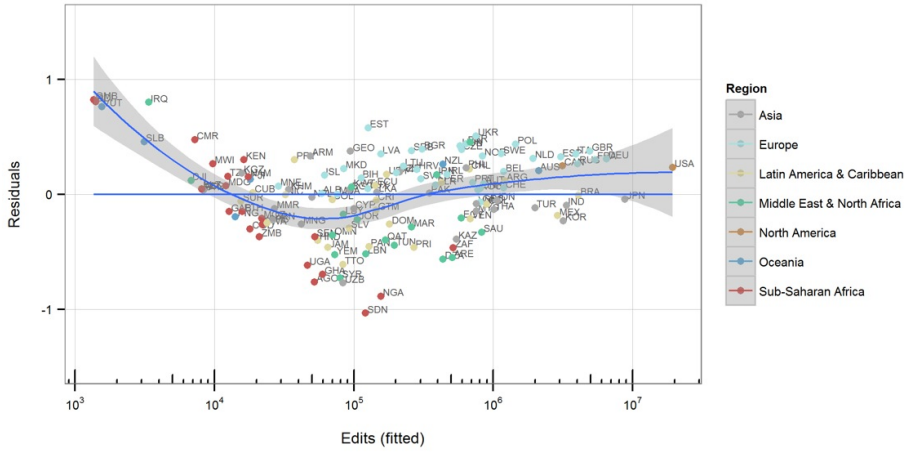

Fig6.1.3.2a. Plot of residuals by edits. Countries are highlighted by continent.

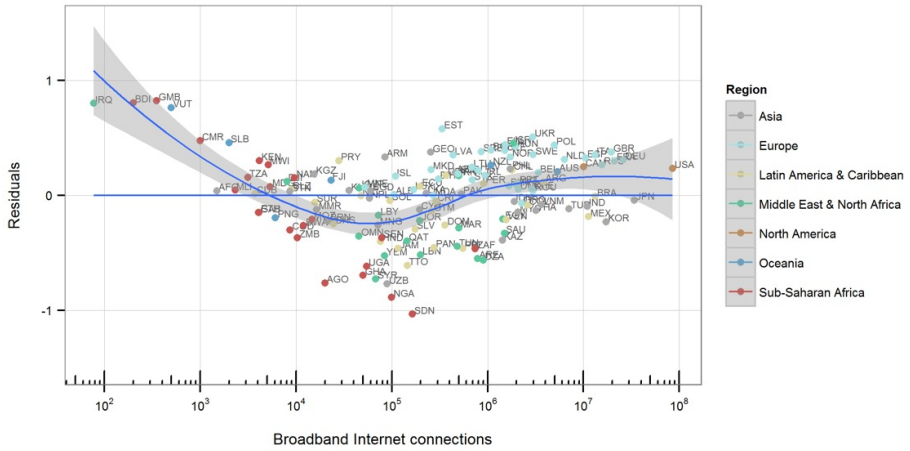

Fig6.1.3.2b. Plot of residuals by number of broadband subscribers. Countries are highlighted by continent.

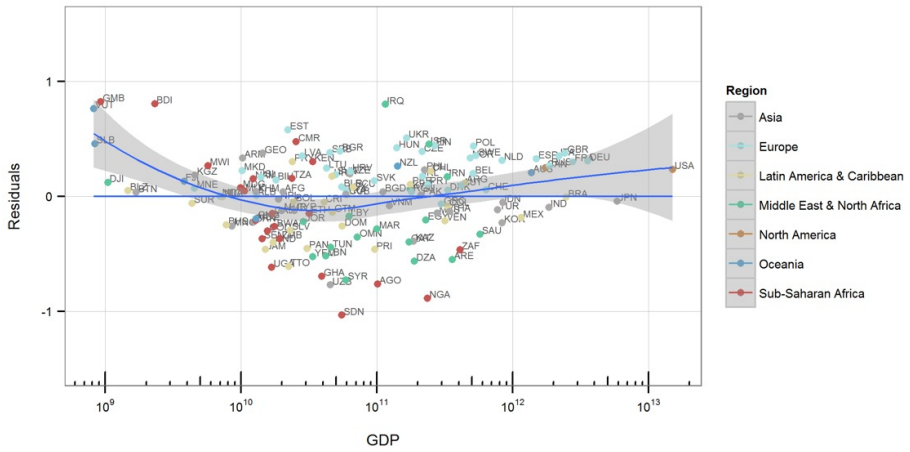

Fig6.1.3.2c. Plot of residuals by GDP . Countries are highlighted by continent.

Attempting to correct for this curvilinear effect yields the simple final model summarized in Table 6.1.3.2b and elaborated in Table 6.1.3.2c.

Table 6.1.3.2b. Regression models (bb denotes broadband Internet connections).

No.	Model	Adj. R^2
1	$log10(\text{edits}) = \beta_0 + \beta_1 log\ 10(\text{Population}) + \beta_2 log\ 10(\text{GDP}) + \beta_3 \text{GER} + e$	0.7719
2	$log10(\text{edits}) = \beta_0 + \beta_1 log\ 10(\text{GDP}) + \beta_2 log\ 10(\text{bb}) + e$	0.8577
3	$log10(\text{edits}) = \beta_0 + \beta_1 log\ 10(\text{bb}) + \beta_2 log\ 10(\text{bb})^2 + e$	0.9049

We attempt to correct for this curvilinear effect by including a squared *broadband Internet* term. Due to the tapering off of the curvilinear effect at the upper end of the distribution such a model will overestimate the number of edits of countries in that region, e.g. of the USA. However, we expect the overall quality of the model to improve.

Table **6.1.3.2c**: Regression statistics of model 3.

	Est.	Std. error	t	p-value
Intercept	4.9195	0.3506	14.031	< 0.001
log10(bb)	-0.7470	0.1391	-5.372	< 0.001
$log10(bb)^2$	0.1418	0.0134	10.556	< 0.001
Adjusted R^2	0.9049			

The simple final model 3 in **Table** regressing the number of edits onto a linear and a squared term of *broadband Internet connections*[27] achieves an R^2 of 0.9 with $p < 0.001$ for both variables (Table). The relative importance (Kruskal 1987) of the simple and the quadratic term are 47 percent and 53 percent, respectively.[28] Figures 6.3.1.3c-e show the plots of the residuals, which are now more linear than before the introduction of the quadratic term.

In the final model, the effect of number of broadband connections is negative, but it is important to remember that broadband connections squared is positive. In a case such as this one, it is not appropriate to simply look at broadband, or broadband squared, but to calculate the value of the dependent variable (number of edits) using both regular and squared values at the same time. What the model shows is that the number of edits to Wikipedia tracks extremely well on to the number of broadband connections, but the effect is somewhat curved. Countries with extremely little connectivity edit more than expected (given that the few connections that exist are probably in universities and among an educated elite) but as new connections are added it takes some time before these connections lead to a proportionate increase in the number of edits made. At some point, there is a 'critical mass' of editors which is associated again with editing more than average.

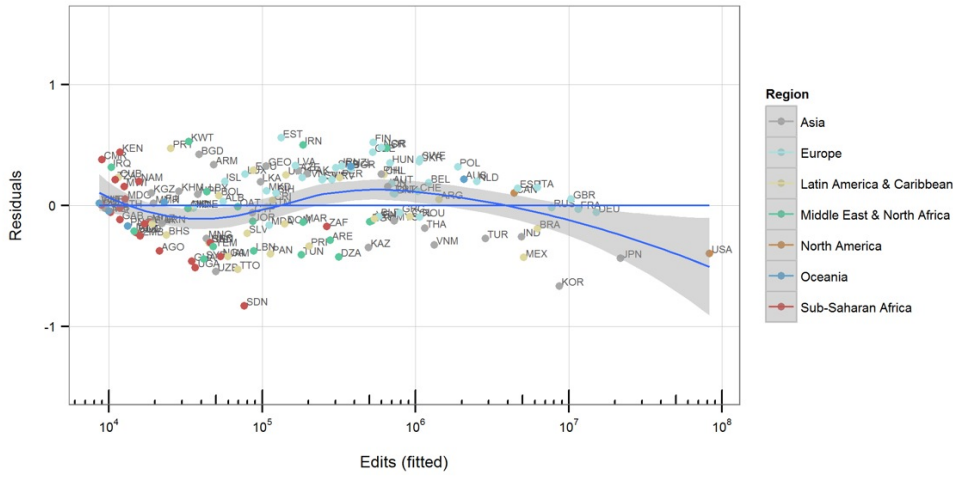

Figure 6.1.3.1d. Plots of residuals vs. fitted values using the preferred model. Note that the curve in the LOESS curve is distinctly more modest than in the previous model suggesting a better fit overall. China is excluded.

[27] By including also *GDP* does not improve this model (*GDP*: $p = 0.19$, VIF of 4.5 and only marginal increase of R^2 to 0.9054).

[28] The same model with China included achieves an adjusted R^2 of 0.87 and $p < 0.01$ for log10(*bb*).

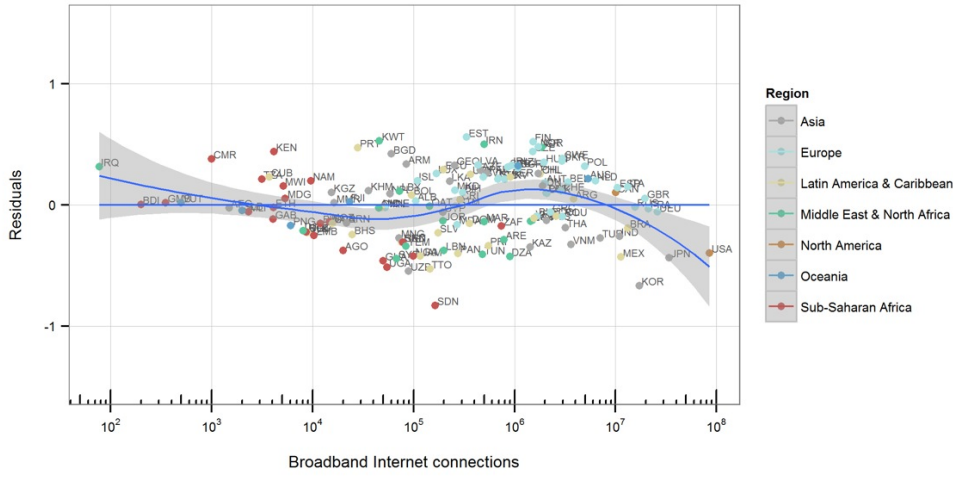

Figure 6.1.3.1e. Plots of residuals vs. number of broadband internet connections using the preferred model (Model 3). Note that the curve in the LOESS curve is distinctly more modest than in the previous model suggesting a better fit overall. China is excluded.

6.1.3.3 Locality of participation

The previous section demonstrated that the geographies of participation are highly uneven, and that availability of broadband connectivity is a central predictor of this spatial unevenness. Yet, it remains that, within those uneven geographies of participation, we know little about the specific geographies of voice. In other words, whilst we now know about who is participating, we know little about who is writing about who and the directionalities of focus.

By obtaining three types of geographic information about Wikipedia (locations of geocoded articles and origin locations of anonymous and registered edits to geocoded articles), we are ultimately able to analyse two important locations attached to each edit: the source location and the target location. Note that we know about the strength of these relationships only for the English-language version of Wikipedia where we are able to geocode the locations of not only the anonymous but also the registered editors.

With these we are able to measure how many edits (anonymous or registered) have originated in a country and been directed into the same and any another country. Thus, we investigate the geographies of local voice in two ways: first, by looking at autochthonous (locally-sourced) to allochthonous (non-locally-sourced) contributions in every region of the world and every country, and second, by looking at trajectories or networks of editing over space.

6.1.3.4. Autochthonous versus allochthonous content

Our first step is to ascertain the volume of autochthonous content by region. We do this here by analyzing the within-region-edits (or self-edits) that happen in every region (i.e. edits from a region to articles within the region). By visualizing the raw number of within-region-edits by both registered and anonymous editors, Figure 6.1.3.3a reveals a few important findings. First, in terms of raw numbers, North America, Europe and (to a somewhat lesser extent) Asia drastically outnumber the remaining world regions. Second, Latin America & Caribbean, Middle East & North Africa, and Sub-Saharan Africa all commit only a very small number of within-region-edits. Thus,

even a relatively small number of edits flowing into those regions from outside (i.e. allochthonous contributions) could easily drown out local voice.

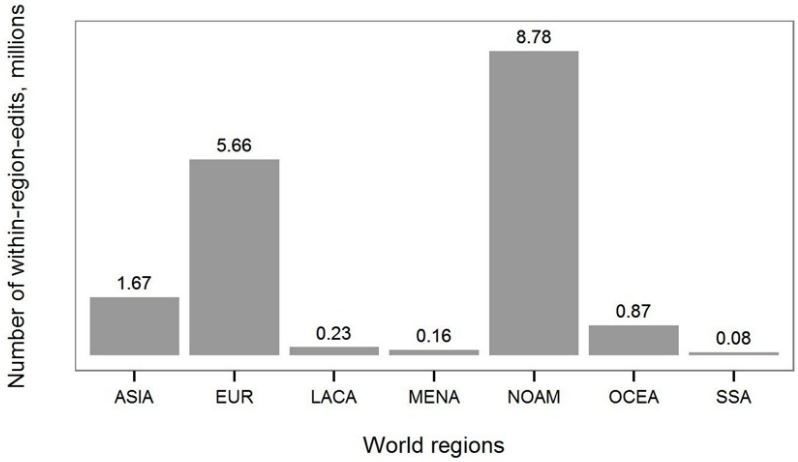

Figure 6.1.3.3a. Absolute number of within-region-edits (both anonymous and registered) to geolocated articles in the English-language Wikipedia, by world region.

Locality of participation and voice is best investigated not just using raw number of edits, but the proportion of within-region-edits. We can examine these proportions in two ways: the proportion of a region's committed edits that stay within that region ["committed edits"] and the proportion of edits to articles in the region that come from that region ["received edits"].

Figure 6.1.3.3b plots these two kinds of data for the seven world regions. The figure places each region according to its proportions for both edit types combined symbolised with a dot. Computing the proportions of only anonymous edits or only registered edits, the characteristics of the regions change slightly.[29]

[29] The overall configuration of the dots (representing both edit types combined) in **Figure 6.1.3.3b** comes with some uncertainty. Keeping in mind that roughly a third of the registered edits and virtually all of the anonymous edits could be geocoded (see **Error! Reference source not found.** in ***Error! Reference source not found.***), the true position of the world regions in **Figure 6.1.3.3b** is likely slightly closer to the × symbols that signify proportions of registered edits. However, the general patterns and insights stay the same.

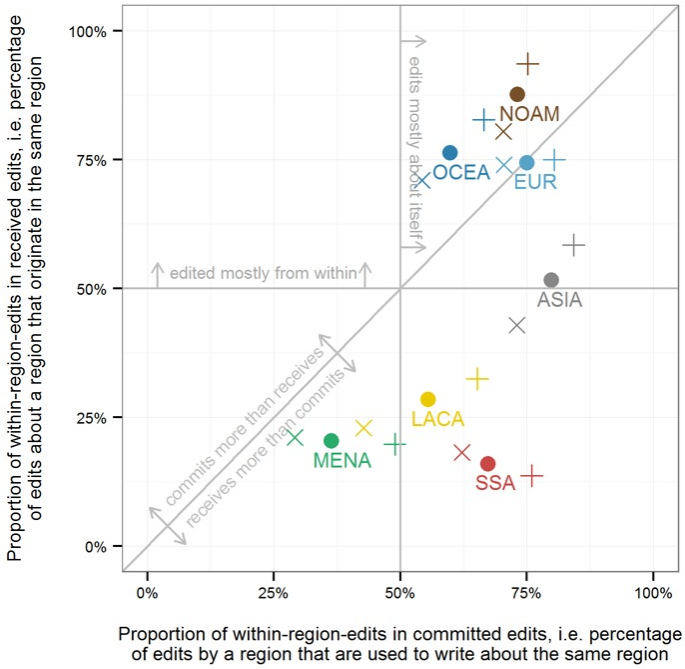

Figure 6.1.3.3b: *Proportion of within-region-edits in total committed edits (horizontal axis) and in total received edits (vertical axis) to geocoded articles, per world region. The data shows anonymous edits only (+), registered edits only (×) and both edit types combined (●).*

On the vertical axis of **Figure 6.1.3.3b** we can see a clear division between regions that are largely able to define themselves and regions that are largely defined by others. The world regions separate into two distinct groups of three (with Asia in the middle): Sub-Saharan Africa, Middle East & North Africa, Latin America & Caribbean receive comparatively few edits from within their territories (around 25 percent). Europe, Oceania and North America on the other hand receive primarily edits from within (around 75 percent). Asia is edited from within and from outside to almost equal degrees.

Asia, Europe and North America all target 73–80 percent of their committed edits stay within their own region. Interestingly, Sub-Saharan Africa commit just slightly less within-region edits at 67 percent. Oceania, Latin America & Caribbean and especially Middle East & North Africa fall behind with 36–60 percent. Overall, the world regions are grouped in more distinct clusters in terms of locality of received edits (the vertical axis) than committed edits (horizontal axis).

It also becomes apparent that only North America and Oceania commit proportionally more edits locally than they receive from elsewhere. For Europe the two proportions are roughly equal. For other regions the ratios are much higher, meaning that the number of edits that a region receives from elsewhere are much higher then the number the region receives from local editors. This ratio is 1.9:1 in Latin America & Caribbean, 1.8:1 in Middle East & North Africa and 4.2 in Sub-Saharan Africa.

What do these findings reveal?

- Even when editors from Sub-Saharan Africa spend most of their edits within region, their small numbers mean that most content still comes from elsewhere.
- The global cores of North America and Europe self-represent very effectively by focusing on their own regions.
- Content appears to be very sensitive to feedback loops. A lot of content on an area in one language leads to more content in other languages as translations rather than similar local content.

The global cores focus their editing primarily within their own territories. Large amounts of geospatial content show no sign of deterring people from further contributions and editing: as more content exists, so too do more articles to amend, augment, update and build upon. It is possible that a stock of good content may be an attractive "editing ground" for Wikipedians, whereas a scarcity of content, beneath a certain unknown threshold, may – somewhat paradoxically – demotivate people to fill in the blanks. A relative lack of content may further reinforce perceptions amongst editors that little content equates to a small audience that is not worth writing for.

Comparing anonymous and registered editing in **Figure 6.1.3.3b** (symbolised by + and ×, respectively) reveals that in all regions a larger share of anonymous edits is about the same region, i.e. "local", than registered edits. This means, registered editors tend to edit less locally than anonymous editors (assessed at the level of world regions).

What could be the reasons for this pattern? – For one, registered editors are likely more senior users of the Wikipedia platform. Then, locality of content may serve as a starting point in the Wikipedia environment, i.e. people begin their editing career for example by correcting a small mistake in the article about their hometown (certainly, in topic or thematic space, one would expect that people most likely start editing topics they are intimately familiar with). Having covered their more immediate environment and wishing to continue to contribute to Wikipedia enough to register, more experienced editors may eventually venture farther in their editing activities. We concede, however, that various confounding variables may also be at play: Maybe people who have registered with Wikipedia belong to a specific group that is more international or travels more (than the average editor) and thus feel confident to write about places in other world regions. Or, being savvier Internet citizens on average, registered editors may feel more assertive to write about places that are not in their immediate environments.

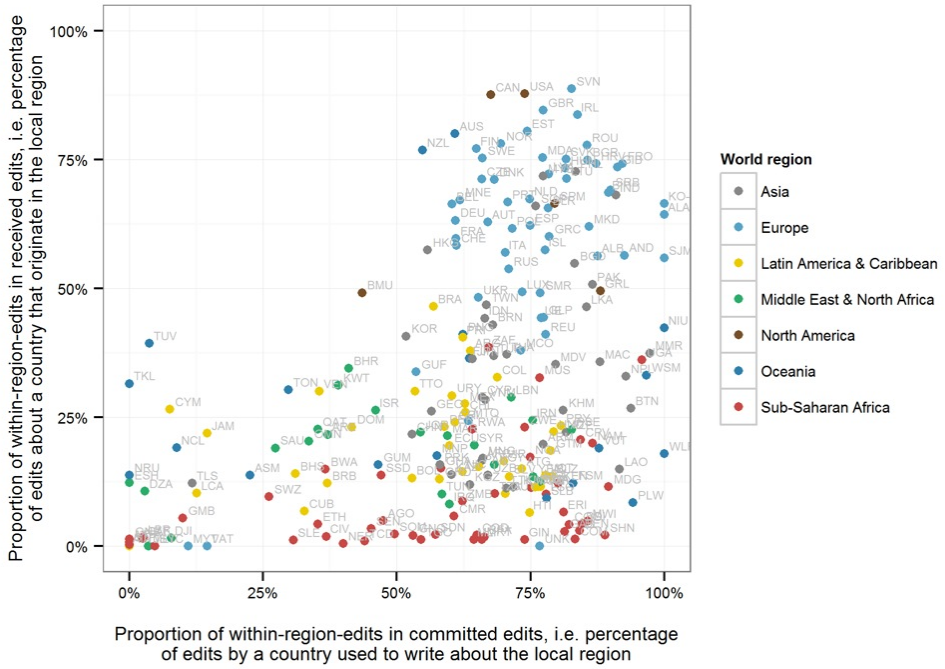

Figure 6.1.3.3e: Proportion of within-region-edits in total committed edits (horizontal axis) and in total received edits (vertical axis) to geocoded articles, per country.

While Figure 6.1.3.3b gives a good overview of core and peripheral world regions in terms of their broad patterns of editing, it lumps together countries that may find themselves on quite different trajectories. Figure 6.1.3.3c thus portrays essentially the same data as Figure 6.1.3.3b but resolves world regions into their constituting countries while still looking at within-*region*-edits. Figure 6.1.3.3c thus betrays a substantial amount of variability at the national level. While the general insights gained Figure 6.1.3.3b remain valid, Figure 6.1.3.3c goes to show that individual countries are characterised by significantly different ways in which they both participate and are defined by people in far-away places. Countries within Sub-Saharan Africa, Latin America & the Caribbean and the Middle East & North Africa can vary substantially, primarily regarding their inside/outside-focus of editing (horizontal axis). For instance, editors in some countries in Sub-Saharan Africa (e.g. Uganda, Madagascar, Namibia, Malawi) are much more likely to edit about regional topics than editors from other countries in the region (e.g. Gambia and Swaziland). Conversely, North America and Europe (with few and usually relatively small exceptions) show a more coherent pattern. With regard to Sub-Saharan Africa it is also important to note that most countries receive a very small fraction of edits from within the region, the 16 percent of within-region edits received that we saw in Figure 6.1.3.3b are primarily due to relatively high numbers for South Africa, Uganda, Mauritius, Rwanda and Zimbabwe. In other words Sub-Saharan Africa's figures would be even lower without the outlier-effects of those five countries.

Figure 6.1.3.3d plots within-country-edits on the axes rather than within-region-edits as in Figure 6.1.3.3e The most striking change between this figure and Figure 6.1.3.3c is certainly Europe that is (mostly) quite drastically lowered and dispersed on both axes. This difference hints at European countries substantially editing about each other, across national boundaries, i.e. most edits by

these countries happen within Europe, less within the respective country. Overall, both Figure 6.1.3.3c and Figure 6.1.3.3d there is considerable overlap between the various world regions, hinting that, while the overall configuration conveyed in Figure 6.1.3.3b is correct, it abstracts the more nuanced patterns some countries experience.

Explanatory factors for participation thus attests to the fact that broad and quite generalised Digital Divide narratives are too schematic to adequately capture the variety of trajectories among countries regarding their participation to Wikipedia.

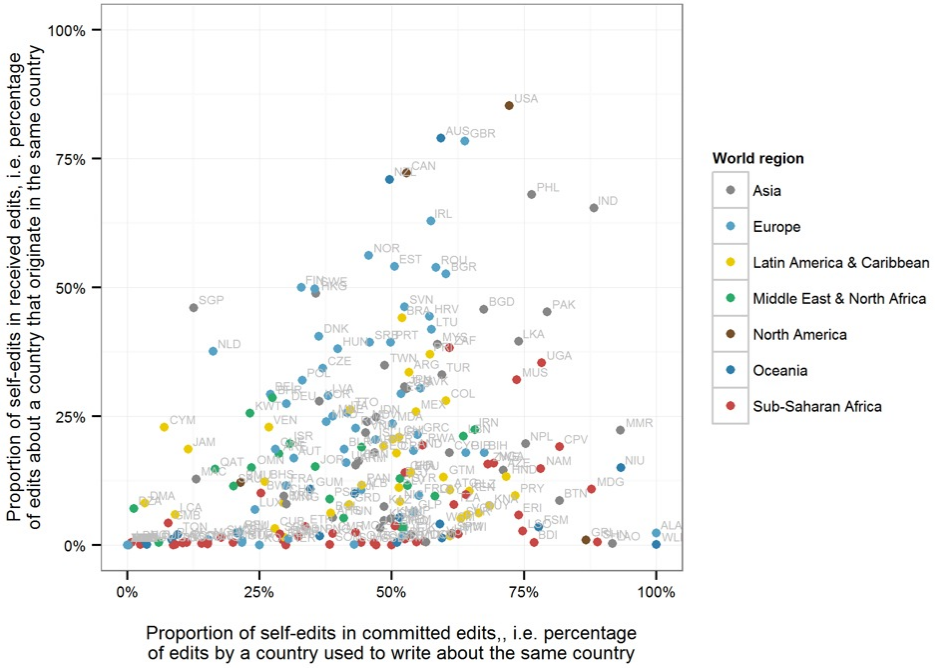

Figure 6.1.3.3d: Proportion of self-edits (within-country-edits) in total committed edits (horizontal axis) and in total received edits (vertical axis) to geocoded articles, per country.

6.1.3.5 Trajectories of attention and focus: Informational Magnetism

In the previous section we employed a simplified distinction of within region or outside region when considering edits. However, it is possible to also look at which regions send edits to which other regions. Thus, instead of a univariate map or a bivariate scatter plot, we can represent the data as a network of flows.

We present two such flows of editing herein. The first shows the data where edge thickness is proportionate to the total number of edits received (Figure 6.1.3.5a) and the second is proportionate to the total number of edits sent (Figure 6.1.3.5b). That is, if a region receives most of its edits from North America then the edge from North America to that region will be thick in the first one. If a region sends most of its edits to North America (even if it sends very few edits) then in the second that edge will be thick.

From this graph, we can consider the MENA among the "net-importing" regions (Asia, Europe, Latin America & Caribbean, Middle East & North Africa, and Sub-Saharan Africa) and "net-exporting" regions (North America, Oceania) in terms of edits. For example, Sub-Saharan Africa receives 10.7 more edits from outside than it commits to the outside. This proportion is much more drastic than for the rest of the regions for which it varies between 3.7 (Asia) and 1.02 (Europe) or for the net-exporting regions even 0.46 (Oceania) and 0.38 (North America).

With the number of edits received, we can again the same pattern of most edits coming from within the region itself, while North America and to a lesser extent Europe sent most of the edits elsewhere. Asia and Oceania send remarkably few edits to the MENA region given that they are relatively active on Wikipedia. The overall configuration clearly reflects the fact that we analyse the English Wikipedia: predominantly English-speaking regions produce more content than they receive. Europe would likely fare worse in terms of number of edits committed if UK and Ireland were omitted from that group.

With the graph normalized by the number of edits sent, we see a striking finding from the MENA region. Not only do less than half of edits in the MENA region stay within the MENA region, but a substantial share go to North America, even if their impact on the overall number of edits within North America is rather minimal. In fact, proportionate to the total number of edits sent from a region, the MENA send more edits to North America than any other regions sends to another.

This leads to what might be considered a 'double whammy' for the MENA region. Not only is this region predominantly written by others, with MENA countries in general writing less than expected, when editors from the MENA region do write they are more likely to send their edits to the rest of the world than to stay within MENA.

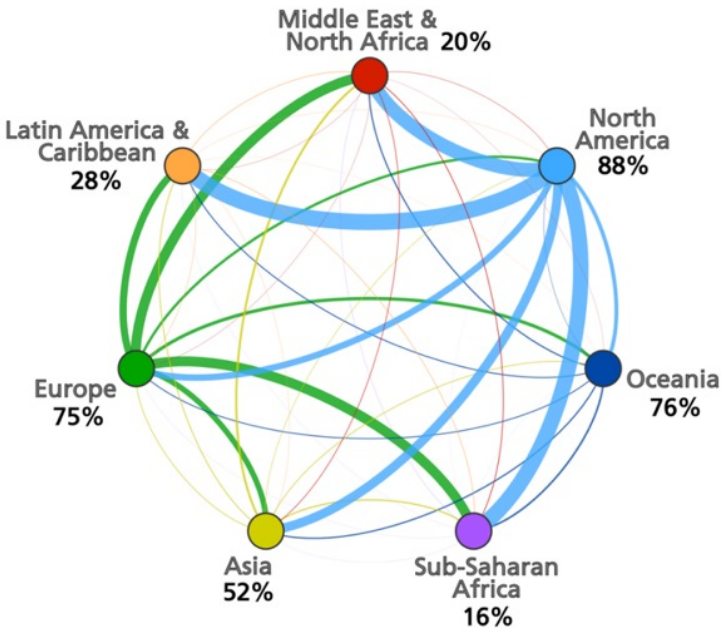

Figure 6.1.3.5a. Network of edits between world regions, normalised for each target region. The edges are coloured according to the source region. Percentages denote self-edits (not depicted).

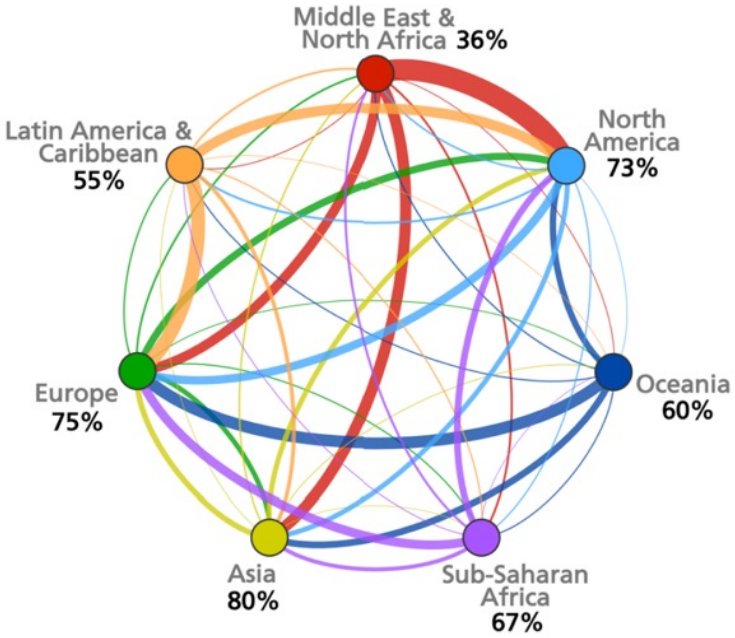

Figure 6.1.3.5b. Network of edits between world regions, normalised for each source region. The edges are coloured according to the source region. Percentages denote self-edits (not depicted).

Summary

The section has revealed novel insights into global geographies of participation. The relative democratisation of the Internet has not brought about a concurrent democratisation of voice and participation. Even on Wikipedia, which is widely touted as one of the Web's most open, most used, and most inclusive platforms, we see highly uneven geographies of participation. Given how strongly these patterns track with broadband, it is not really the fault of Wikipedia for being exclusionary (either intentionally or not). Rather, the existing dominance of Western nations on the Internet manifests itself in Wikipedia in raw numbers.

Some countries can account for the majority of the world's participation. 57.6% of all edits to Wikipedia come from just five countries (Italy, France, UK, Germany, and the US). We also see shocking comparisons like the fact that there are more Wikipedia editors from Australia than all of Africa combined, or more articles about Antartica than any country in Africa.

The goal of highlighting these inequalities is not to suggest that they are insurmountable. Our regression shows that the availability of broadband is a clear determinant of the propensity of people to participate on Wikipedia. However, the relationship is not a linear one. Countries with modest broadband penetration (i.e. 10k to 100k subscribers) tend to underperform, but over that level countries begin to have a critical mass of editors. Because the relationships between broadband access and participation are non-linear, countries with high broadband penetration tend to have exponentially louder voices on Wikipedia than countries with low penetration rates. This inequality in voice manifests itself in the stark differences between world regions outlined above.

Because of the nature of Wikipedia, relative levels of participation are not always as important as absolute levels. Said differently, countries that are home to large blocks of editors have the ability to dominate the production of knowledge about smaller countries. However, as we have shown in some research outside this report (particularly the paper uneven openness), the presence of local content tracks very strongly with local editing populations. So while most of the work might be done from outside, the simplest form of visibility – the creation of the article, tends to be a much more local affair.

Finally, we see a noteworthy practice among the MENA region. Even though this region has a comparatively small number of both articles and editors, these editors remain focused on fleshing out articles within the global core more so than in their own region. The many reasons for this practice (such as starting on pre-existing articles, being attracted to controversy, etc…) illustrate the powerful effects of informational magnetism that appears to be difficult to break free of.

6.2 Wikipedia articles in MENA countries

The previous section focused on positioning the activity of the MENA region in relation to the global pattern of edits. There was an unsurprising lack of content in the region, but a surprising set of findings: given the level of broadband penetration there is a lower than expected level of editing, first, a focus on editing more in North America than in the local region, and a relative lack of interest from areas other than North America. In the following sections we focus more specifically on the MENA region itself, first looking at the number of articles across various sociodemographic factors, across key language groups and in terms of national editing and identity. This way, we can look at differences within the MENA region rather than merely differences between the MENA region and elsewhere.

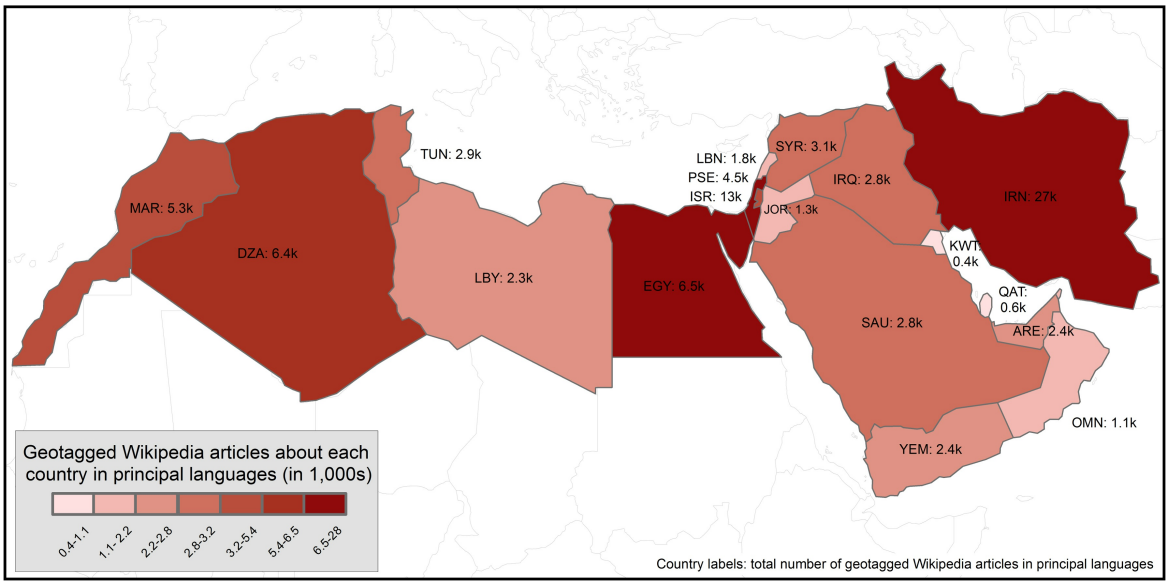

Figure 6.2a: Distribution of Wikipedia articles across principal languages by country within MENA.

Out of a total of 88,813 Wikipedia articles in all considered languages for all MENA countries, Iran and Israel represent almost half of this total (46%). The map (Figure 6.2a) shows this distribution. If we were to ignore those two outliers the presence of information about the region appears even more sparse.

To help explain the variation in this region and bring some perspective to the distribution of articles, in this section we explore the significant independent variables from the previous section. Each one is accorded its own subsection.

Section	Metric
6.2.1	Wikipedia articles and Population

6.2.1 Wikipedia articles and Population

Population, in absolute terms, appears to be able to explain some of the distribution of articles in the region.

The correlation existing between the number of Wikipedia articles and the population of each country is statistically significant (*rPearson=0.64*), even if not especially strong.

Figure **6.2.1a** depicts the distribution of Wikipedia articles by population for each country. The two countries that stand out here are Israel and the Palestinian Territories. In spite of their small territorial dimensions and of a population count that is in both the cases strongly under the MENA average, these two countries present over 1.7 articles for 1000 inhabitants.

We also see a long tail of countries, which bring down the regional average to 0.42 articles per every 100 people. Iraq and Egypt have the lowest values with 0.09 and 0.07 articles, respectively, per 1000 inhabitants.

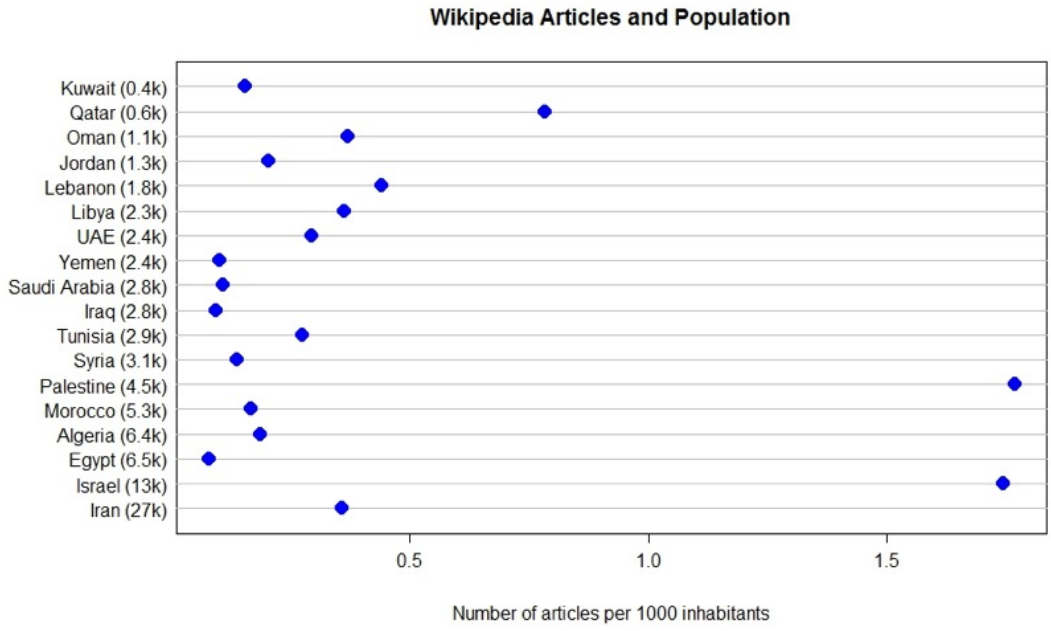

Figure 6.2.1a: Number of Wikipedia articles per 1000 inhabitants in MENA countries.

6.2.2 Wikipedia articles and GDP

Compared to population, Gross Domestic Production (GDP) doesn't appear to be able to explain the distribution of Wikipedia articles in the MENA region. Even though a positive correlation exists, it is not too strong (*rPearson=0.34*).

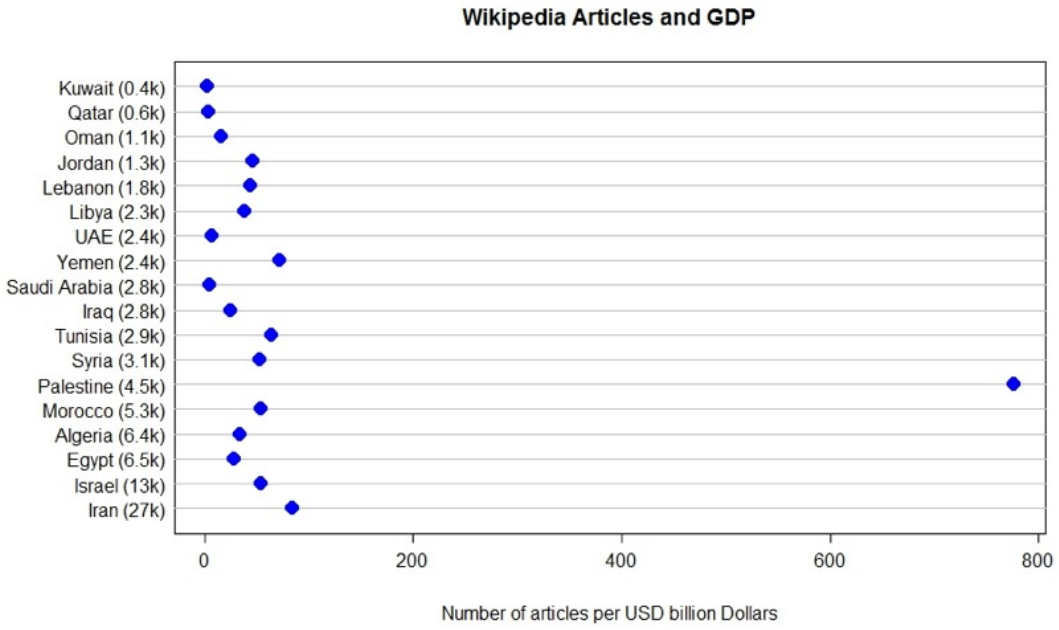

Figure 6.2.2a: Number of Wikipedia articles per US Billion dollars GDP.

The graph (Figure 6.2.2a) shows better this pattern expressing the number of Wikipedia articles per USD million dollars. It is evident that the Palestinian territories behave in this distribution as an outlier, but this condition doesn't significantly influence the correlation value that, without the Palestinian Territories, is almost unchanged ($r_{Pearson}=0.35$).

Behind the Palestinian Territories, which has 775 Wikipedia articles per every million dollars of GDP, the next largest country is Iran with 84 articles per billion dollars. All others Mena countries then sit below the regional average of 78 Wikipedia articles per billion dollars (without the Palestinian value, the average drastically decreases to 38 Wikipedia articles per billion dollars).

Interestingly, the countries at the bottom of the distribution are the wealthy oil and gas exporters of Kuwait with 2.3 articles, Qatar with 3.8 articles and Saudi Arabia with 4.9 Wikipedia articles per billion dollars. Just above them we find the UAE with 6.74 articles, Oman with 15.6 articles, Iraq with 24.8 articles, Egypt with 28.5 articles and Algeria with 34.3 Wikipedia articles per billion dollars.

Countries with relatively higher values include Libya with 38.3 articles, Lebanon with 43.3 articles, Jordan with 46.2 articles, Syria with 52.8 articles, Israel and Morocco with 53.6 and 53.8 articles, Tunisia with 64 articles and Yemen with over 72 Wikipedia articles per billion dollars.

In sum, we see that it isn't wealth that is able to help us explain the geography of articles in the MENA region. This is quite remarkable in that GDP is a relatively reliable predictor in the world overall.

6.2.3 Wikipedia articles and Internet users

The final variable considered is also the one that has the strongest correlation with the number of Wikipedia articles. The value of the correlation between the number of Wikipedia articles and the number of Internet users is very strong ($r_{Pearson}=0.81$) and (as we described earlier in the report) is likely to be a causal factor (or at least a necessary precondition). However, the variation within this relationship is also important to pay attention to.

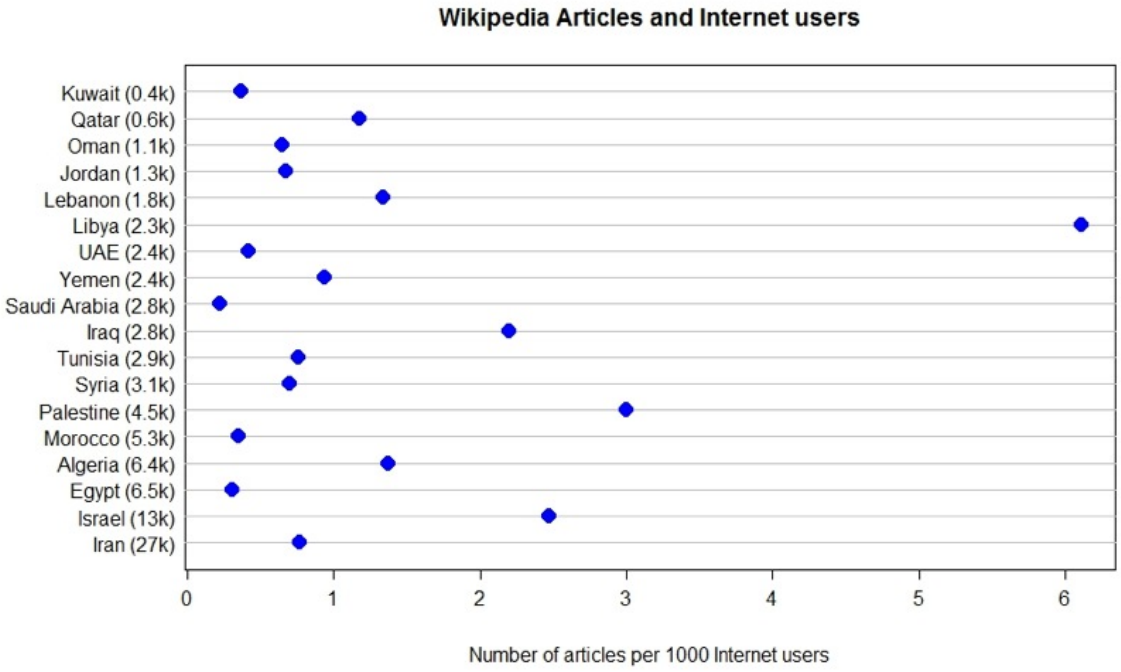

Figure 6.2.3a: Number of Wikipedia articles per 1000 internet users.

The outlier in this case is represented by Libya that with more than 6 Wikipedia articles per 1000 Internet users presents the best value among the Mena countries of this distribution.

Considering that the average value is 1.32 articles per 1000 Internet users, there are very few countries able to exceed this value. If we were to omit the value of the outlier Libya, we would only have Jordan with 1.33 articles, Algeria with 1.37 articles, Iraq with 2.19 articles and the Palestinian Territories with 3 Wikipedia articles per 1000 Internet users that show higher than average values.

The lowest values, in contrast, are Qatar with counts 1.17 articles, Yemen with 0.93, Tunisia with 0.76, Syria with 0.70, Jordan with 0.67, Oman with 0.64, the United Arab Emirates with 0.41, Kuwait with 0.36, Morocco with 0.34 and Saudi Arabia with only 0.22 Wikipedia articles per 1000 Internet users.

While we can say with confidence that local broadband subscribers is a globally important variable, this variation between MENA countries points to the fact that it is not a perfect correlation and local factors play a significant part in influencing this variation.

6.3 Linguistic representation in MENA countries

The total number of articles gives us some information about the capacity of places to be represented on Wikipedia. But looking at the relative amount of content in different languages gives us a sense of the varying amount of focus given to each place from different geolinguistic communities. Here we describe the variation at multiple scales for several languages of interest. First at the country level (Section 6.3.1), then the provincial level (Section 6.3.2) and finally, normalized by area (Section 6.3.3).

Through these analyses we focus on similar questions as above but now with specific linguistic and geographic details. These details will be relevant for both targeting particularly underdeveloped areas and particularly underdeveloped language groups.

6.3.1 Underrepresented languages - a country-level analysis

In this section we highlight which languages have the most coverage of a country as a percentage of all principal languages. Thus, if a country has 32 percent of articles in English, it means that 68 percent of articles about that country are in another language. Since we are analysing 42 principal languages (having excluded 2 synthetic languages), in a "perfect world" the percentage should be 2.4 for all languages – that is, the articles are equally distributed across all languages. However, we know that is not the case. Articles are not translated into all languages. The MENA region is virtually absent in Polish or Swedish, for example.

Iran and Yemen are the countries with the highest percentage of articles about the country in English. In Yemen there are a total of 977 (44% of the total) English Wikipedia articles, whilst Iran has 13,456 (48% of the total) Wikipedia articles in English. This strong presence could be explained in two ways. There is clearly a varying capacity of these countries (as there is in all countries in the Mena region) to participate in Wikipedia. However, some countries are also more likely to attract attention from outside their borders. In the case of Iran, as will be noted in Section 6.4, it is not the case that articles are solely being written by outsiders. The Iranians have a very strong local editing presence in Wikipedia. Whether this is due to a focus on self-determination, state-sponsored editing or merely a highly connected population is still somewhat ambiguous. It is likely a combination of all three. Yemen on the other hand is almost entirely written about by outsiders. Presently there is a high degree of interest in Yemen due to its presumed position in wider geographic and international conflicts.

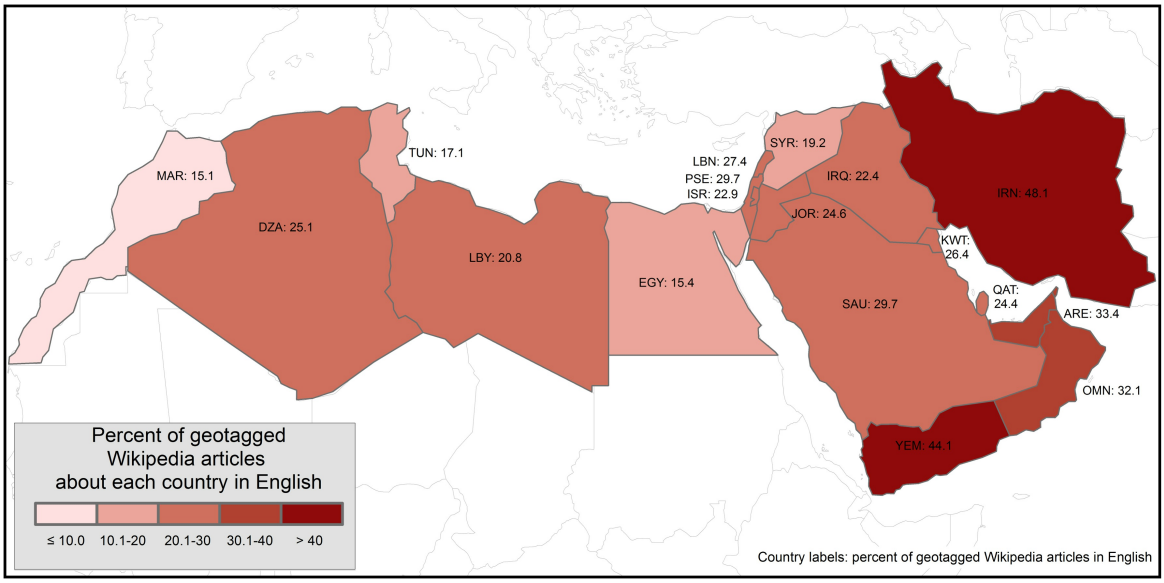

Figure 6.3.1a: The percentage of articles about a country in the MENA region that are written in English.

The average percent of articles that occur in the English Wikipedia is 25.2%. A few other countries (mostly in the Gulf) are above this average. Countries with lowest counts are Morocco and Egypt, respectively with 15.2 and 15.5 per cent of the total number of Wikipedia articles across all principal languages. In both of these cases, it is because the have a more even distribution of articles across all languages. That is, both Morroco and Egypt have more than average in French and Arabic, and also relatively high counts in other languages such as German, Spanish and Italian.

Out of 18 Mena countries, 14 have English as their dominant language in Wikipedia articles. The only exceptions are Morocco, Egypt, and Tunisia that have most articles in French, and Syria, which has most articles in Arabic.

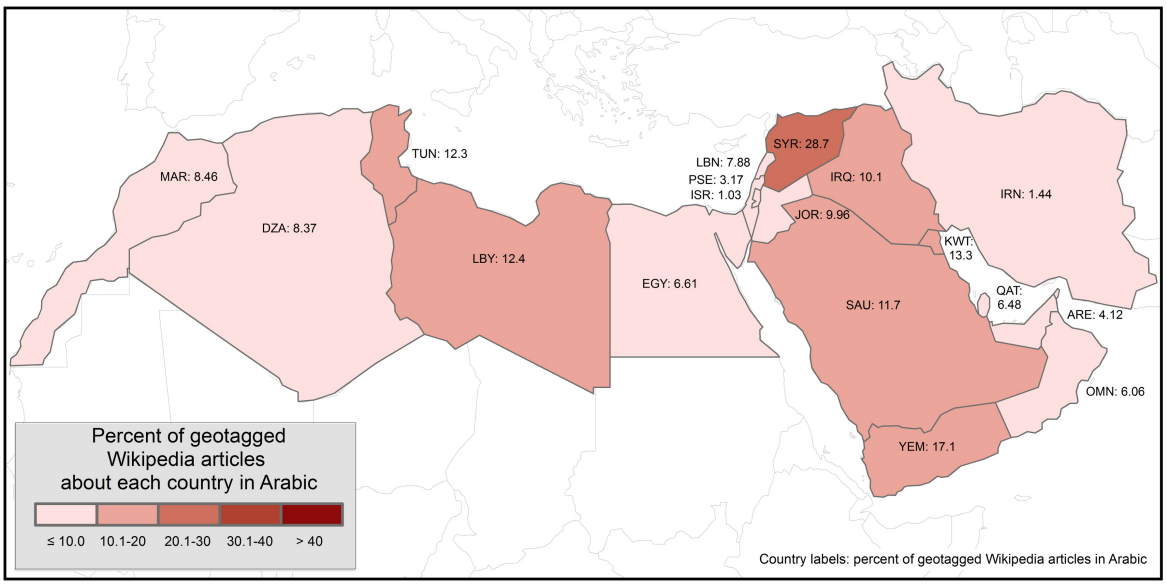

Figure 6.3.1b: The percentage of articles about a country in the MENA region that are written in Arabic.

The geography of Arabic Wikipedia articles shows how most countries with Arabic as a dominant language have roughly nine percent of their geocoded articles in Arabic, with substantial variation. Syria, as noted above, is extensively written about in Arabic. Yemen is also extensively written about in Arabic. Unsurpsingly, both Iran and Israel have a much smaller representation in Arabic as a percentage of the total. In the case of Iran, there is a strong presence in Farsi and in Israel there is a strong presence in Hebrew. Nevertheless, the fact that these countries where Arabic is the dominant language and it is not the dominant Wikipedia prompts us to ask – what is Wikipedia for? Is it for outsiders looking in or insiders looking out. For most of the world, it is for insiders looking out. Swedes use Wikipedia to document their towns, villages, and landmarks in Swedish. The Japanese use Wikipedia to further document train stations and railway routes. But in the Arab world, Wikipedia may feel much more like an outsider's project.

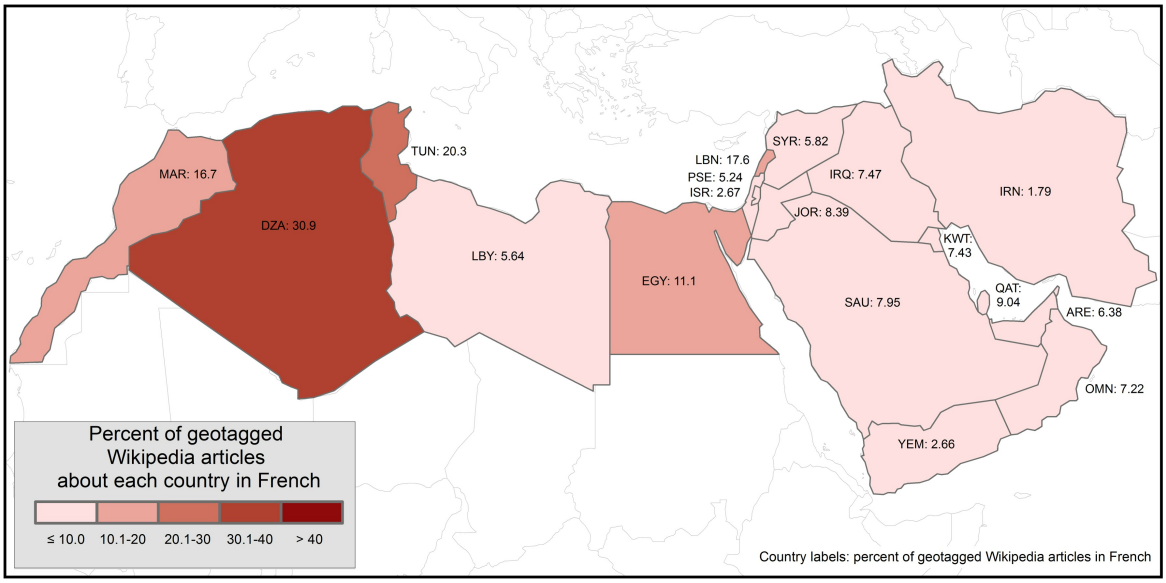

Figure 6.3.1c: The percentage of articles about a country in the MENA region that are written in French.

Languages are often symbols and signals of broader patterns of the geographies and histories of political power. We see this presented to us when mapping the relative coverage of French Wikipedia (fig. #). With the 18 countries in the MENA region, French is dominant in three: Algeria, which shows the highest value (30.9%), Tunisia with 20.3%, and Morocco with 16.7%. Lebanon is also noteworthy, because even though it has a higher proportion of English articles, it still has a relatively high proportion of French content at 17.7%. This certainly attests to French as a former colonial power and a residual interest in French, both within and beyond the borders of these countries. Although English is the world's current lingua franca, French is also a common tongue especially in these countries. If we are to consider the Wikipedia as a means of representing countries to the world, in these countries in particular it is both unsurprising and useful to see an extensive representation in French. It is highlights the potential for local actors who are more likely to speak French than English (especially in Tunisia and Algeria) to contribute meaningfully to self-representation on Wikipedia.

The Hebrew Wikipedia demonstrates the pattern we find elsewhere in the world about a national language being primarily used to represent within national borders. Hebrew-speaking Wikipedians clearly have a limited interest in representing the rest of the MENA region with a high degree of granularity. The average is exceptionally low (2.9%), and would be even lower if it weren't for Israel and the Palestinian Territories. Israel and the Palestinian Territories show the highest values, respectively 19.3% and 15.8%, of Hebrew Wikipedia articles. Israel and the Palestinian territories both have extensive numbers of articles in Hebrew, but in the rest of the MENA region the numbers are so low as to really only constitute a handful of articles, primarily about major cities. This suggests that a predominantly Hebrew speaker will have a limited capacity to learn on Wikipedia about specific places in Iran, Kuwait, Qatar, or Saudi Arabia even if such a user can learn a substantial amount about Israel and the Palestinian territories. Fortunately, most Hebrew

speakers are also fluent in English, and thus are probably likely to turn to the English Wikipedia for such representation.

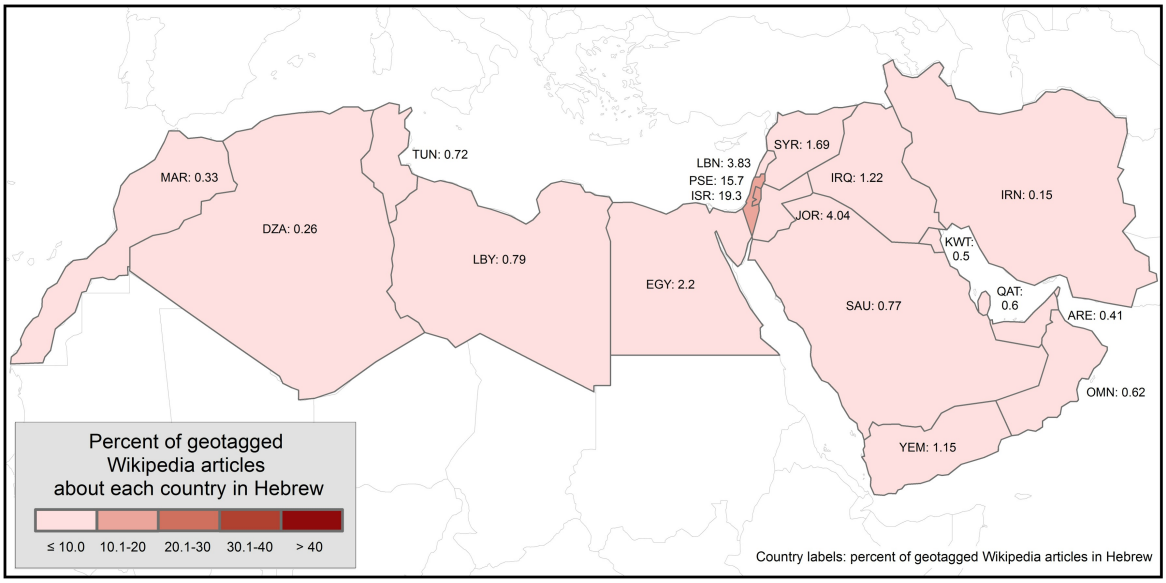

Figure 6.3.1d: The percentage of articles about a country in the MENA region that are written in Hebrew.

The distribution in Farsi in Iran shows how the country is particularly focused on self-representation. Even more so than Israel, where 19.3% percent of all articles in principal languages are in Hebrew, in Iran 28.8% of Wikipedia articles are written in Farsi. This high percentage may reflect the fact that whereas in Israel where there most people also speak English, in Iran, Farsi is not only the national language, but the primary language for most residents along with Pashto and Dari (the Afghani variant of Persian). Ethnologue does not even list English. Actually, to this point, given that English is not listed as a common language in either Ethnologue or Wikipedia, the fact that Iran is so well represented in English is a notable curiosity, moreso than the fact that Persian is so common. The only other moderately high values are found in Oman (6.4%), Iraq (6%), and Kuwait (5.8%).

Despite heated political rhetoric between leaders, we see a state of mutual disinterest in Wikipedia between Iran and Israel. Only 0.9% of Israel's Wikipedia articles are in Farsi (and 0.2% of Iran's Wikipedia articles are in Hebrew).

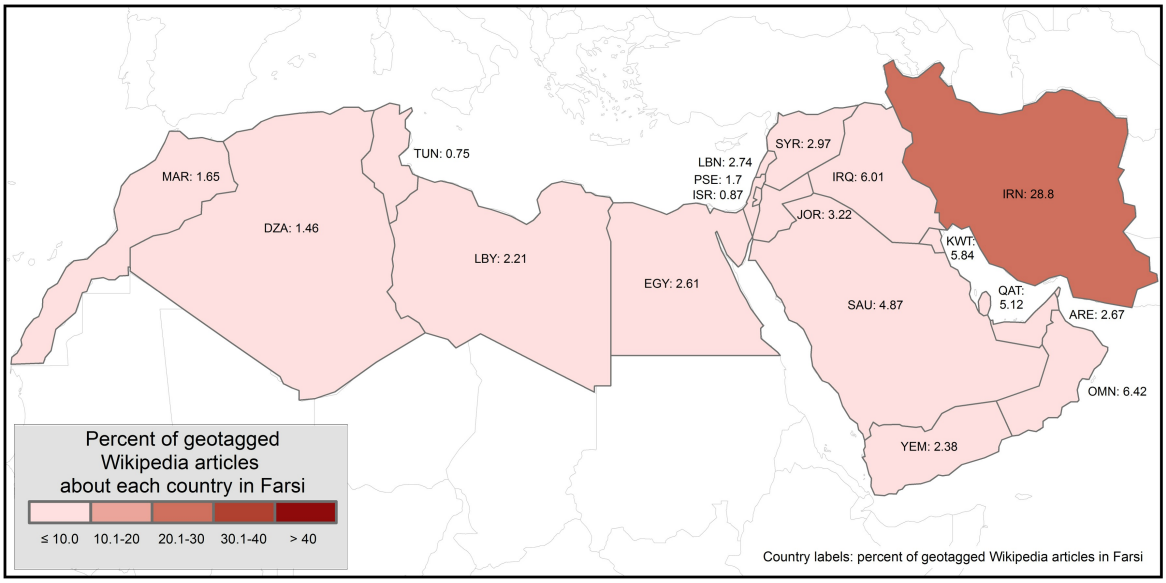

Figure 6.3.1e: The percentage of articles about a country in the MENA region that are written in Farsi.

6.3.2 Underrepresented languages: sub-national unevenness

Many of the patterns that we are interested in can also be elucidated by exploring the coverage of articles at the provincial level.

The first map (Figure 6.3.1a) shows the distribution of Arabic geotagged Wikipedia articles among MENA provinces. Israel, the Palestinian territories, parts of the Arabian Peninsula, and Iran all tend to have the highest counts. However, because some areas are simply larger than others, these data are perhaps not as revealing as they could be. To better understand these patterns it could be more useful to look at the number of articles per square kilometre (Figure 6.3.1b).

The densest layers of information in Arabic are again over Israel and the Palestinian Territories. Much of the Mediterranean coast in Morocco, Tunisia, and Algeria as well as the Nile valley and parts of the UAE seem also to have relatively dense clouds of content about them.

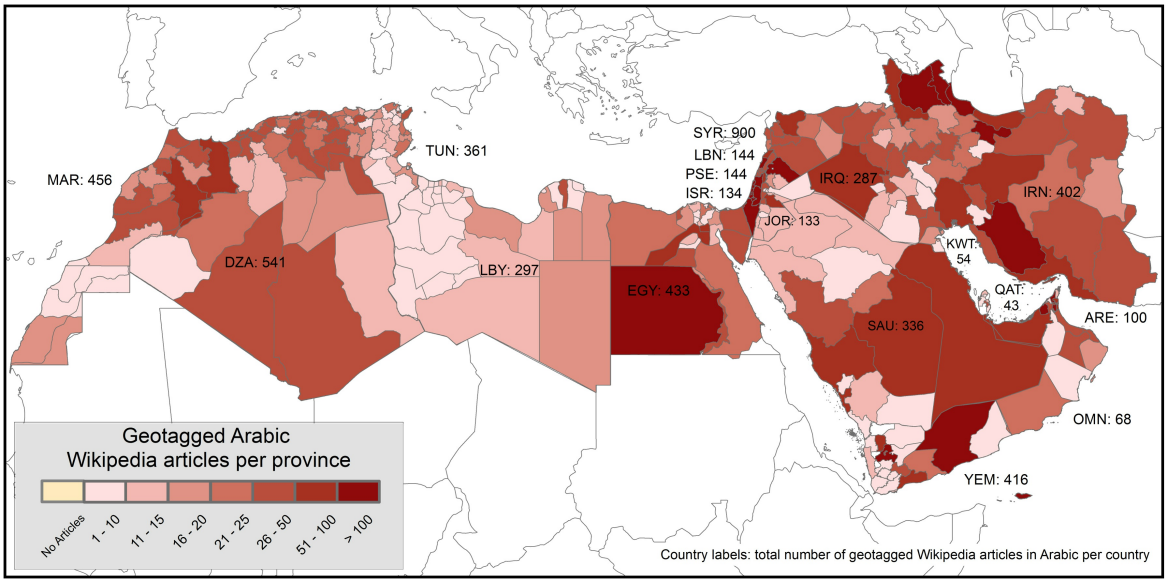

Figure 6.3.1a: The distribution of articles written in Arabic at the sub-national level within the MENA region.

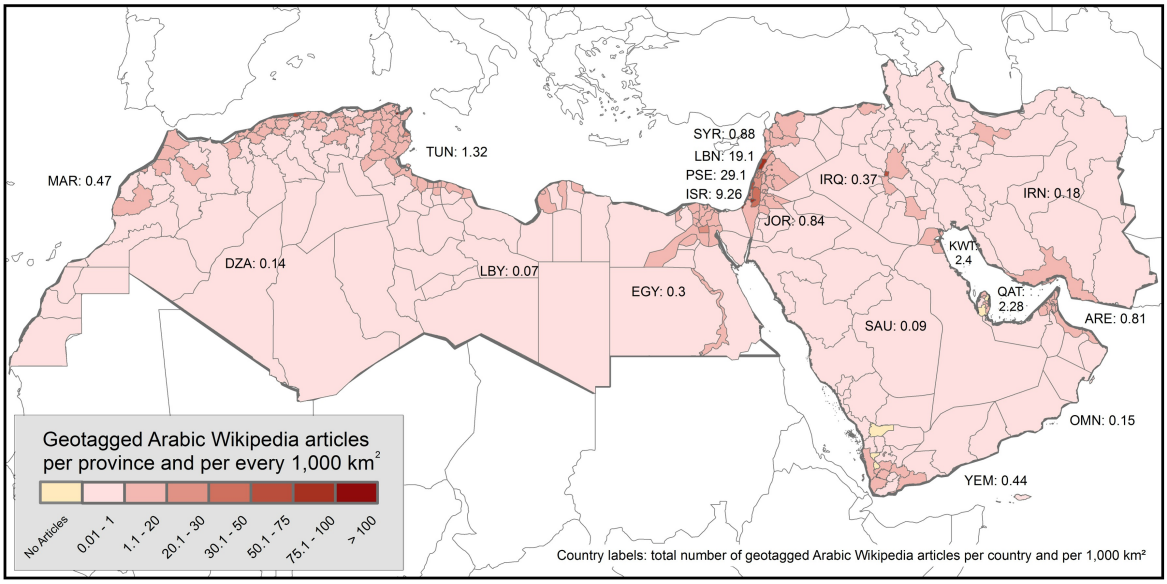

Figure 6.3.1b: The distribution of articles written in Arabic at the sub-national level within the MENA region normalized by area. Counts represent the number of articles per 1000km.

Looking at the distribution of English Wikipedia articles at a provincial level, the scale and scope of English language penetration into the region is evident. There is in fact only one province that has no English Wikipedia article at all: Quneitra in Syria (Al Qunaytirah in Arabic).

The concentration of English Wikipedia articles is especially strong in the Al Wadi al Jadid province in Egypt as well as in the Iranian provinces of Ardebil and Fars. High concentrations are also recorded in some provinces of the United Arab Emirates and Saudi Arabia, in the Hadramawt province of the Yemen and, despite their relatively small sizes, in Israel and the Palestinian Territories. In contrast, low concentrations of English Wikipedia articles are detected in Tunisia, Libya, and the Southern territories of Morocco.

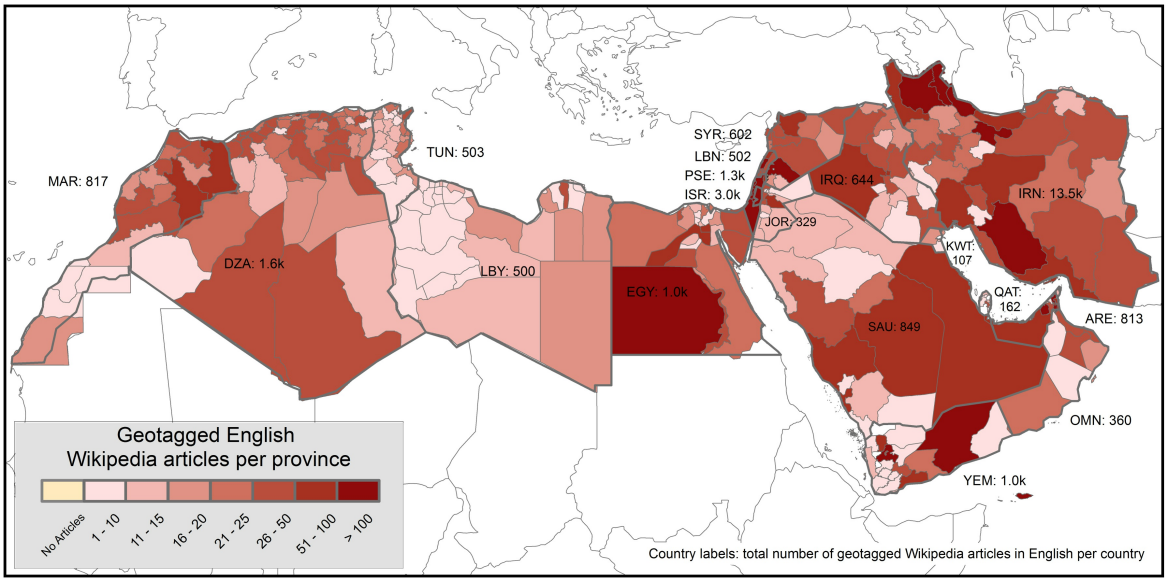

Figure 6.3.1c: The distribution of articles written in English at the sub-national level within the MENA region.

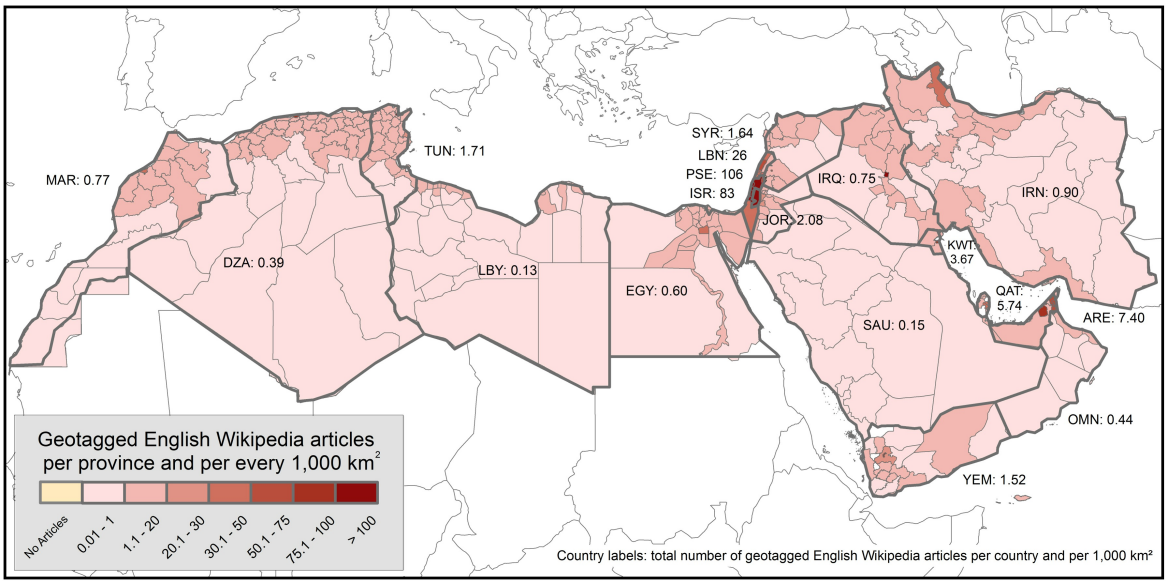

Figure 6.3.1d: The distribution of articles written in English at the sub-national level within the MENA region normalized by area. Counts represent the number of articles per 1000km.

We see a different picture when looking at the 'article density' in Figure 6.3.1d. Areas of high informational density include the provinces of Al Kuwayt and Hawalli in Kuwait, Baghdad in Iraq and Tunis in Tunisia, as well as Gaza and the West bank in the Palestinian Territories and Haifa in Israel. The very highest concentrations recorded in Beirut province, in Lebanon (2,275 per 1000km^2), and in Israel in Tel Aviv (1,554 per 1000km^2) and Jerusalem (757 per 1000km^2). These high densities likely exist because of the not only the large and historical cities (and thus presence of much to write about) in those provinces, but also the attention that those cities get in global media (thus encouraging non-locals to write about them).

The distribution of French Wikipedia Articles at the sub-national level again demonstrates post-colonial linkages between the French language are countries that are, or were once, within the Francophone sphere. One might also note the curiously high number of articles within Egypt's New Valley Governorate. This is primarily because of the extensive detailing of faraway Oases in both English and French. When we normalize by geography we again see the expected pattern of higher information density around Cairo, Giza, down the Nile and up to Alexandria.

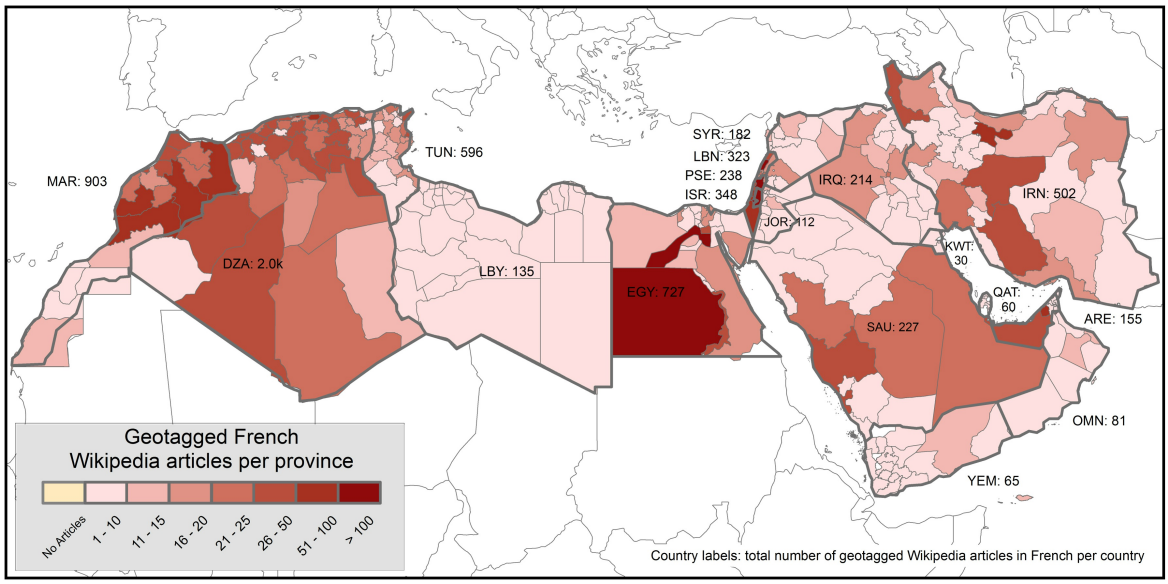

Figure 6.3.1e: The distribution of articles written in French at the sub-national level within the MENA region.

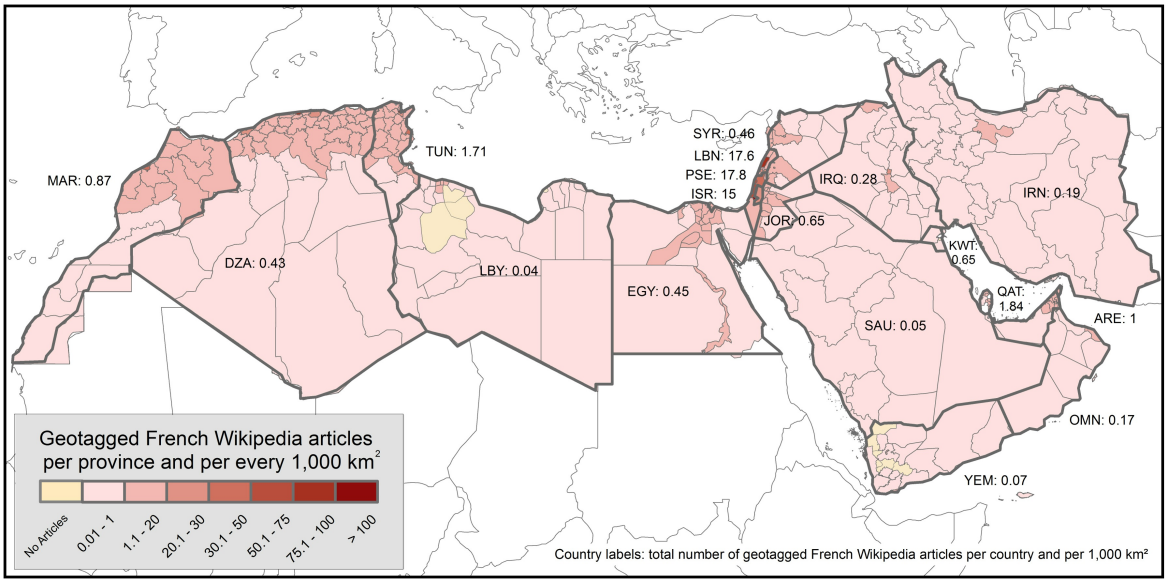

Figure 6.3.1f: The distribution of articles written in French at the sub-national level within the MENA region normalized by area. Counts represent the number of articles per 1000km.

North Africa, with the exception of Libya has a relatively broad spread of Wikipedia articles in French. The province with the highest amount of French content is the Mont Lebanon province (in Lebanon). This could be explained by the fact that Lebanon was French protectorate from the end of the First World War until 1946 and French is still widely spoken in the country.

This interpretation is strengthened by the next figure that shows the informational density of French content. The strong presence of French content in northern and more populated areas of Morocco, Algeria and Tunisia is clear. The low values recorded in Libya, by virtue of a very different historical and cultural heritage, are also notable. Moving East, we can see a strong focus of the French Wikipedia in Lebanon where Beirut has 900 articles per 1,000 square kilometres, followed by Tel Aviv with 288 articles and Al Kuwait province, in Kuwait, with 216 French Wikipedia articles in ever 1,000 square KM.

The distribution of Hebrew Wikipedia articles at the sub-national scale shows how Hebrew does not exhibit the sort of completest approach to documenting geography found in larger Wikipedias. Out of a total of 301 provinces in the Mena region, a full 101 don't have a single Hebrew article about them. These provinces are fairly scattered across the region. As mentioned earlier, the main object of attention in Hebrew is Israel, the Palestinian Territories, and parts of the neighbouring countries.

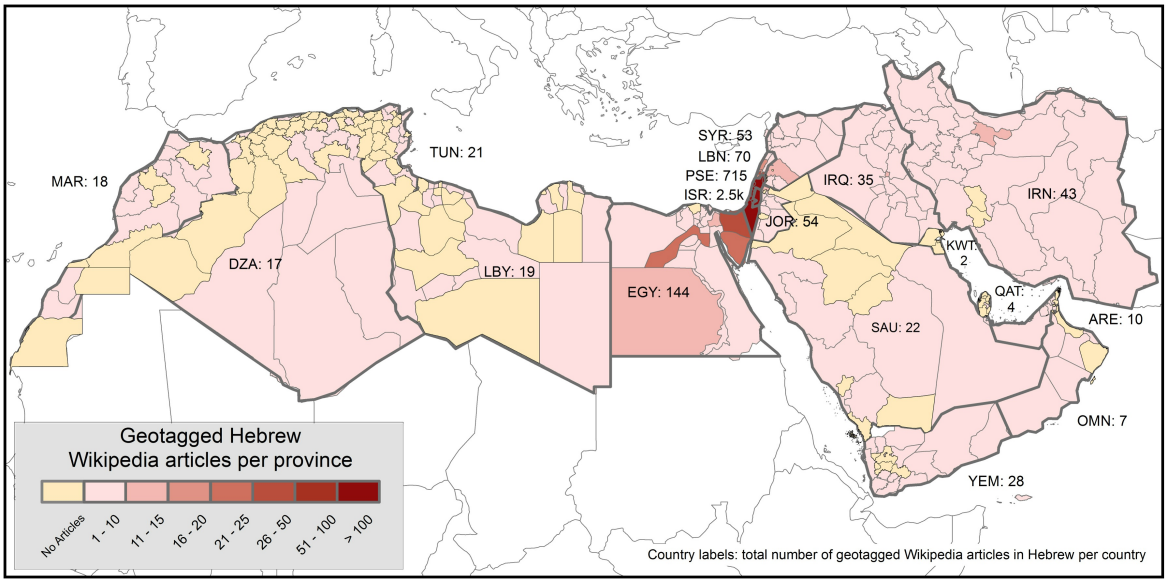

Figure 6.3.1g: The distribution of articles written in Hebrew at the sub-national level within the MENA region.

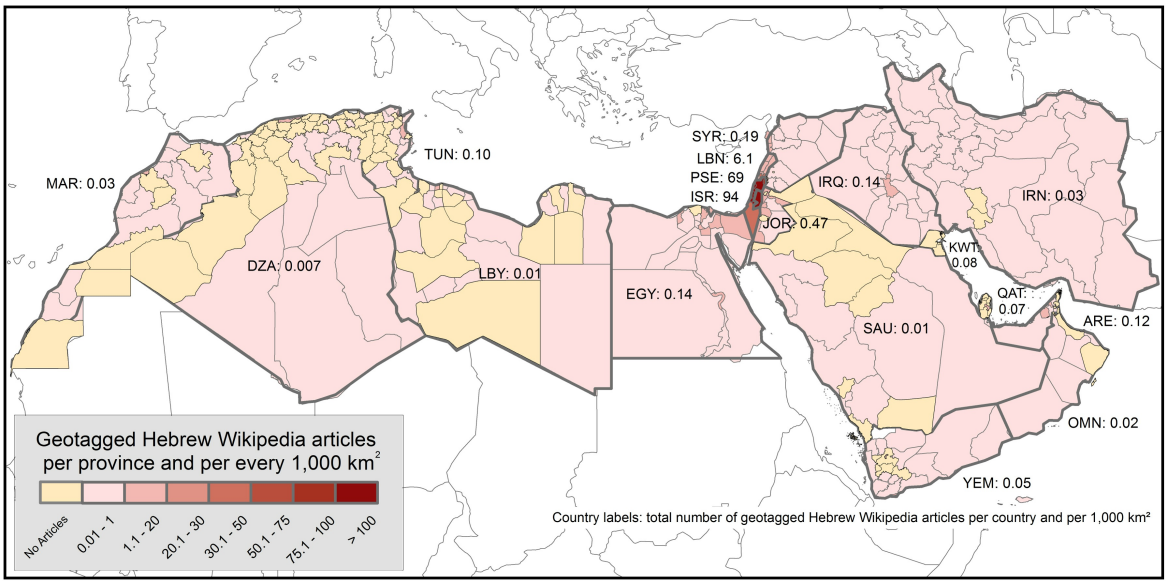

Figure 6.3.1h: The distribution of articles written in Hebrew at the sub-national level within the MENA region normalized by area. Counts represent the number of articles per 1000km.

That pattern is even more pronounced when looking at the information density of Hebrew Wikipedia (see Figure 6.3.1h). The *district* that with the highest Hebrew informational density is Tel Aviv with 2,880 articles per 1,000 square kilometers. The Palestinian territories also exhibit very high information densities. In the middle of this distribution there is the province of Beirut, in Lebanon, with 634 Hebrew Wikipedia articles per 1,000 square kilometers.

6.3.3 Final considerations

The data and maps presented in this section offered some insight into the geographic spread of content in different languages. In other words, it allowed us to better understand which parts of the world are covered in dense informational layers, and which are largely left out of the digital augmentations that are beginning to cover much of our planet.

To do this we have analysed the distribution of English, Arabic, French, Hebrew, and Farsi Wikipedia articles. English clearly occupies a dominant role in much of the region, and other languages are characterised by stark pockets of under-representation. Even though Arabic is widely spoken throughout the region, we see a surprising lack of content produced in that language.

At the same time, we see some countries and provinces that have healthy amount of content about them in Wikipedia. A goal of our work is now to not just highlight these stark inequalities, but also to try to understand them.

6.4 How MENA editors represent their region to the world

The previous sections have focused on the representation of the MENA region on Wikipedia, with most of the emphasis on what content is there, and how that content relates to visibility. In 6.1.3.3 we discussed editing and participation, but only between MENA and the rest of the world. In this section, we investigate the participation by MENA actors themselves and articulate the extent to which local content is in fact produced by locals. In the subsequent sections, 6.5 and 6.6 we look at how content is policed and the cultural / qualitative factors explaining these findings. We find that MENA actors are focused on local content, particularly within national boundaries. There is little evidence for a pan-Arabic culture of editing. Ultimately the bulk of content still comes from English-speaking Western nations. This is not the case for other countries around the world, including some where English is not the dominant language, such as Estonia and Bulgaria, highlighting that this is not a necessary situation.

In this section, we focus primarily on the English language Wikipedia. This is in part because we are looking at identified registered editors. As we discussed in Section 3.2 ["User geolocation methods"], it is currently not feasible to look at registered editors in other languages, although we believe that this will be a ripe field for future study. Further, we believe the methods we created in this project will be germane in helping these future researchers. The linguistic challenges for looking at identified editors in Arabic are non-trivial, but give Arabic's steady growth in use of Wikipedia and Wikipedia's steady growth in use elsewhere, this is an obvious important next step. There are fewer challenges in examining French editors – something that will be especially useful for the Francophone countries of North Africa.

In the subsections that follow we describe the patterns of inequality for both edits and editors. We begin with edits, first examining the sampled data that comes from Wikipedia across all languages and then the precise data, where we have complete information on all edits, but less complete information on all editors. Even though our means for identifying editors is incomplete, there is at least one distinct advantage to our approach: by using data from editors who have self-declared an association with MENA countries we are able to model those who wish to self-identify to the Wikipedia community their association with the MENA region. Such self-identification is part of the community building process as well as a potential source of conflict.

Table 6.4a summarizes the maps and analyses that follow. We focus six distinct qualities that we have deduced through this data and our discussions with Wikipedians themselves: influence, magnetism, potential, interest, activity and agency. The first two describe patterns of edits – are local editors influential and what areas of the world are considered attractive as sites for editing? The next four concern the editors themselves – how many are there in raw and weighted numbers, how much do these editors contribute and what characterises a culture of contribution?

Table 6.4a. Summary of subsections and their focus.

Focus	Topic	Data source	Interpretation
Edits			

	Number of committed edits per country across all languages	Sampled Wikipedia data	Editing influence (global)
	Number of committed edits per country in English, by both registered and anonymous editors.	Geolocated users and edit histories	Editing influence (English)
	Number of received edits per country by registered and anonymous editors: Local, MENA and Global	Geolocated users and edit histories	Editing magnetism (English)
Editors			
	Number of identified and anonymous editors per country	Geolocated users	Editor potential (English)
	Number of identified and anonymous editors per country per capita	Geolocated users, World Bank statistics	Editor interest
	Number of edits per identified editor by country	Geolocated users	Editor activity (English)
	Number of edits per number of views across all languages per country	Sampled Wikipedia data	Editor agency (global)

In the analysis of editors and edits, unless otherwise specified we use the unweighted geolocated data. This means that an editor could be double-counted if they indicate that they are from Egypt and work in Jordan. In the alternative, weighted, version such an editor would be counted as .5 for Egypt and .5 for Jordan. We provide rationales for both where relevant.

To briefly summarize our findings, we find stark differences in the number of editors between countries in the MENA region. Editors do tend to edit more locally than abroad as noted by previous research on anonymous editors (Hardy 2013). Nevertheless, even if editors are editing locally, the fact that there are orders of magnitude more editors from the rest of the world also editing, this means that it may be very difficult to ensure local representation. The good news, however, is that editing tracks extremely closely to views. Thus, if Wikipedia itself becomes more popular in a country as a site for the consumption of information, it will almost certainly become a site for the production of local information. Thus, in addition to the key driver of broadband is a need for greater interest in local geography. People who are interested in reading on Wikipedia are editing Wikipedia. In places where an interest in local geography is especially salient (such as Israel and the Palistinean territories), local editing follows suit.

6.4.1 Editing Influence: Global and Local.

In previous sections we discussed the patterns of representation across the globe with an implicit, and sometimes explicit, assertion that local editors are the key drivers of local content. Of course, this is not entirely true. Wikipedia is driven primarily by educated editors in the Global North. In Section 6.2 we demonstrated how North America, Europe, and Oceania do not merely edit locally at very high rates, but also export a great deal of edits to the rest of the world. Asia tends to receive as many edits as they export, but the MENA region and Sub-Saharan Africa tend to import far more edits than they produce locally.

Because of the substantial differences in broadband diffusion, we are dealing with fundamentally different orders of magnitude for participation. That said, even though editors from the Global North and particularly the USA and the UK dominate editing in general, when it comes to towns and cities beyond obvious national capitals and landmarks, local editors dominate. Furthermore, local editors help to clarify and flesh out articles whereas global 'power-users' (i.e. editors who edit a lot) tend to focus on administration, clean-up and more generic tasks. For this reason, it is important to evaluate the extent to which people are editing from any given country.

6.4.2: Editor activity across the MENA region.

The statistics from our identified editors tell a rather stark picture of editors not identifying themselves in the English language Wikipedia. This can happen for a number of reasons. One potential reason is that editors can have multiple editor pages in different languages. Hence an editor from the MENA region who is bilingual (or multilingual) and writes about the MENA region in English may be more inclined to self-identify on the Arabic user page or the French user page but not the English one. In this case, such a user would not have been identified. Another reason is a concern that the edit is coming from a politically motivated or defensive stance. A third is potential concern over state surveillance.

All three of these reasons emerged during our workshops and are noted in Section 6.6. To large extent, however, we believe that Wikipedians from the MENA region are self-identifying as often as editors from elsewhere. This is because the ratio of identified editors to edits from IP addresses does not vary significantly and we can be certain of the location of IP address edits, even if we cannot be sure which editor made which edit behind an IP address. Thus, we believe it is fair to compare the numbers of identified editors both between MENA countries and between MENA and the rest of the world.

Recall that our identified editors represent only approximately a tenth of all registered users, but these editors account for half of all edits from registered users. That is to say, identified editors are mostly power users, and power users clearly tend to articulate where they came from and where they live. This issue becomes a problem for editors in the MENA region who are looking for support or community. Visibility is not simply a matter of users indicating that they are representing their country, but the sort of effects that happen at scale, such as meet-ups, funding to go to conferences, trips to collect pictures, knowledge sharing activities about best practices and votes for Wikipedia's governance.

Of all our maps that show the stark asymmetries in the various regions, Figure 6.4.2a might be the starkest. That said, it does certainly help to explain the challenges involved in organizing editors from the region. For example, we were only able to identify 4 editors from Yemen who edit local content, 11 from Tunisia and 53 from the United Arab Emirates. We were able to identify several

hundred from Iran and Israel. With so few identifiable editors from these regions, we believe it is very difficult for such editors to have a strong sense of community or cohesion. Recall that we capture every user page from the beginning of 2012, but these pages might have been dormant for several years. As we discussed in our strategy for inviting editors from Wikipedia, many highly active editors from the MENA region have now ceased writing Wikipedia.

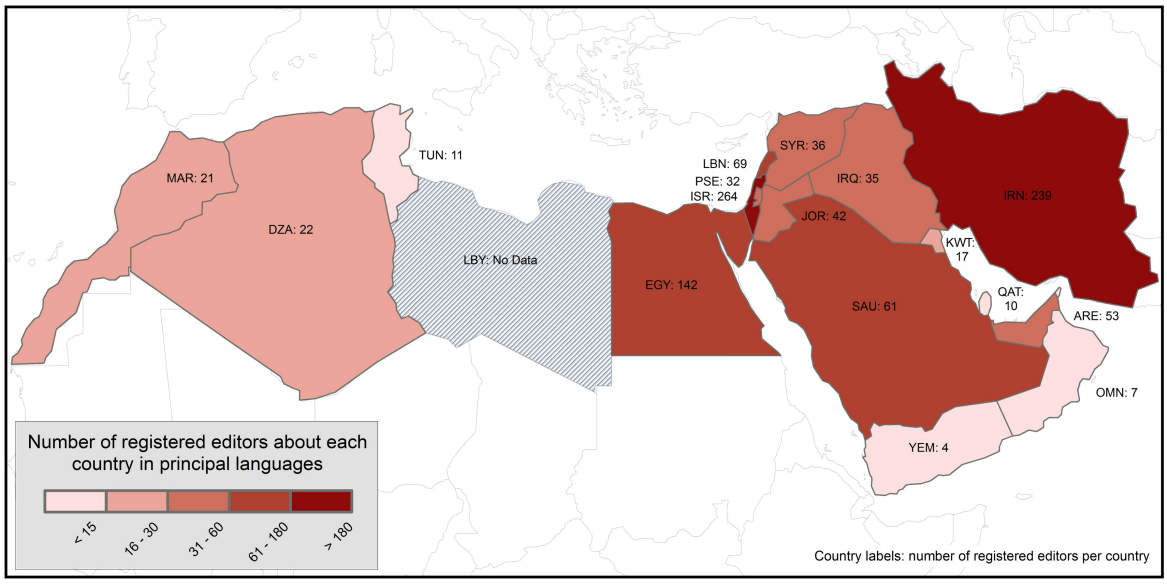

Figure 6.4.2a. Number of identified registered editors from MENA countries who have ever edited content about the MENA region in English.[30]

While the map only shows the number of registered editors, we summarize this information in Table 6.4.2a. Here we can see that registered editors to Wikipedia in the MENA region tend to be deeply committed to their work. Editors in the Palestinian territories in particular are incredibly busy.

Table 6.4.2b. Summary of edits and editors in the MENA region (unweighted).

Country	Registered Editors	IP addresses	Edits from Registered	Edits from IP addresses	Edits per registered Editor	Edits per unique IP address
ARE	53	1980	5583	8493	105.34	4.29
DZA	22	587	2322	923	105.55	1.57
EGY	142	3512	9325	6466	65.67	1.84
IRN	239	5459	16105	13300	67.38	2.44
IRQ	35	351	2986	662	85.31	1.89
ISR	264	8731	29283	18786	110.92	2.15

[30] Due to a programming error, data on Libya was not available for mapping at the time of the map generation. That said, data is available in Table 6.4.2b.

JOR	42	1752	1765	3442	42.02	1.96
KWT	17	1206	878	2451	51.65	2.03
LBN	69	1198	6176	4169	89.51	3.48
MAR	21	1283	3941	2263	187.67	1.76
OMN	7	649	105	1118	15.00	1.72
PSE	32	599	18480	1415	577.50	2.36
QAT	10	920	334	2906	33.40	3.16
SAU	61	2995	2692	6992	44.13	2.33
SYR	36	178	8009	651	222.47	3.66
TUN	11	479	167	817	15.18	1.71
YEM	4	147	241	326	60.25	2.22
LBY	16	233	2670	548	166.88	2.35

In Table 6.4.2a, the first thing to notice is the huge contrast between the number of edits per unique IP address and between identified editors. Whereas the number per unique IP address ranges from 1.5 to 4.3, the number of edits per identified editor ranges from 15 (Tunisia) to almost 600 (the Palestinian territories). That there are 32 editors who lived or were born in the Palestinian territories who made almost 600 edits each on average –to articles in the MENA region- indicates the extent to which such editors feel committed to local representation. To note, for those Palestinian editors, there is a clear focus on editing local content. These editors made 288 edits to articles in Egypt and 277 to articles in Iran. But they made 7433 edits to articles in Israel and 8590 edits to articles in the Palestinian territories. This is similarly reflected the edits from Palestinian IP addresses. Anonymous editors made only 6 edits to Egyptian articles and 12 to Iranian articles, but 254 to Israel and 1060 to articles in the Palestinian territories.

From the Israeli side, we see a similar pattern of focus in the area with an emphasis in the home country. Identified Israeli editors made 12,074 edits to articles located in Israel and 3519 to articles located in the Palestinian territories, but only 220 edits to articles in Iran and 349 to articles in Egypt. While there has been some concern about concerted editing by countries towards other countries, at least in these cases, it appears that local authors (either logged in or anonymous) are much more interested in writing about their country, with Israel and the Palestinian territories' longstanding conflict being a notable exception.

Such local focus does not necessarily imply that the bulk of edits to a country come from editors in that country. As we noted in Section 6.1.2a, in many countries, the locally dominant language has the greatest geographic coverage. There are more articles in Italian about Italy and more articles about Japan in Japanese, but in the MENA region there are more articles in English or French than in Arabic (with Syria being the only exception). However, that the articles are in English or French does not imply that the articles are solely written by foreigners or those with no local association. With data in hand about the location of editors we can explore this issue in greater detail below.

The three maps shown in Figures 6.4.2b-d detail the percentage of edits (either identified or anonymous) to a country from one of three areas: the country itself, the other countries in the MENA region and the rest of the world. The first thing to notice is that the majority of edits still come from the rest of the world. This is not necessarily the case for other countries. For the dominant English-speaking countries as well as a handful of others around the world, the majority of edits come from the country itself. Eighty-six percent of edits to articles about the USA come from there. Seventy-nine percent of edits to Australia come from Australia. In descending order the

next ten are Great Britain (78%), Canada (72%), New Zealand (71%), The Philippines (68%), India (65%), Ireland (63%), Norway (57%), Estonia (55%), Romania (54%), and Bulgaria (52%).[31] Notably, it is not necessarily the case that a country needs to have a predominantly English speaking population in order to have a majority of Wikipedia edits in English come from that country, as observed in Norway, Estonia, Romania, and Bulgaria. Indeed, there are fifty-six countries with a greater percentage of local edits than Iran (the highest value in the maps shown).

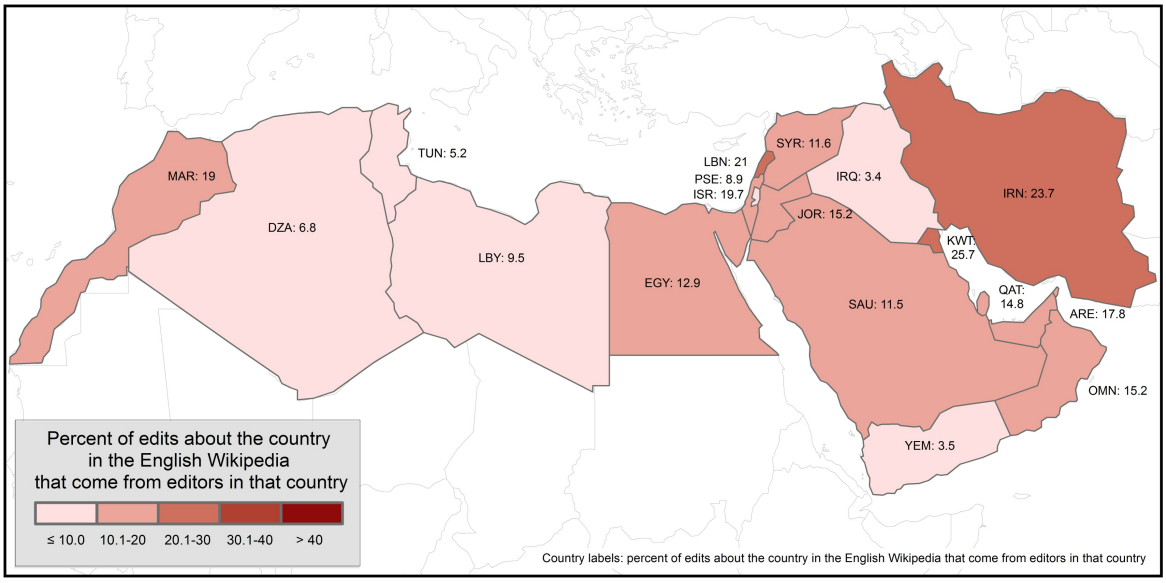

Figure 6.4.2b. Percent of edits about a country that come from editors associated with that country.

[31] These numbers were generated using the weighted edit data so that all edits total to 100%. By using the unweighted numbers we would have greater than 100% of edits accounted for since one could be from multiple countries. Nevertheless, the weighted and unweighted counts of editor-to-country have a Pearson correlation of 0.95.

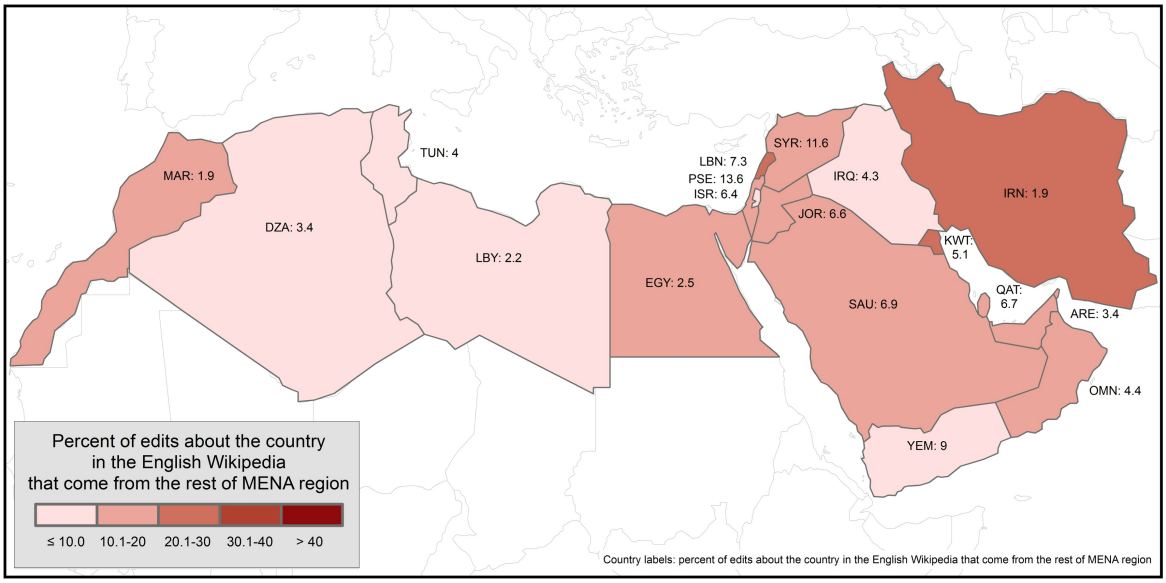

Figure 6.4.2c. Percent of edits about a country that come from editors associated with the rest of the MENA region.

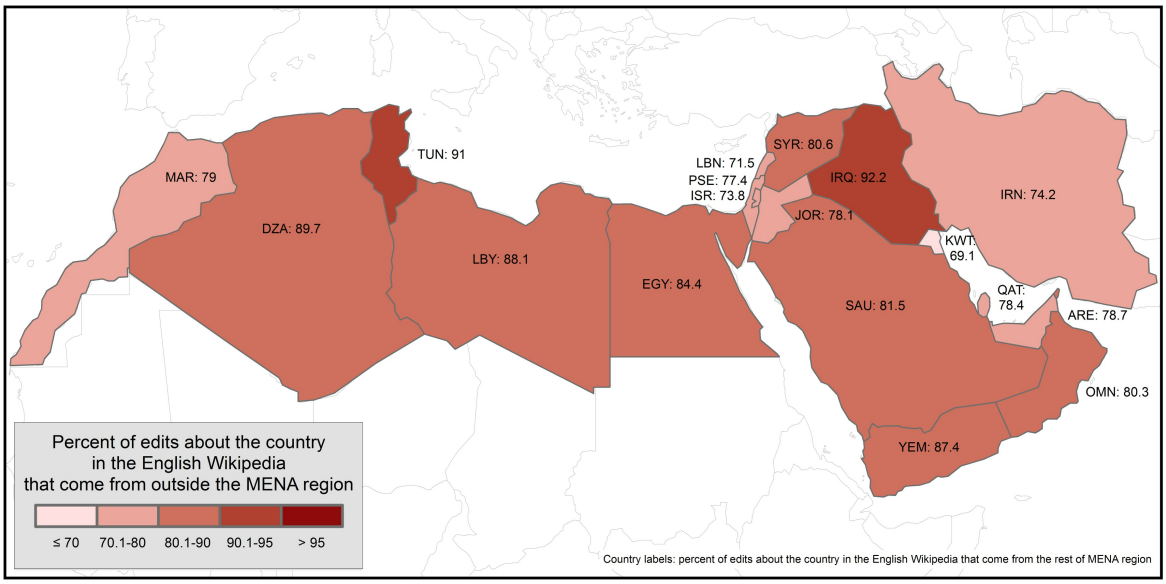

Figure 6.4.2d. Percent of edits about a country that come from editors associated with countries outside the MENA region.

Two possible (and proximate) reasons for the MENA region's lack of activity in the English language Wikipedia include a lack of interest by the population in general, and a lack of interest (or agency) for those who consume Wikipedia articles.

To explore this issue further, we return to the analysis in Section 6.2 about the number of editors per million inhabitants. To be particularly generous, we are counting editors who have edited any Wikipedia content, not merely those who edit content about MENA. That said, in virtually every country in the MENA region (and most all countries for that matter) about one third of editors will

edit local content. Thus, we display the numbers for all identified editors to highlight those would could be considered fellow Wikipedians. Moreover, these are the people who could help with community building since there are many commonalities among Wikipedians whether they write about molecular gastronomy or MENA politics. Local wisdom about vying for sources, articles and the favour administrators operate under broadly the same principles, regardless of the subject matter. In this figure we can see that per million people, there are strikingly few identifiable editors in the MENA region. Thus not only are few people within a country speaking for it, but that the likelihood of personally knowing an author is also small. Of course authors seek each other out, but with so few editors in any given city there is the risk that any specific community becomes cliquish, hard to find or skewed demographically (such as predominantly representing men, a dominant ethnic group or religion).

Once again, we note Israel dominating the statistics. There are 61 identifiable editors from Israel for every million people. Lebanon comes in second with 27 identified editors per million. On the other end of the scale, Yemen and Algeria have less than one editor per million people. A Yemeni's task of fleshing out articles on local cities, attractions, and political matters is a lonely task (especially because some of the editors with accounts registered in Yemen may not even live in the country any more).

One of the challenges that come from a lack of critical mass may be the lack of agency among consumers of Wikipedia. If the website is to be considered a primarily English enterprise, a Western one, or one that reflects views that depart from mainstream sensibilities in the region, individuals may not feel as if it is worth contributing. A final example demonstrating this returns to the data from Wikipedia themselves by exploring the ratio of views to edits. Again, these data are sampled data so they potentially underestimate the values in smaller countries. On the other hand, they operate across all languages. What this graphic shows is not the differences in editors or places, but variations in *agency*. That is to say, in countries that have a low number of edits to views, Wikipedia can be seen as primarily perceived as an external site for consumption rather than a site for local production.

The MENA region has a lower average edits-to-view ratio than the rest of the world. The average for MENA countries is 0.50 edits for every 1000 views, whereas it is 1.05 in the rest of the world. This difference is robust to outliers (i.e. by removing highly active countries such as Israel). While this average appears substantially lower, using a standard one-tailed t-test, it is not actually statistically significant ($p=0.15$, $\alpha=0.05$). Thus, as greater interest in Wikipedia emerges, it is likely that edits will increase in these countries as well. Nevertheless, it is still worth considering countries where Wikipedia is seen more as a form of consumption within the MENA region. In this case, the lowest edit-per-view ratios are in Morocco (0.27), Tunisia (0.29), and the United Arab Emirates (0.32), while the highest are in Israel (1.19), Iran (0.94), and Yemen (0.74). The remaining MENA countries all have a ratio between 0.37 and 0.56.

While Tunisia or Morocco may never have as many edits as Israel or Iran on account of their small population, it is still possible to increase the edits-to-views ratio. A larger edits-to-views ratio implies that Wikipedia users are shifting from being consumers of content to producers, even if it is happening at a small scale.

To put this figure into perspective, it is worth noting that the edits-per-view ratio is not particularly high among individuals in the English-speaking West. The ratio is 0.35 for America, 0.65 for the United Kingdom and 0.49 for Canada. Israel, Libya and Iran all have much higher edit to view ratios than these countries.

This is partially an optimistic consideration. Countries in the MENA region are roughly as likely as Internet users around the world to contribute to the website given the same level of interest in the site. Given the previous log-scale-sized differences in views and edits between the MENA region and the rest of the world, as well as the fact that the encyclopedia is presently written by non-locals, one might assume that users from the MENA region might consider Wikipedia as merely a site for (Westernized) content consumption, when necessary. However, we find that similar level of consumers turn towards production. Thus, we believe the site is actually hampered by less interest fostered by a lack of access.

In this section we have noted a clear conundrum: Wikipedians have a keen interest in writing about their local area. From past sections, we can also see that this interest tracks with Broadband diffusion very well. We have also noted Wikipedia's use as a generative technology that inserts itself into much of the architecture that powers the web, such as Google, Facebook, and Travel guides. Yet, for many countries and most certainly within the MENA region, it is not simply that they are not well represented on the site, but what representation exists tends to be done by foreigners. How then to elevate local content to create greater visibility, empowerment and representation?

Before turning to this question explicitly, we want to review a final piece of the quantitative Wikipedia puzzle: the act of policing content through reversions. While all editors can create or revert content, the act of monitoring what others have written and asserting a form of quality control is also not shared equally among editors. So who polices who?

6.5 Policing Wikipedia: National pattern of Reversions

It is not merely the case that Wikipedia is a bucket holding all activity from editors. Articles shrink, are vandalized or reverted to earlier versions. In its early days, Wikipedia suffered from a stigma that it was untrustworthy, primarily because anyone could edit it. The response from the Wikipedia community has been thorough, technical and careful. Many articles are watched by editors who review every change. Articles are monitored by scripts ('bots') that have an strong track record on identifying vandalism (Smets et al. 2008). They are also monitored by people who place their preferred articles on a 'watch list'. Editors who consider a specific edit to be a problem, either as factually incorrect, in bad faith, biased or any other of a host of reasons can roll back any changes most recently made. This act of rolling back is called reverting.

It is now well established that reversions are a central part of the Wikipedia editing process. Such an act is recorded in the editing record as a reversion. If the author of the reverted content believes this was unfair they can revert the revert. When two authors repeatedly do this, it is referred to as a 'Wikiwar' (Brandes and Lerner 2008). Wikipedia have instituted policies that inhibit such wikiwar behavior, including a cooling off period and the ability to seek arbitration from administrators.

The analysis of reversions takes on a slightly different character than the analysis of other aspects of Wikipedia. This is because there is always a reverter and a "revertee". Since we have identified the locations of many editors, particularly power editors, we have been able to construct country-specific networks of reversion patterns. These networks indicate who is putting in the most edits that get reverted as well as who is doing the most policing of content coming from other countries.

Reverting has serious consequences for Wikipedia. It is now known that first-time editors who have their content reverted are far more likely to never come back to the site. Halfaker has not only demonstrated this statistically, but also experimentally. Using the 'be nice' plug-in, editors are warned to be nice to first time editors so as not to scare them away (Halfaker et al. 2011). It worked. Implicit in this practice is that reverting is a form of power. Rather than writing over the content as if one is working with a collaborator, reverting is a blanket and symbolic rejection of an author's work, as if one is grading the other person.

In this section we examine the reverting patterns as an expression of power. Without reviewing every revert, we are not able to make strong value judgments on the specific content that is reverted. Thus, we cannot say that people from a certain country are more likely to create misunderstood content or that others are particularly likely to target content made by people from that country. We assume it is a little of both.

Previous figures have projected values from Wikipedia on to a map. However, in the case of reversion patterns, a map is not necessarily the most effective way to visualize these patterns. Instead, we use a network analysis approach. In this approach, each country is denoted by a circle. If a country reverts another country, then there is a line with an arrow going from the country who reverts to the country who is reverted. We have tweaked these networks in a number of ways. In general, we show a pattern of "self-policing" on Wikipedia, where members of a given country tend to be the ones doing the most reverting, whereas those being reverted tend to come from elsewhere. The ones being reverted are often from the main English-speaking countries on Wikipedia, but not always, as is seen below.

6.5.1 – Reversions in the MENA region: an Overview

In the following work we first show a global map of reversions to articles in the MENA region. In all cases, the size of the node represents the number of reverts that they send. The colour of the node represents the number of reverts that they receive. Thus, a larger node sends more reverts (I.e. does more 'policing') and a smaller node does less. A darker node is reverted more often (I.e. is "policed"), a lighter node is reverted less often. In many cases we also eliminate edges if there are too few reverts. For example if the editors associated with Tunisia only revert one person from Australia this is not really enough to suggest a strong signal of reverting from Tunisia to Australia. If there are ten reverts on the other hand, there is good cause to suggest a specific relationship. If we were to show every edge between two nodes (or every revert between two countries), then the graphic would look overly busy.

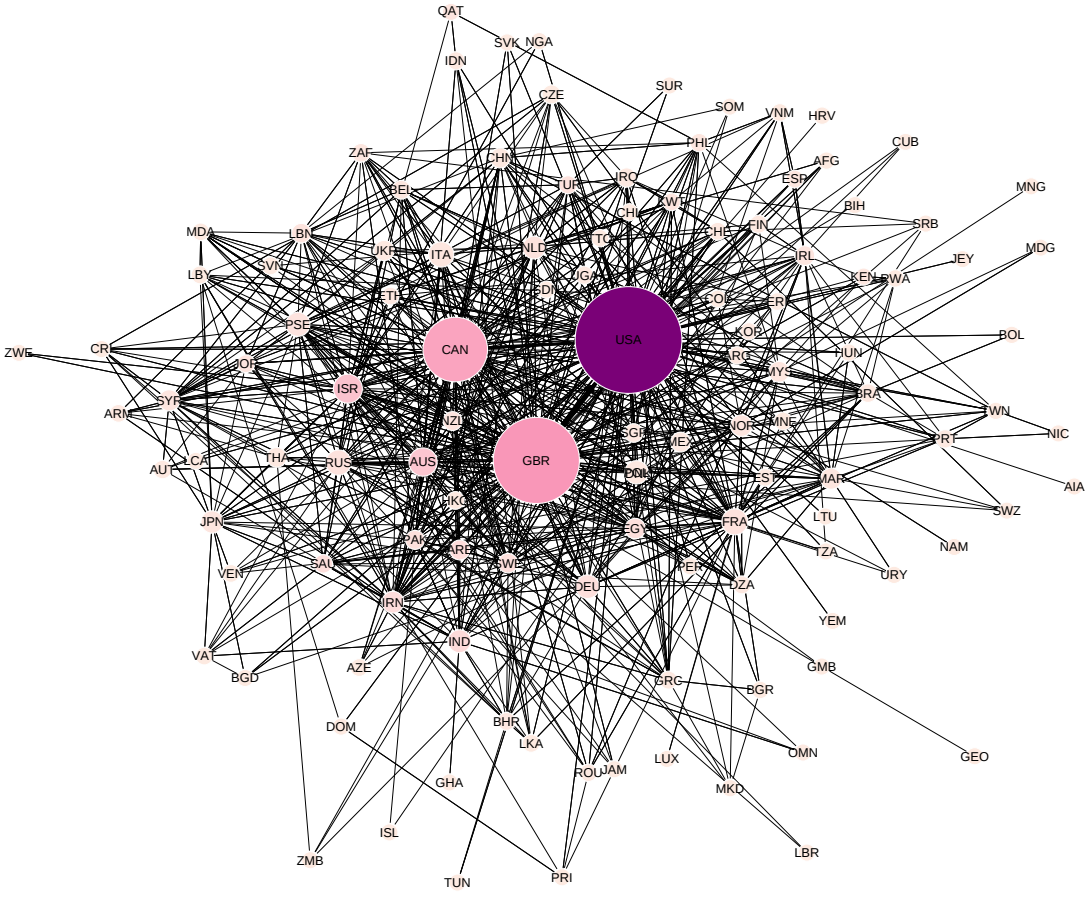

Figure 6.5.1a. Network of reversion patterns to articles referring to the MENA region in English.

In Figure 6.5.1a we can see that Great Britain a substantial amount of reverting (and more than average), despite not being the most active on Wikipedia, whereas America receives the most

reverts. That America receives the most reverts is partly just a function of the fact that the most edits come from America.

Reversions to the English Wikipedia are inevitably going to be done more by English speakers, and from English countries, but there is still a notable story in here about the cultures (or lack thereof) of editing within MENA regions. However, for the sake of brevity here, we do not feature and detail the reversion patterns within every country, but select specific countries out of an academic research interest and that fact that such countries have heretofore had notable histories on Wikipedia. Below we show the reversion patterns for Egypt, Israel, Iran, Yemen and Algeria.

6.5.2 Country-specific examples of reversion patterns

Reversions to articles on Egypt in the English Wikipedia tell a characteristic story found in many other countries. That is, the United States tends to be the country that is reverted the most, Great Britain is the country that does the most reverting, and in between all of the English speaking countries is the host country (obviously Egypt here) that tends to revert much of the content from others. That is to say, Egyptians are indeed somewhat protective of articles about Egypt and are ready to undo certain content that they consider inadequate.

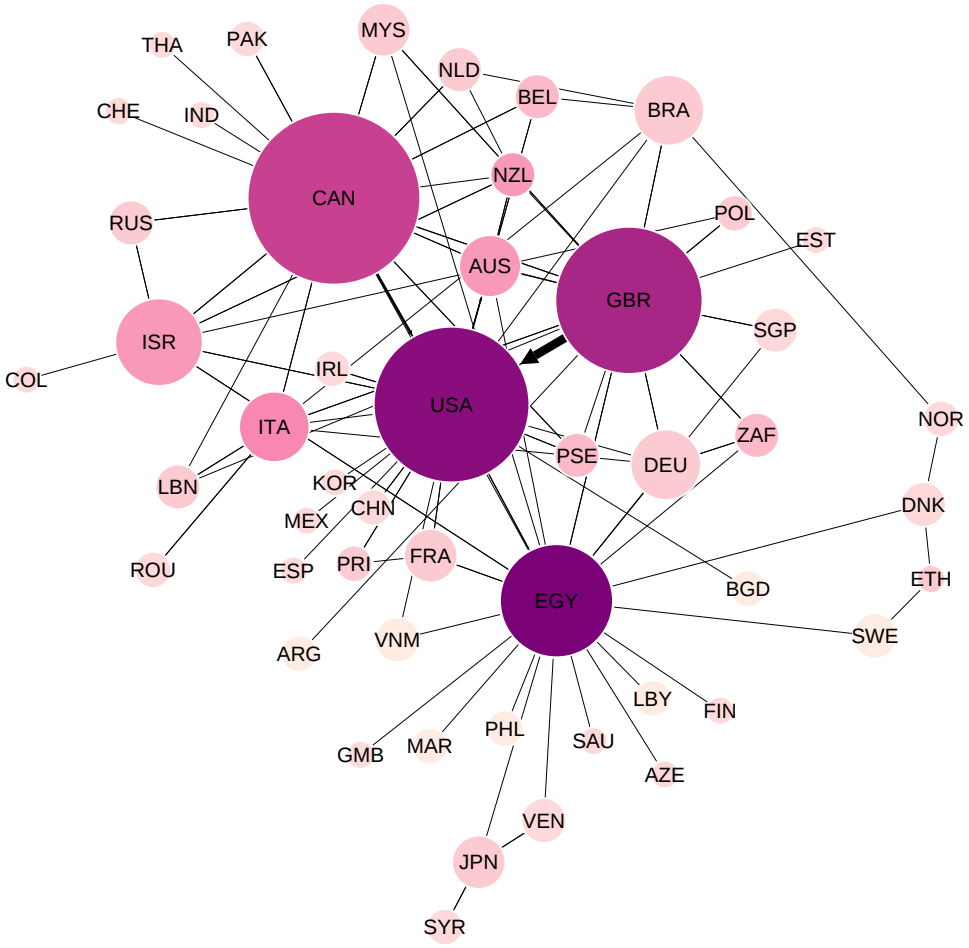

Figure 6.5.2a. Network of reversions from and to countries on articles in Egypt. Larger nodes do more reverting, darker nodes are reverted more often. Layout is automatic based on connectivity (using Force Atlas 2 in Gephi), with slight adjustments for readability.

By looking at the reversion network for Iran, one can see a story that unfolds in a similar way, with a modest exception. Perhaps owning to a great deal of migration from Iran to the Netherlands, there is in fact a large amount of reverting of Iranian content coming from Dutch editors. That this is happening in the English Wikipedia is indeed a curiosity that demonstrates how Wikipedia manifests migratory and geopolitical events beyond merely a story of the English core and the non-English periphery.

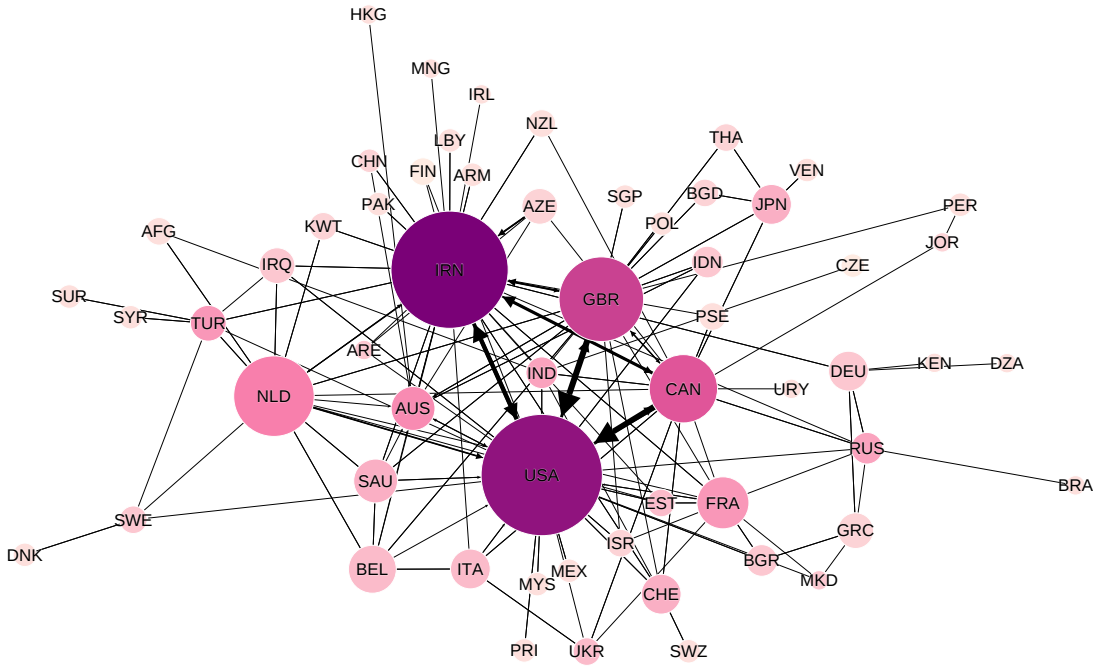

Figure 6.5.2b. Network of reversions from and to countries on articles in Egypt. Larger nodes do more reverting, darker nodes are reverted more often. Layout is automatic based on connectivity (using Force Atlas 2 in Gephi), with slight adjustments for readability.

The story in Israel is roughly similar to that of Egypt and Iran except in this case, unsurprisingly, a second MENA country appears to be very central to the reversion networks. Palestinians do a substantial amount of reverting to articles in Israel. However, in this case, it is just as likely that Palestinians are as protective of content in Israel.

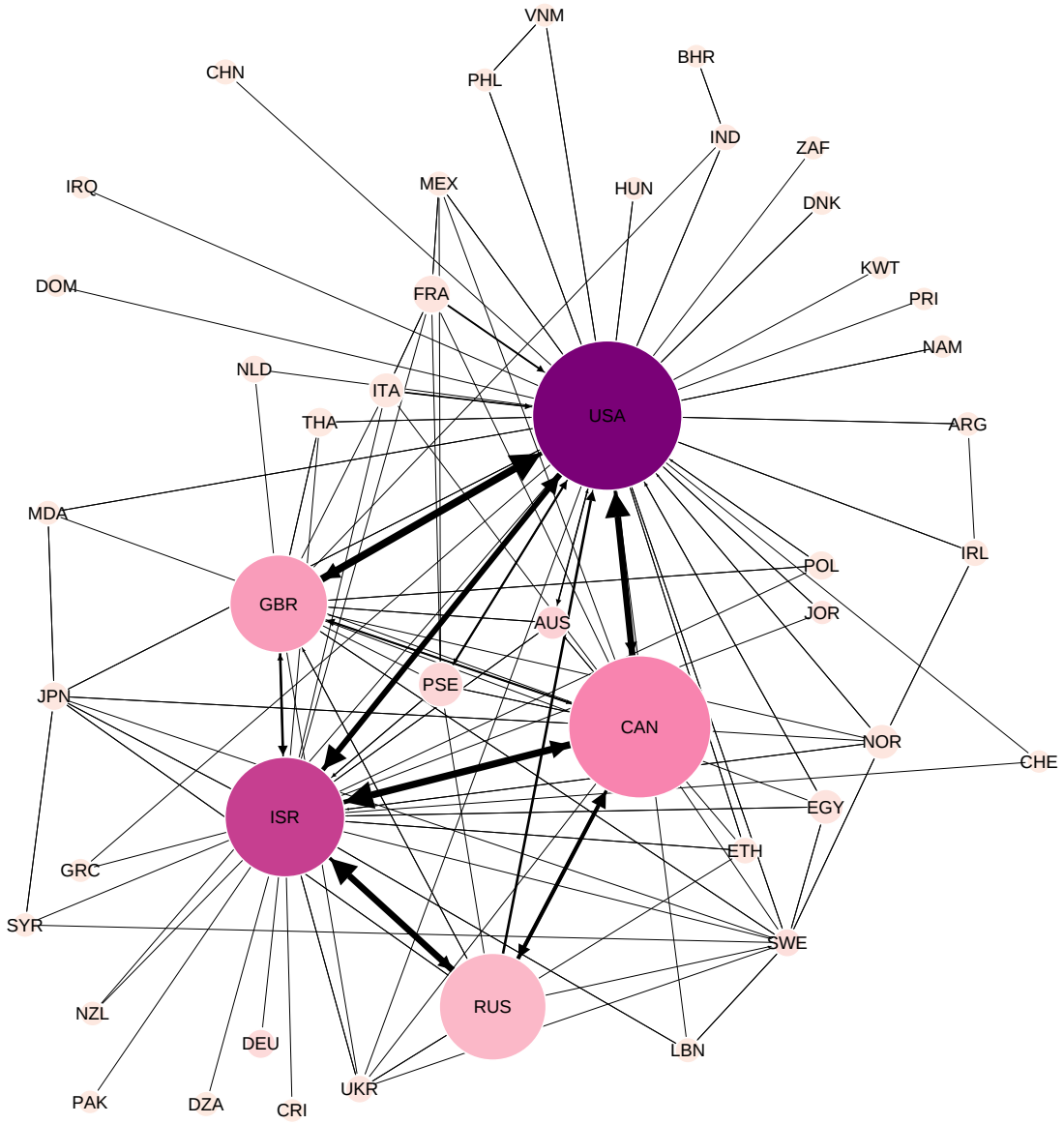

Figure 6.5.2c. Network of reversions from and to countries on articles in Israel. Larger nodes do more reverting, darker nodes are reverted more often. Layout is automatic based on connectivity (using Force Atlas 2 in Gephi), with slight adjustments for readability.

Tunisia shows an interesting contrast to Israel and Iran. Not only because of the size of the reversion network, but for the fact that no identified Tunisians were either reverted or reverted others. Again, we see the pattern of most reversions coming from the UK and most reversions going to the U.S., with Canada, Australian and New Zealand also reverting the Americans.

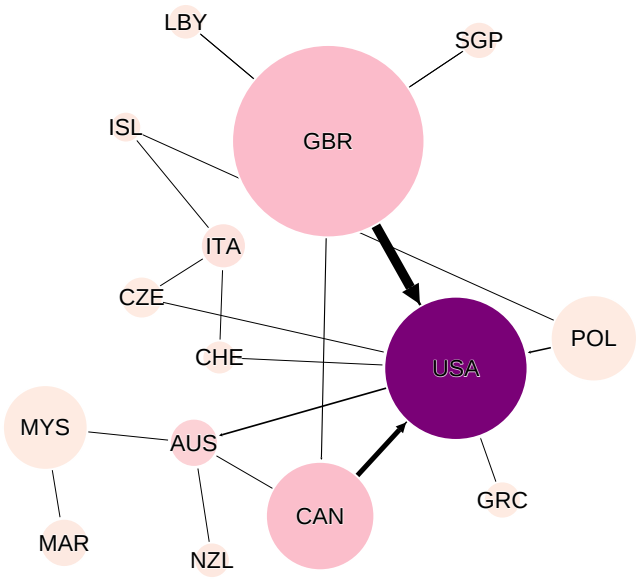

Figure 6.5.2d. Network of reversions from and to countries on articles in Tunisia. Larger nodes do more reverting, darker nodes are reverted more often. Layout is automatic based on connectivity (using Force Atlas 2 in Gephi), with slight adjustments for readability.

In general, within the reversions, we can see evidence of a core-periphery of Wikipedia's administrative power, with the United Kingdom being especially central to the overall governance structure of the site. Further analysis should look at whether the U.K Wikipedians are being judicious with their use of reversions, especially given the strong organizational presence of the United Kingdom Wikipedia. Being able to provide outreach to foreign Wikipedians and bringing them more further into the fold, rather than merely regulating their content is likely to have many positive benefits for the development and maintenance of high quality content.

On the other side, it appears that in every single country, most reversions are targeted towards America. However, as we note below, this is not because Americans are particularly poor editors of Wikipedia, but more because the sheer number of Americans editing Wikipedia lead to more reversions.

6.5.3. Volatility of edits

Editing on Wikipedia is a more volatile process than one might think from the outside looking in. Reverting content is an everyday occurrence. In Table 6.5.3a we describe the volatility of edits within countries in the MENA region, taking into account both identified and anonymous edits as a ratio of total edits. Thus, a score of .2 means that 20% of all edits are reverted. Here we can see that while there are only a small number of edits coming from the MENA region to other countries in the MENA region in English, these edits are far less likely to stick.

Israel stands out as highly volatile in this regard. Four countries from the MENA region have had at least half of their edits to articles in Israel reverted: Iran, Iraq, Yemen (which had every edit reverted)

and Tunisia. Several others have over forty percent of their edits reverted. All of Yemen's edits to the Palestinian territories were also reverted.

By contrast, content about Tunisia and Libya is much less volatile. For the most part, roughly one in ten articles to these countries from other MENA editors are reverted.

The network diagrams suggested United States is the biggest source of volatile content. Yet, when we look at the ratio of edits to reverts, content from the U.S. is not as likely to be reverted as content from many other countries in the MENA region. That said, content from editors in the United States is still more likely to be reverted than content from editors in Canada, the United Kingdom or Ireland. Interestingly, when French editors edit about the MENA region in English they are the least likely to be reverted of the countries featured here.

Country the editor is writing about

Country of editor being reverted	ARE	DZA	EGY	IRN	IRQ	ISR	JOR	KWT	LBN	LBY	MAR	OMN	PSE	QAT	SAU	SYR	TUN	YEM	Average
ARE	0.14	0.21	0.26	0.13	0.27	0.41	0.22	0.19	0.18	0.13	0.23	0.27	0.36	0.19	0.34	0.35	0.09	0.28	0.24
DZA	0.50	0.12	0.06	0.24	0.14	0.10	0.00	0.50	0.06	0.00	0.07	0.00	0.03	0.00	0.09	0.13	0.07	0.20	0.13
EGY	0.33	0.17	0.10	0.18	0.19	0.16	0.11	0.05	0.10	0.11	0.06	0.15	0.09	0.20	0.19	0.16	0.05	0.11	0.14
IRN	0.20	0.00	0.14	0.09	0.14	0.52	0.06	0.05	0.11	0.07	0.00	0.00	0.28	0.15	0.09	0.14	0.07	0.04	0.12
IRQ	0.11	0.00	0.11	0.18	0.12	0.56	0.00	0.20	0.12	0.07	0.00	0.00	0.27	0.00	0.29	0.17	0.00	0.00	0.12
ISR	0.13	0.12	0.09	0.18	0.14	0.09	0.13	0.06	0.18	0.08	0.08	0.22	0.16	0.08	0.15	0.21	0.08	0.01	0.12
JOR	0.17	0.10	0.30	0.25	0.13	0.37	0.14	0.09	0.16	0.25	0.09	0.33	0.29	0.06	0.21	0.29	0.00	0.06	0.18
KWT	0.26	0.50	0.29	0.30	0.29	0.47	0.18	0.18	0.26	0.00	0.20	0.24	0.35	0.25	0.28	0.30	0.00	0.30	0.26
LBN	0.28	0.38	0.38	0.11	0.15	0.13	0.22	0.15	0.08	0.29	0.18	0.25	0.18	0.36	0.25	0.16	0.27	0.57	0.24
LBY	0.11	0.16	0.07	0.53	0.05	0.35	0.27	0.10	0.09	0.04	0.21	0.25	0.26	0.13	0.10	0.09	0.17	0.06	0.17
MAR	0.42	0.18	0.15	0.11	0.18	0.20	0.14	0.25	0.06	0.11	0.11	0.00	0.23	0.00	0.15	0.20	0.11	0.15	0.15
OMN	0.15	0.00	0.38	0.40	0.39	0.47	0.46	0.00	0.12	0.17	0.00	0.25	0.08	0.40	0.33	0.25	0.20	0.03	0.23
PSE	0.02	0.00	0.04	0.08	0.09	0.05	0.07	0.00	0.04	0.02	0.25	0.00	0.05	0.50	0.13	0.03	0.00	0.07	0.08
QAT	0.21	0.63	0.34	0.28	0.19	0.49	0.24	0.60	0.18	0.00	0.63	0.25	0.21	0.23	0.34	0.09	0.05	0.56	0.31
SAU	0.36	0.35	0.13	0.14	0.16	0.32	0.26	0.45	0.13	0.05	0.10	0.16	0.34	0.19	0.15	0.28	0.29	0.16	0.22
SYR	0.18	0.00	0.02	0.49	0.19	0.14	0.29	0.11	0.05	0.00	0.25	0.00	0.09	0.06	0.38	0.06	0.00	0.00	0.13
TUN	0.02	0.06	0.50	0.22	0.00	0.50	1.00	0.00	0.29	0.00	0.06	0.00	0.17	0.07	0.05	0.50	0.12	0.00	0.20
YEM	0.11	0.00	0.20	0.08	0.20	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.20	0.00	0.00	0.17	0.16
USA	0.16	0.23	0.29	0.13	0.23	0.18	0.23	0.22	0.14	0.13	0.18	0.23	0.17	0.19	0.24	0.22	0.19	0.22	0.20
GBR	0.14	0.09	0.10	0.09	0.12	0.11	0.12	0.13	0.07	0.05	0.10	0.11	0.12	0.13	0.18	0.13	0.07	0.09	0.11
CAN	0.18	0.19	0.23	0.12	0.18	0.08	0.22	0.14	0.17	0.12	0.14	0.15	0.13	0.18	0.19	0.15	0.13	0.11	0.16
NZL	0.18	0.06	0.26	0.03	0.14	0.12	0.20	0.06	0.08	0.16	0.03	0.10	0.19	0.15	0.09	0.10	0.05	0.07	0.11
AUS	0.17	0.21	0.25	0.16	0.18	0.17	0.27	0.15	0.18	0.12	0.19	0.22	0.22	0.21	0.23	0.28	0.16	0.20	0.20
IRL	0.19	0.18	0.15	0.18	0.10	0.17	0.07	0.18	0.13	0.04	0.10	0.17	0.21	0.13	0.16	0.09	0.13	0.08	0.14
FRA	0.12	0.09	0.05	0.07	0.08	0.09	0.04	0.15	0.11	0.18	0.08	0.07	0.07	0.13	0.09	0.07	0.10	0.08	0.09
Max	0.50	0.63	0.50	0.53	0.39	1.00	1.00	0.60	0.29	0.29	0.63	0.33	1.00	0.50	0.38	0.50	0.29	0.57	
Min	0.02	0.00	0.02	0.03	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.05	0.00	0.00	0.00	

Table 6.5.3a. Ratio of reversions to edits between MENA countries.

6.5.4 Summary

The patterns of high edit to reversion ratio stands in for a multidimensional series of factors: vandalism, inexperience, politically charged content and lack of neutral point of view are all legitimate reasons to revert content. When we discussed earlier the notion of critical mass within countries, it is clear that this is an issue within the MENA region. As we noted in Section 6.4, editors from the MENA region are editing slightly less per view than elsewhere. Now we can see that they are also being reverted slightly more. Thus, visibility is not merely a matter of a lack of interest or a lack of will but also external forces that are signaling to editors in the MENA region that their content is not as legitimate.

To make legitimate content that sticks is a part of the acculturation process on Wikipedia. Without broad community support it is difficult for editors to naturally understand what constitutes legitimate content, and for editors to blow off steam when their content is being reverted. As Dr Tarawneh said about the Arabic Wikipedia in Section 2.1.3, many authors move over to the Arabic Wikipedia to edit when their content is reverted in the English or French Wikipedias to "blow off steam". This is not necessarily a positive outcome as it simultaneously pushes local editors away from contributing to the dominant site of geographic representation while fostering content that might not be as neutral or fair in Arabic as it is in English.

Part of this issue is simply a matter of misunderstanding. As was very evident in our Amman workshop, there is a great deal of misunderstanding about how to fairly edit about neutral topics. People tend to get very emotional about certain topics and delve into the minutae of arguments on a matter of principle. Partly, this is a psychological matter. As noted in the social psychology of computer-mediated communication, text-based interaction on the Internet can be very 'hyper-personal', meaning that individuals infuse a great deal of emotion in content in the absence of the sort of social cues found in everyday life (Walther 2007). When this happens on Wikipedia, it can lead to intense and tedious reversion battles. Also, when one consistently sees such reversions happening from British Wikipedians, or content that is not of the appropriate quality from Americans, one can infuse this content with a sort of emotional power, when in fact it is a misunderstanding.

We strongly recommend additional face-to-face interactions among Wikipedians in order to help minimize this sort of hyper-personal tit-for-tit controversy and the building and maintenance of community structures that allow people to better understand the cultural norms on Wikipedia rather than assume bad faith.

Section 6.6 – Perspectives from Wikipedians in the MENA region

The vast majority of the analysis to this point have been quantitative work drawing upon data from Wikipedia. However, as noted in our project activities section, a significant portion of the project activities were dedicated to considering the voices of Wikipedians themselves both as a form of research and as a form of capacity building within the region. Although it was merely implicit in the other analysis in Section 6, these voices have contextualized a substantial portion of the direction of our work, including an interest in focusing on a lack of critical mass, policing by external Wikipedians, and a focus on local content.

Drawing on a wiki questionnaire, interviews, and focus groups with editors, this section outlines some of the primary justifications for the spatial unevenness in content. While our quantitative work has demonstrated that, in general, production on Wikipedia tracks very closely to diffusion of broadband, the MENA region is a particular outlier for in terms of content production. That is, given the current level of broadband diffusion within the region there is typically less edits than expected, either in the English Wikipedia or in all Wikipedias combined.

We do not believe that there is a single explanation for this lower-than-expected level of content production, and the editors tend to agree. However, neither do editors believe there is a single magic bullet that will lead to an increase in development of increased and more representative geospatial content. Instead, we saw a range of barriers that together served to hinder the creation of content. Some of the issues identified can be rectified with more effective local organization and more interaction among active and potential editors in the region. Other issues point to longstanding cultural trends, and other issues still are in the process of being rectified, such as the lack of accessible technology for typing in Arabic.

Below we offer brief summaries of some of the core barriers before exploring some of the specific stories told us by participants.

Infrastructure: This includes issues around internet access, but also things like the difficulty of Arabic keyboards; and crucially importantly a relative lack of digitized secondary sources to draw from.

Literacy and education: Literacy is a necessary pre-condition for editing on Wikipedia. But herein we also refer to digital literacy. Digital literacy includes the ability to critically evaluate sources, locate requisite Wikipedia policies, include references in the often tedious Wiki syntax, and confidently assert one's statements rather than assume the other person is correct simply because they are more argumentative.

Institutional support from Wikipedia: Wikipedia's policies often make it particularly difficult to include pictures in the Wikimedia commons. While Arabic internet users are very active users of Facebook, photos of the MENA region on Facebook are not suitable for use on Wikipedia, and there is little to no work in trying to fix this. Wikipedia also uses fonts in Arabic that many editors consider to be difficult to read and a poor choice for Arabic script. This is likely to dissuade people from using Wikipedia. Finally, Wikipedia is often caught in between local political struggles. For example, many editors in the highly popular Arabic Wikipedia believe that the choice to create a secondary "Egyptian Arabic" Wikipedia as a snub and inappropriate.

Governance and community: Here there was a lot said about some of the often quite vicious edit wars that go on in Arabic Wikipedia. Because of this, the Arabic Wikipedia is heavily moderated, meaning that anonymous edits in Arabic Wikipedia do not automatically appear instantly. This is in contrast to most popular Wikipedias such as the English or German Wikipedias. Such a practice has led to the failure of heavily moderated alternatives to Wikipedia such as Citizendium and Google Knol. In the larger Wikipedias, reverting and use of administrative tools such as protecting articles are seen as a reasonable compromise. However, the administrators of the Arabic Wikipedia believe that this will not work for them.

There also seems to be a culture of administrators deleting articles that they do not think are locally appropriate. These often relate to politically sensitive topics.

Finally, there is also an issue of internal governance. A few editors felt that people from outside the region tended to win arguments and debates about topics they knew relatively little about, and that the rules and structures of the encyclopaedia encouraged this.

Culture: Wikipedias reflect the culture of those who edit them. The English Wikipedia and especially the German Wikipedia are very bureaucratic but also very tolerant. They seek compromise through arbitration. The Arabic Wikipedia similarly reflects many of the cultural values and issues of the Arabic world, as does the Farsi and Hebrew Wikipedia.

In general, our team did not push very strongly on this issue for fear of being seen as judgmental. Nevertheless, many of the editors were quite open about what they consider to be cultural issues. Several such issues came up. For example, several editors believed that the Arabic world was characterized by an education system that encouraged rote learning thereby feeding into a notion that only experts ought to edit a topic rather than amateurs with local familiarity.

In a more positive light, many editors considered the project of an Arabic Wikipedia to be a force to help signal how Arabic people across different countries and cultures could come together. Editors also talked about the notable gender disparities on the site. That said, a lack of women editing Wikipedia is a longstanding issue in parts of the world characterized by less stringent gender norms.

Politics: Editors had revealing stories about meddling from state actors. People point at both the fear of local authorities (we had stories from both Tunisia [with the Arabic Wikipedia] and Iran [with the Persian Wikipedia] about this); and a general lack of trust with a platform seen as an outside, foreign, and sometimes explicitly American tool.

Language: Here people spoke about the power of communication in English and French, and the reach afforded to anyone doing so. The many quite different dialects of Arabic also seem to act as a hindrance to people in some cases. Additionally, language issues multiply issues of governance as many of the extensive policy discussions in other Wikipedias have not been translated, leading to a replication of such discussions or ambiguity about the appropriate forms of conduct in case of disputes.

Below we outline some of these themes in more detail.

6.6.1. Governance and Sourcing

Perhaps the greatest barrier to content production is a lack of vetted sources for content. Wikipedia is meant to be a forum for the aggregation and summary of external sources, and not a site for primary material. It is very common to see articles tagged with "citation needed" or deleted because the article does not meet notability requirements.

In order to legitimate content on Wikipedia, editors look to external sources and particularly government sources and news media. In both cases there is a paucity of *legitimated* information in the MENA region. As one editor noted: "...[V]ery few magazines and journals relating to that topic are available online, and I honestly think the number and quality of major news sources were rather less impressive back then than now as well". For example, editors detailed how citing Al Jazeera was up until a few years ago considered to be insufficient as many Western Wikipedias considered Al Jazeera to be ideologically biased (whether or not this is the case).

This is compounded with the twin issues of authority/culture and language. If a source exists in English and an English-speaking author feels especially passionate about the topic, it is likely that the claim drawn from that source will persist. But if the source is in another language it is already difficult to persuade sceptical editors that it is a legitimate source or that it properly reflects what is claimed in the article. It is compounded with authority/culture as alluded to above in that native English speaking editors may make strong assertions, even if ungrounded, but be more persuasive due to a culture of adversarial debate rather than respect for / acquiescence to authority as found in Arabic cultures.

The combination of these issues was well expressed in one editor's story of Heliopolis. Very little survives from the Ancient city of Heliopolis. One of the largest remaining artefacts is the Al-Masalla obelisk. The editor created an article about the city and gave the coordinates of the obelisk. Later, another editor (from America) moved the geocode location of the obelisk several miles away, citing a book published in the 1930s with the location of the obelisk. The editor retorted by saying that he knew where the obelisk ought to be since it was right outside his window. Yet, because American editor had a source material and the Egyptian editor did not, it took some time and help from additional Wikipedians to rectify the geocode.[32]

A secondary issue with sources that was alluded to in the discussions was a lack of official government sources and open content. Whether it concerns statistics on health, population or demographics or whether it is the presence of official maps and gazeteers, if these sources are not easily accessible on the web, then Wikipedians cannot draw upon them for the encyclopedia. As an editor noted, "a successful article needs sources and interested editors. Sources are abundant for places in the Western World. The same cannot be said about developing countries".

When an editor is especially tenacious he or she can overcome this issue with much hard work. We asked one editor if he drew upon a great deal of sources for his favourite article. "[I]n the case of the economy of Iran the answer is a clear 'NO'. The US Library of Congress said in 2008 that there was no comprehensive coverage of this topic in *any* language. I had to start from scratch, with motivation to learn for myself and share my findings with others as my sole motivation and resources. Little by little, I was able to grow this topic to a large book (~ 1,200 pages approximately). Since then, other books have been written but in my honest opinion none of them

[32] This story can be confirmed by looking at the talk page for Heliopolis: http://en.wikipedia.org/wiki/Talk:Heliopolis_(ancient) where Cairo workshop attendee Wikipedian "Ashashyou" discusses this issue.

(to the best of knowledge) comes close". Having to research such a wide ranging topic with little source material is not a task many editors want to face. This is especially the case for articles in areas that may be obscure. As one editor noted, "It can be incredibly resource-intensive to write a good Wikipedia article. You need plenty of time and lots of good reference sources. That being the case the articles that develop well tend to be in subject areas that can attract a group of knowledgeable editors who can collaborate reasonably well. I would guess that some subject areas just don't attract the critical mass of editors needed".

We believe that there is much room for knowledge sharing with MENA governments about open access to knowledge. We believe that Wikipedia is an excellent example of how a small number of people can repurpose official open access documents for the benefit of not just the local citizenry but for the rest of the world. The issue with sourcing highlights how Wikipedia is a system for the organization of public knowledge, not a site for the generation of this knowledge. It needs fuel in the form of official and validated sources to move beyond superficial coverage of well known sites towards deep coverage of local sites.

6.6.2 The (lower) prestige of editing in Arabic

Editors both in the Cairo workshop and on our wiki noted that there is somewhat of an image issue with editing in the Arabic Wikipedia. That is to say, editors would consider editing in the Arabic Wikipedia as unglamourous because it would always be overshadowed by the dominant English Wikipedia.

This has led to what we consider some ironic unintended consequences. Presently, there are multiple efforts afoot to translate articles from the English Wikipedia to the Arabic Wikipedia, but no such efforts in the other direction. From the naïve point of view that it is good to have representation in Arabic, such initiatives appear to be worthy and in good faith. However, the unintended consequence is that it leads to editors believing that the Arabic Wikipedia is merely going to be a shadow or poor translation of the English Wikipedia. As one editor suggested "*Articles on the places in the Arab World are just a subset of the total number of articles in Wikipedia, and the English Wikipedia is vastly larger than its Arabic counterpart, so it is not unthinkable that there is more content, even about Arab-world subjects, in English. From my (unscientific) observation, many times, content in Arabic about a place or a tribe is not very encyclopedic, but promotional, and lacks citations*".

In this respect, translating articles to Arabic then becomes seen as menial and unrewarding work, when the exciting debates about an article are happening elsewhere.

We would actually recommend in this regard an unconventional strategy of mining and discovering articles in Arabic (or Hebrew or Persian) that are not in the English Wikipedia, translating them to English and promoting this activity among Arabic Wikipedians. In doing so, we believe that such activities will help signal that there is local knowledge from Arabic speakers that they should be proud of that can be shared with the world. Doing so may also might help to reinforce the sentiment expressed by one editor about the contradictions of the Arab world as irritated by cultural misunderstandings or ignorance from outsiders and pride at the past accomplishments of the Arab World, while signalling how the rest of the world can learn from the MENA region: "*MENA is obviously important in geopolitical terms and attracts a great deal of interest from those concerned with contemporary manifestations of the Great Game and the associated claim that the Arab world and Islam pose an existential threat, but beyond that it's fascinating to anyone with an interest in the history of mathematics, astronomy, medicine, chemistry, physics, cartography and architecture. Most of all though the Arab world is close to the*

West geographically, but remote intellectually and ideologically. It's a fascinating juxtaposition with an obvious attraction to those interested in the history of civilisation".

6.6.3 The (unhelpful) governance structure of Wikipedia

Wikipedia has been jokingly described as the opposite of communism: "it works great in practice and terrible in theory". One of the key ways in which Wikipedia succeeds is through its extensive consensus-based governance structure. However, this structure is neither obvious or immediately accessible. Moreover, different Wikipedias from Arabic to Chinese to Afrikaans all have local rules, norms and quirks that differentiate them from each other. Thus, one might have to learn additional rules when switching languages.

While the governance of Wikipedia is meant to enable content that is high quality and fair, it also puts up barriers for those seeking an entry into this project. Long time editors know specific and often tedious rules, know which rules trump other rules based on past disputes and which administrators are likely to sympathetic to which kinds of arguments when resolving disputes. These forms of cultural capital can sometimes steamroll over good intent and push out otherwise willing editors. As one editor noted: "*Consensus can be used to make decisions on Wikipedia. While this is normally a really good way to come to a group agreement, on Wikipedia I think there can be problems. I've seen examples where only a few editors who have knowledge in another era or area of knowledge, but have little or no knowledge of the area under discussion, can form a negative consensus*".

This adherence to rules over good faith can lead to territorial editors using the rules to marginalize or exclude content. Project member Ms. Ford previously wrote about this in relation to Africa using the example of local folk hero Makmende. An editor at the Amman workshop introduced a very similar case in his own editing: "*Yes I have faced a lot of challenges in writing the articles that contain a little notability. I would like to share one example. I have written articles on supertall skyscraper namely, Marina 101, that were planned to be built in Dubai [in 2009]. The article was very soon tagged that says (I am explaining precisely) 'This article has very little notability, and will be deleted upon consensus of other editors*'." Reviewing the history of the article (now called "Dream Dubai Marina", one can see an editor from the UK, Astronaut, expressing scepticism about the legitimacy of the article and another suggesting that more photos are needed, while another, anonymous editor, complaining about the use of the phrase "supertall". We can also see that workshop attendee Nabil_rais2008 has done substantial work on this article and persevered. At present, the article is now considered of 'high importance' within Wikipedia's group on skyscrapers, even though it was charged with not being notable when it first started. Nabil was a tenacious and dedicated editor, but it is clear that other editors were being overly dismissive and less experienced editors would likely have been extremely intimidated by the way fellow Wikipedians wielded their authority and often with haste.

Unfortunately, this is not an issue that can be resoled outside of Wikipedia, but one that must come from within. However, we believe that two things can be done. The first is in further capacity building. Nabil's story was not uncommon. In fact, informally observing the workshop participants, both P.I.s noticed that the editors were frequently exchanging strategies on how to deal with tedious and recalcitrant administrators. This sort of community building helps to provide the support editors require when faced with direct opposition. The second is in providing translations of materials on governance. Wikimedia already provides such materials, but they are often buried and difficult to locate. Funding of school-based efforts and Wikipedia curricula can circumvent this.

6.6.4 The (unhelpful) technical architecture of Wikipedia

One issue that kept coming up among editors was the notion of emotional energy. As mentioned earlier about the hyperpersonal model of online interaction, it is fair to say that editing can be a draining and frustrating task. From the outside looking in Wikipedia almost appears to have emerged *ex nihilo*. Yet, inside there are many heated disputes and arguments, where often the editor with the greatest reservoir of tenacity and most connected peers wins out, not the editor with the best argument.

This is in contrast to a site that claims to be open and accessible to all. As one editor noted, "[*g]etting started on Wikipedia is very difficult with so much to learn about style and markup. Not all advice is accurate either and reading all the directions can be confusing and difficult to understand. Would it be possible to develop an interactive video to talk new editors through the first few edits and another to build on the basics later? I'm still getting used to all the different quote boxes and etc. Not to mention all the tags etc."*

Between the fact that the site's fonts are not the best Arabic fonts, something noted by participants in both workshops, that the policies are varied and that help items are scattered and not obvious when an editor clicks "edit", it appears that Wikipedia is more unwelcome than it portends to be.

Wikimedia themselves recognize this and have spent significant amounts of time and effort on a WYSIWYG editor for the site. Currently it is available in a few Latin languages. It was met with outrage among the community because it undermines many of the carefully organized infoboxes, templates and other niceties that benefit from working within the Wiki syntax. We believe their efforts might have been better served in a rethink of the user interface for editing and the creation of additional affordances for edting *especially from mobile devices*.

6.6.5 Systematic bullying efforts and (non)neutrality

As noted in Section 6.5, editing Wikipedia can be a tough challenge. Editors from some countries have had almost every edit reverted. Edits to articles about Israel are especially volatile and countries vary significantly in the extent to which editors will produce the sort of edits that are in keeping with Wikipedia. When almost one fifth of all edits are reverted, it is clear that there is a serious issue of regulation happening.

Wikipedia editors may consider this a just process that is required to ensure Wikipedia maintains a high quality presentation of reference material. But it is very likely that it goes beyond this as editors feel possessive of their content and of the sort of views that they consider appropriate. Such territoriality has made international headlines. For example, User:NewsAndEventsGuy was featured in Popular Science for his aggressive assertions that climate change had no impact on the devastating Hurricane Sandy despite much evidence to the contrary.[33] But it can go further than this as editors with power arbitrarily ban other users, protect articles and refuse moderate points of view. Essentially, editors can bully.

On Wikipedia many editors here claim to have been the victim of concerted efforts to minimize or undermine their activity. For example (from three different editors from three different countries:

[33] http://www.popsci.com/technology/article/2012-11/wikipedia-sandy

"*As a matter of fact, YES. I have been bullied (if that is the right expression here) by a group of Jewish and Massonic editors (as per their user pages' info) when I tried to edit important articles, and by using only reliable resources such as the Financial Times or similar sources*"

as well as:

"I understand your anger, in my case, I was heavily bullied, BANNED :) , by a group led by this Saoudi guy who happened to be a BUREAUCRAT! including medical doctors in that group, ha!"

and,

"*A number of articles I have edited with quality sources, have been subjected to editors cutting information that doesn't fit their ideas or they add information which changes the meaning without adding a source. These activities are more of a waste of time; instead of going on to other things I spend a lot of time going back to reinstate information. Today's examples are in the 'Battle of Nablus (1918)' and the 'Third Transjordan attack' articles. Bullying does occur from time to time to the summaries attached to some of these edits. Having tried the disputes process I wouldn't recommend it. So I don't know what can be done to fix this particular problem*".

In the last extensive quote we can observe the relationship between sourcing issues and bullying, where some individuals dismiss sources that ought to be considered based on the neutral criteria, but are excluded because they are perceived by some in power as not conforming to a specific ideology.

This is one issue where making recommendations is especially challenging. We cannot make recommendations to Wikimedia in particular and as people from Wikimedia note, the community themselves often have a great deal of power and are quick to dismiss many of Wikimedia's recommendations. To this extent we maintain an emphasis on community building. That is, moving away from exclusively online interaction has in our workshops led to greater understanding. It is likely that this will continue to be the case. Wikipedians are remarkably generous with their time. Their efforts are helping to document much of the world where documentation is scarce. But if they are scared off by unruly mobs who 'weaponize' policies to fit a specific agenda, then it is a loss for everyone interested in a more extensive and accessible means for learning about the world. One editor notes that he often sees "*editors parroting sensationalist rhetoric while conveniently leaving out (or deleting) essential information that ultimately leads readers to reach false conclusions*". This becomes very disheartening for the editors and the consequences are a loss for everyone else.

6.6.6 Surveillance and state intervention

Within the West there have been scandals of state actors writing on Wikipedia pages to clean up text, to minimize information on a certain topic or to skew a page in specific ways.

Several editors preferred to speak to us off the record about the specific details of such meddling, but noted that officials of different states knew that they were active on the Arabic Wikipedia and asked them how to remove unflattering or critical content. Several editors noted being approached by the Syrian government, in particular, about how to identify the editors who are making these critical comments. This fact actually strongly reinforces the legitimacy of Wikipedia's policy not to keep the IP addresses of logged in editors for more than 90 days.

Beyond this is a chilling effect of rumors that several states, particularly, Israel and Iran have concerted efforts to work on content on Wikipeda. As we have seen from the editing practices in section 6.4 and 6.5, if this truly is the case, then it is clear that these actors are not interested in

undermining each other's content so much as cleaning up content about their country. While we can neither confirm nor deny that these practices exist, the mere mention of them creates a chilling effect for new editors who might feel that editing certain pages is simply intractable.

We can only make the most modest recommendations for development in this regard. We believe that to undermine this chilling effect Wikimedia should demand that any state actor is clearly labelled. We also recommend the broader diffusion of privacy-enhancing technologies and knowledge of how they work with and through Wikipedia. For example, while one cannot create a Wikipedia account through TOR, the anonymizing browser, one can create such an account, leave it dormant for 90 days and then only edit through TOR. This practice is not widely known despite the fact that it might be particularly useful for editors working in sensitive areas.

6.6.7 A counter point to the barriers

Between such barriers, from poor fonts to bullying to surveillance, it is a wonder editors bother with Wikipedia at all. Yet, the editors we spoke to considered Wikipedia as not merely a hobby but as a project in the grandest sense. Deeply embedded actors were especially optimistic about the site and considered it not merely a self-serving activity but one that is meant to serve the greater good, within the Arab World, the broader MENA region and the world in general. While many Internet evangelists have also spoke about sites with unbridled enthusiasm in this case, Wikipedia has the success to back it up. It *is* one of the most significant cultural forces on the internet and it does embed itself into the core technologies and digital information spaces of everyday life. It is used by journalists, investors, tourists, and students. We consider it one of the largest value adds of the Internet age.

In the initial grant document we came to this project with a sense that this exclusion was unjust and that editors were probably working to rectify this situation. What we came to realize is that editors do not merely want to self-represent for its own sake, but because they consider it a path to a more just society. Editors in the MENA region said that while there is exclusion and bullying on Wikipedia, it pales in comparison to actual physical and armed conflict. Some went so far as to say Wikipedia could be an engine for teaching the process of democritization through consensus, voting, and an aspiration to neutrality.

These editors see Wikipedia as both a potential site for conflict and also a site for pride. As one editor noted, "*Everyone has equal rights on Wikipedia to edit the article, but within the limits of the rules and policies of Wikipedia. e.g: If someone is adding new information than he has to provide the references for citations in the article*", but also on the same subject "*It would be frustrating if the information is incorrect or not sourced, but otherwise I am okay with anyone editing an article about my home town, regardless of his location*".

Another editor (from Egypt) spoke of how he met and became friends with several Syrians through Wikipedia, saying grimly "*I have good remember and dear friends in Syria before it's civil war*". He has not heard from them for some time.

Editors forge bonds and express pride in self representation. As one notes, "*[t]here is several articles I have edited regarding the region, and I appreciate them so much because they are related to my own culture, and because I had put a great efforts*". Or as another flatly stated when asked why he writes geographic content, he said because "It's my own town".

7. Overall Assessment and Recommendations

We end this report with four summaries: (1) a short reflection on our links to Canadian researchers and institutions; (2) our reflections on what we would do differently; (3) the capacity building that emerges from our work; and (4) a summary of the overall value of this project.

7.1 Canadian partnerships

While this project was not designed as one to make Canadian partnerships a central feature of the work, we should nonetheless highlight four important linkages to Canada and Canadian institutions.

- Ilhem Alagui, a core project partner is a joint Canadian/Tunisian citizen and has a PhD from the University of Montreal (she is now a faculty member in the UAE).
- Bernie Hogan, the Co-investigator of the project, is a Canadian with a PhD from the University of Toronto.
- Ali Frihida, an early project partner also has a Canadian PhD.
- Mark Graham spoke about the project findings at a 2012 conference organised by the Canadian Security Intelligence Service (CSIS).

We, however, see the core value of this work in the sections below.

7.2 The project through the lens of a time machine

We would consider doing a few things differently if we were to be offered the opportunity.

First, much of the time invested in the project was devoted to iterative analysis of datasets created with bespoke software. This is because we repeatedly encountered situations and configurations of data that we had not anticipated. In other words, mapping Wikipedia was a slow and messy affair.

However, we have learnt from our mistakes, and our final results (published as both academic papers and freely downloadable code and data) have transcended many of the specific problems that we encountered. These 'best practices' would allow anyone emulating our methods to circumvent challenges that we faced.

We have also learnt much from our outreach meetings and attempts to work with Wikipedia editors based in the Middle East that might have altered some of our recruitment strategies. Specifically, we faced significant hurdles convincing Wikipedians from the MENA region to trust our academic intentions. Here we ultimately discovered that by constructing a project wiki, we could spark some initial debate and allow potential participants to better understand the themes and questions that we wanted to address. Furthermore, by using a wiki that was in both Arabic and English, we were essentially 'speaking the same language' as other Wikipedians. We would encourage other

researchers studying participation to similarly slowly build trust using platforms and languages as the communities that they wish to better understand.

On the technical side, we now better appreciate many of the complexities involved in text mining and natural language processing. Several of our collaborators had sold us on the power of experimental software. This meant that the project had to serve as both a testing ground for this software as well as a substantive research project. Being more sceptical about such software and weary of those who would oversell things based on very preliminary results would have made it easier for us to estimate the actual work involved.

On the political side, we did come into this project knowing that specific political conflicts in the MENA region would be a salient factor that involves tact and care. However, we were surprised that even some of our collaborators at academic institutions felt so strongly about the involvement of people from Israel that it made working either with Israelis or on the Hebrew Wikipedia challenging. Although we would not change our analysis, in the future we would be more up front about the need to engage both Arabic partners and Israeli partners from the outset.

7.3 Capacity building

Our capacity building has come in a few forms. Some are specific and traceable; others are broader, but less quantifiable.

Some of work has had direct impacts on organisations attempting to address these informational divides. For instance, the WikiAfrica project, has taken our work as a starting-point to begin their outreach work (as explicitly stated in a WIPO magazine article from June 2013).[34] The WikiAfrica project is specifically designed to address the paucity of information about Africa in Wikipedia by facilitiating and encouraging the creation of 30,000 new articles about the continent. As our maps earlier in this report demonstrate, this is still a relatively small number, but would be a vast improvement on the current situation. Our maps feature centrally on the project's homepage (see Figure 7.3a below) and we have written letters of support to help the project sustain itself.

[34] http://www.wipo.int/wipo_magazine/en/2013/02/article_0006.html

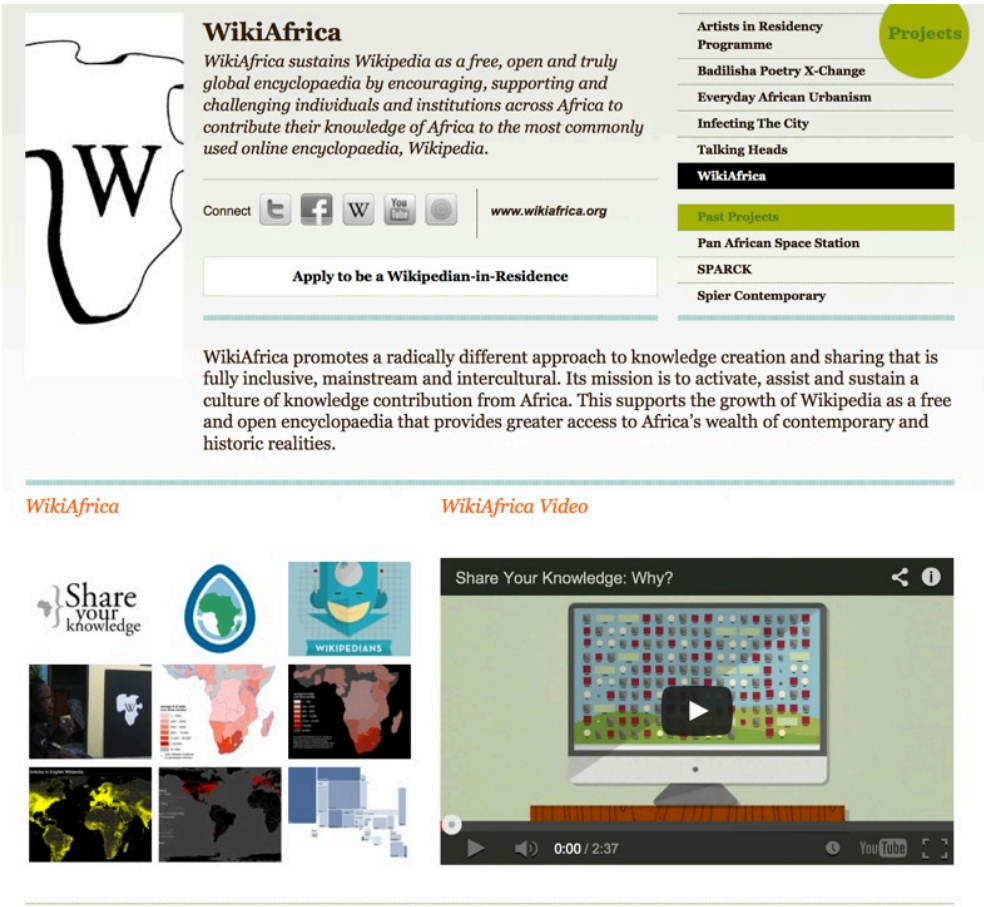

Figure 7.3a. A screenshot of WikiAfrica <http://www.africacentre.net/wikiafrica>. Dec 13, 2013. Note that our maps are prominently featured in the lower left corner.

More broadly, our work and results appear to have become relatively well-known to the Wikipedia community. For instance, Sue Gardner, the out-going Executive Director of the Wikimedia Foundation, has mentioned our work as an impetus for the creation of more content about the world's economic margins (for instance in an 2012 interview with CBC's 'Spark'). Jimmy Wales, a co-founder of Wikipedia, also regularly refers to global inequalities in Wikipedia (very little other empirical work other than ours points to these global inequalities). Wikipedia's Global Development team has also drawn on some of our outputs (specifically our maps) to stress the need for a less Western-centric Wikipedia (in this case, they specifically contacted us to request our maps [and new/bespoke maps] of the encyclopaedia).

In addition to presenting project work to the Executive Director of the Wikimedia Foundation and their core outreach and research team at the Wikimedia headquarters in San Francisco, Dr Graham has had repeated follow-up discussions with staff at both Wikimedia Headquarters and Wikimedia UK about the uneven geographies of content contributed to the user-generated encyclopaedia. This work has informed some of the outreach carried out by Wikimedia Foundation in the developing world. For example,

project results were used in a keynote speech by Wikimedia's Chief Development Officer, Barry Newstead, at the 2011 Wikiconference India[35], and these results were similarly used to inform the Wikipedia Arabic Catalyst[36] project (an initiative to address the paucity of Wikipedia content from the Middle East). Our work is also repeatedly featured in Wikipedia blogs and newsletters in multiple languages[37].

Our 2012 Cairo-based and 2013 Amman-based workshops for Arab Wikipedians might also be considered key capacity-building exercises. Workshop attendees were all key contributors to the Arabic Wikipedia and the workshop brought them together to talk about barriers to participation from the Middle East. As a result of the workshop, the attendees now have an active Facebook group (which has been joined by many other Arab Wikipedians) and a community wiki (http://menawiki.oii.ox.ac.uk/) on which they discuss topics of concern (both of which were set up by our team). That Facebook group, in turn, grew into a more ambitious collective who have called themselves 'Wikipedians Without Borders.'

The combination of this research and public outreach related to Wikipedia resulted in the Oxford Internet Institute being awarded (by Wikipedia UK) a 2012 *UK Wikipedian of the Year Award* for being the "Educational Institution of the Year."

Our work has contributed not just to the debate about inequalities in Wikipedia, but also to more general conversations, debates, and strategies about information inequalities. Some of these contributions have taken place through lectures that we gave to influential organisations such as:

UNCTAD (CSTD), Dec 2013

Internet Africa Summit (Lusaka, Zambia), June 2013

DFID Ministerial Advisory Seminar, May 2012, March 2012

RaceOnline, March 2012

United Nations Student Association, March 2012

London Conference on Cyberspace/UNDP/Oxfam/DFID, November 2011

Wikimedia Foundation, October 2011

---

[35] http://meta.wikimedia.org/wiki/WikiConference_India_2011/Programs

[36] http://blog.wikimedia.org/2011/10/04/wikimedia-foundation-to-launch-arabic-catalyst/

[37] E.g.:
http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Regards_sur_l%27actualit%C3%A9_de_la_Wikimedia/2011/46
http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2011-11-14/In_the_news
http://meta.wikimedia.org/wiki/Research:Newsletter/2012/April
http://news.yahoo.com/blogs/technology-blog/mapping-undead-invasion-201428007.html

The common thread in all of these talks has been our empirical demonstration of: (a) the presence of enormous information inequalities on the internet; and (b) the fact that these inequalities cannot be simply bridged by investing in greater connectivity. All of those meetings were relatively small settings in which ministers, policy makers, and executives listened and asked questions about our work.

Other presentations have been given at venues with more popular (and larger) audiences. This has included talks at *TEDx Life Online* (2012, Bradford, UK), *South by Southwest* (2013, Austin, USA), and *Re:Publica* (2013, Berlin, Germany), and a 90 minute *Department for International Development (DFID) seminar* that was also live-streamed to six DFID offices in Asia and Africa (2012, London, UK).

This was coordinated with our efforts to spread our findings widely in the media. It is here that we believe that we have done the most to influence public understanding of the Internet and its geographies. Our work has been featured in media outlets around the world including *The BBC, The Atlantic, The New York Times*, *The Guardian, The Economist, The Telegraph, Wired, Der Spiegel, Il Sole 24 Ore*. Mark Graham has also written a few articles which specifically highlight our findings in *The Guardian*. We regularly blogged about our findings on Graham's Internet geography blogs (zerogeography.net and floatingsheep.org) (which, in turn, were also frequently referenced in the media and by other academic blogs), and have now had over one million views.

7.4. Summary of our own perspectives on the value and importance of the project

This report has described our concerns, our questions, our methods, our findings, and ultimately our interventions. However, it might be important to end with a brief reflection on the core value and importance of this project.

First, this is work that has never before been carried out. We did not merely empirically investigate questions about voice, participation, and representation in, and from the Middle East and North Africa region , but also developed new methods to help us achieve those goals. Relatively little is known about contemporary geographies, patterns, processes, and networks of online participation (about anywhere in the world), and we were therefore able to make significant traction in a relatively uncharted area.

We were able to successfully bring these issues into the public and policy consciousness with work based on rigorous quantitative and qualitative research and data rather than speculation and conjecture. We witnessed the ways in which we have been able to actively change the debate about voice, representation, and information inequality amongst Wikipedians, amongst policy makers, and amongst the general public. We have highlighted some of the core barriers and constraints to generating information, and we have developed replicable methods to measure and interact with people who contribute user-generated content.

As the internet becomes integrally embedded into everyday life for a majority of humanity, it will become ever more important to understand not just what the internet enables, but also what it omits and who it excludes. This project has been a first step to understanding

some of these crucial issues, and we hope that future work can use our findings as a based from which to launch further inquiry into specific patterns and processes of visibility, representation and voice in the world's informational margins.

Bibliography

Ahern S., M. Naaman, R. Nair, and J. Yang. 2007. World Explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, 2007*, 1–10. New York: ACM.

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*.

Andrienko G., N. Andrienko, P. Bak, S. Kisilevich, and D. Keim. 2009. Analysis of community-contributed space- and time-referenced data (example of Flickr and Panoramio photos). In *IEEE Symposium on Visual Analytics Science and Technology, Atlantic City, 2009*, 213 –214. Washington, DC.: IEEE.

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

Backstrom L., E. Sun, C. Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web, Raleigh, 2010*, 61–70. New York: ACM.

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

Bosca A. and L. Dini. 2011. Automatic gazetteer generation from Wikipedia. In *LNCS 6699: Advanced Language Technologies for Digital Libraries*, Italy, 2009, 61–71. Berlin: Springer.

Brandes, U., & Lerner, J. (2008). Visual Analysis of Controversy in User-Generated Encyclopedias*. *Information Visualization*, *7*(1), 34-48.

Brandes U., P. Kenis, J. Lerner, and D. van Raaij. 2009. Network analysis of collaboration structure in Wikipedia. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web, Madrid, 2009*, 731–740. New York: ACM.

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.

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.

Cheng Z., J. Caverlee, and K. Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, 2010*, 759–768. New York: ACM.

Craig, W. J., & Elwood, S. A. 1998. How and why community groups use maps and geographic information

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.

Davis C. A. Jr., G. L. Pappa, D. R. Rocha de Oliveira, and F. de L. Arcanjo. 2011. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS* 15(6): 735–751.

Davis, A., Gardner, B. B. and M. R. Gardner (1941) *Deep South,* Chicago: The University of Chicago Press.

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

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

Flöck F., D. Vrandečić, and E. Simperl. 2011. Towards a diversity-minded Wikipedia. In: *Proceedings of WebSci '11, Koblenz, 2011*, 1–8. New York: ACM.

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.

Geiger, R. Stuart, and Aaron Halfaker. "Using edit sessions to measure participation in wikipedia." *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900-901.

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

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

Graham, M. 2011b Wiki Space: Palimpsests and the Politics of Exclusion. In *Critical Point of View: A Wikipedia Reader*. Eds. Lovink, G. and Tkacz, N. Amsterdam: Institute of Network Cultures, 269-282.

Graham, M. 2013a. 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. 2013b. The Virtual Dimension. In Global City Challenges: Debating a Concept, Improving the Practice. Eds. Acuto, M. and Steele, W. London: Palgrave (in press).

Graham, M. 2014. The Knowledge Based Economy and Digital Divisions of Labour. In *Companion to Development Studies, 3rd edition*, eds v. Desai, and R. Potter. Hodder. 189-195.

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

Halavais, A., and D. Lackaff. 2008. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication* 13(2) 429-440.

Halfaker, A., Kittur, A., & Riedl, J. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (pp. 163-172). ACM.

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.

Hollenstein L. and R. S. Purves. 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 1: 21–48.

Holloway T., M. Bozicevic, and K. Börner. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity* 12(3): 30–40.

http://www.itu.int/wsis/docs/pc2/visionaries/lessig.pdf (last accessed 17 April 2013).

Ito M., K. Nakayama, T. Hara, and S. Nishio. 2008. Association thesaurus construction methods based on link co-occurrence analysis for Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, 2008*, 817–826. New York: ACM.

Jankowski P., N. Andrienko, G. Andrienko, and S. Kisilevich. 2010. Discovering landmark preferences and movement patterns from photo postings. *Transactions in GIS* 14(6): 833–852.

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.

Kittur A., B. Suh, B. A. Pendleton, and E. H. Chi. 2007. He says, she says: Conflict and coordination in Wikipedia. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, 2007*, 453–462. New York: ACM.

Kriplean T., I. Beschastnikh, D. W. McDonald, and S. A. Golder. 2011. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work, Sanibel Island, 2007*, 167–176. New York: ACM.

Kruskall W. 1987. Relative Importance by Averaging over Orderings. The American Statistician 41(1), 6-10.

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).

Lessig, L. 2003. An Information Society: Free or Feudal (talk given at the *World Summit on*

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.

Milne D., O. Medelyan, and I. H. Witten. 2006. Mining domain-specific thesauri from Wikipedia: A case study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, 2006*, 442–448. Washington, DC.: IEEE.

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

O'Brien R. M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673–690.

Ortega F. and J. M. Gonzalez-Barahona. 2007. Quantitative analysis of the Wikipedia community of users. In *Proceedings of WikiSym'07, Montréal, 2007*, 75–86. New York: ACM.

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

Purves R., A. Edwardes, and J. Wood. 2011. Describing place through user generated content. *First Monday* 16(9).

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.

Scellato S., A. Noulas, R. Lambiotte, and C. Mascolo. 2011. Socio-spatial properties of online location-based social networks. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, 2011*, 329–336. Palo Alto: AAAI.

Shirky, C. 2010. *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.

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).

Smets, K., Goethals, B., & Verdonk, B. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (pp. 43-48).

Stadler B., R. S. Purves, and M. Tomko. 2011. Exploring the relationship between land cover and subjective evaluation of scenic beauty through user generated content. In *Proceedings of the 25th International Cartographic Conference, Paris, 2011*, CO-062. International Cartographic Association.

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.

Walther, J. B. 2007. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior*, 23, 2538–2557.

Warncke-Wang M., A. Uduwage, Z. Dong, and J. Riedl. 2012. In search of the Ur-Wikipedia: Universality, similarity, and translation in the Wikipedia inter-language link network. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration, Linz, 2012*.

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

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

Zittrain, J. 2008. The future of the Internet. *London: Allen Lane*.

Appendix I. News articles regarding project output or featuring project members discussing Wikipedia.

Number. Title Source | Date | URL		
1. Wikipedia wants more contributions from academics		
The Guardian	29-Mar-11	http://www.theguardian.com/education/2011/mar/29/wikipedia-survey-academic-contributions
2. This Map Shows the World of Wikipedia Broken Down by Languages		
Gizmodo US	11-Nov-11	http://gizmodo.com/5858668/this-map-shows-the-world-of-wikipedia-broken-down-by-languages
3. The world of Wikipedia's languages mapped		
Guardian Datablog	11-Nov-11	http://www.theguardian.com/news/datablog/2011/nov/11/wikipedia-map-world-languages
4. Wikipedia Language Maps Created By Oxford Internet Institute's Mark Graham		
Huffington Post	13-Nov-11	http://www.huffingtonpost.com/2011/11/13/wikipedia-language-maps_n_1091241.html
5. Without Wikipedia, where can you get your facts?		
BBC News	18-Jan-12	http://www.bbc.co.uk/news/magazine-16601517
6. Wikipedia world: an interactive guide to every language. Infographic map		
The Guardian	4-Apr-12	http://www.theguardian.com/news/datablog/interactive/2012/apr/04/wikipedia-world-language-map?newsfeed=true

6. OII Recognised as Educational Institution of the Year at Wikimedia UK's Annual Conference

Oxford Internet Institute	15-Jun-12	http://www.oii.ox.ac.uk/news/?id=715

7. Geography, Big Data, and Augmented Realities

Oxford Internet Institute	1-Aug-12	http://www.oii.ox.ac.uk/news/?id=736

8. Twitter Map Predicts 2012 Presidential Election: Will It Be Right?

Huffington Post Technology (US)	6-Nov-12	http://www.huffingtonpost.com/2012/11/06/twitter-map-predicts-election_n_2082000.html

9. Election 2012: Twitter map predicts presidential race results

Syracuse.com	6-Nov-12	http://www.syracuse.com/news/index.ssf/2012/11/election_2012_twitter_map_presidential_race_results.html

10. Big data and the death of the theorist

Wired	25-Jan-13	http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory

11. Mathematical model 'describes' how online conflicts are resolved

University of Oxford	20-Feb-13	http://www.ox.ac.uk/media/news_stories/2013/130220.html

12. Who Writes the Wikipedia Entries About Where You Live?

The Atlantic	26-Mar-13	http://www.theatlanticcities.com/technology/2013/03/who-writes-wikipedia-entries-about-where-you-live/5085/

13. Free for all? Lifting the lid on a Wikipedia crisis

New Scientist	17-Apr-13	http://www.newscientist.com/article/mg21829122.200-free-for-all-lifting-the-lid-on-a-wikipedia-crisis.html?full=true#.Ulu38lA3uCk

14. Catalan Wikipedia Reaches 400,000 Article Milestone

Global Voices	19-Apr-13	http://globalvoicesonline.org/2013/04/19/catalan-wikipedia-how-to-place-a-stateless-nation-on-the-global-network/

15. Why Wikipedia's Millionth Russian Page Is Worth Celebrating

Simulacrum	11-May-13	http://simulacrum.cc/2013/05/11/why-wikipedias-millionth-russian-page-is-worth-celebrating/

16. Gütesiegel für Wikipedia

Technology Review	13-May-13	http://www.heise.de/tr/artikel/Guetesiegel-fuer-Wikipedia-1836701.html

17. OPINIÓN: El acceso generalizado a internet, ¿es una meta alcanzable?

CNN Mexico	17-May-13	http://mexico.cnn.com/opinion/2013/05/17/opinion-el-acceso-generalizado-a-internet-es-una-meta-alcanzable

18. □基百科不自由 (Wikipedia is not free)

Caijing.com.cn	21-May-13	http://column.caijing.com.cn/2013-05-21/112805891.html

19. The Controversial Topics of Wikipedia

Wired Science Blog	30-May-13	http://www.wired.com/wiredscience/2013/05/the-controversial-topics-of-wikipedia/

20. Wikipedia 'Edit Wars': The most hotly contested topics

Live Science	31-May-13	http://www.livescience.com/37034-wikipedia-controversial-topics.html

21. The Most Controversial Article in all of English Wikipedia is George Bush's

The Huffington Post	31-May-13	http://www.huffingtonpost.com/2013/05/31/controversial-wikipedia-articles_n_3367573.html

22. Wikipedians most likely to war over 'Israel,' 'God'

The Times of Israel	3-Jun-13	http://www.timesofisrael.com/wikipedians-most-likely-to-war-over-israel-god/
23. Wikipedia 'Edit Wars': The most hotly contested topics		
NBC News online	3-Jun-13	http://www.nbcnews.com/technology/wikipedia-edit-wars-most-hotly-contested-topics-6C10167271
24. Chile, el tema más controvertido de Wikipedia en espaňol		
BBC Mundo	3-Jun-13	http://www.bbc.co.uk/mundo/noticias/2013/06/130603_tecnologia_conflictos_wikipedia_espanol_aa.shtml
25. Wikipedia's most controversial pages include Jesus and George W. Bush		
Toronto Star	5-Jun-13	http://www.thestar.com/news/world/2013/06/05/wikipedias_most_controversial_pages_include_jesus_and_george_w_bush.html
26. Blockbuster-Prognose mit Wikipedia		
Deutschlandfunk	13-Jun-13	http://www.dradio.de/dlf/sendungen/forschak/2142379/
27. Edit wars		
The Economist	5-Aug-13	http://www.economist.com/blogs/graphicdetail/2013/08/daily-chart-1

Appendix II – ISO 3166-1 Codes

This report often uses three letter codes to represent country names. These codes are the International Standards Organization codes for country naming, code 3166-1. They are presented below for the reader's convenience. Source: http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3

ABW	Aruba	GIB	Gibraltar	NLD	Netherlands
AFG	Afghanistan	GIN	Guinea	NOR	Norway
AGO	Angola	GLP	Guadeloupe	NPL	Nepal
AIA	Anguilla	GMB	Gambia	NRU	Nauru
ALA	Åland Islands	GNB	Guinea-Bissau	NZL	New Zealand
ALB	Albania	GNQ	Equatorial Guinea	OMN	Oman
AND	Andorra	GRC	Greece	PAK	Pakistan
ARE	United Arab Emirates	GRD	Grenada	PAN	Panama
ARG	Argentina	GRL	Greenland	PCN	Pitcairn
ARM	Armenia	GTM	Guatemala	PER	Peru
ASM	American Samoa	GUF	French Guiana	PHL	Philippines
ATA	Antarctica	GUM	Guam	PLW	Palau
ATF	French Southern Territories	GUY	Guyana	PNG	Papua New Guinea
ATG	Antigua and Barbuda	HKG	Hong Kong	POL	Poland
AUS	Australia	HMD	Heard Island and McDonald Islands	PRI	Puerto Rico
AUT	Austria	HND	Honduras	PRK	Korea, Democratic People's Republic of
AZE	Azerbaijan	HRV	Croatia	PRT	Portugal
BDI	Burundi	HTI	Haiti	PRY	Paraguay
BEL	Belgium	HUN	Hungary	PSE	Palestinian territories
BEN	Benin	IDN	Indonesia	PYF	French Polynesia
BES	Bonaire, Sint Eustatius and Saba	IMN	Isle of Man	QAT	Qatar
BFA	Burkina Faso	IND	India	REU	Réunion
BGD	Bangladesh	IOT	British Indian Ocean Territory	ROU	Romania
BGR	Bulgaria	IRL	Ireland	RUS	Russian Federation
BHR	Bahrain	IRN	Iran, Islamic Republic of	RWA	Rwanda
BHS	Bahamas	IRQ	Iraq	SAU	Saudi Arabia
BIH	Bosnia and	ISL	Iceland	SDN	Sudan

	Herzegovina
BLM	Saint Barthélemy
BLR	Belarus
BLZ	Belize
BMU	Bermuda
BOL	Bolivia, Plurinational State of
BRA	Brazil
BRB	Barbados
BRN	Brunei Darussalam
BTN	Bhutan
BVT	Bouvet Island
BWA	Botswana
CAF	Central African Republic
CAN	Canada
CCK	Cocos (Keeling) Islands
CHE	Switzerland
CHL	Chile
CHN	China
CIV	Côte d'Ivoire
CMR	Cameroon
COD	Congo, the Democratic Republic of the
COG	Congo
COK	Cook Islands
COL	Colombia
COM	Comoros
CPV	Cape Verde
CRI	Costa Rica
CUB	Cuba
CUW	Curaçao
CXR	Christmas Island
CYM	Cayman Islands
CYP	Cyprus
CZE	Czech Republic
ISR	Israel
ITA	Italy
JAM	Jamaica
JEY	Jersey
JOR	Jordan
JPN	Japan
KAZ	Kazakhstan
KEN	Kenya
KGZ	Kyrgyzstan
KHM	Cambodia
KIR	Kiribati
KNA	Saint Kitts and Nevis
KOR	Korea, Republic of
KWT	Kuwait
LAO	Lao People's Democratic Republic
LBN	Lebanon
LBR	Liberia
LBY	Libya
LCA	Saint Lucia
LIE	Liechtenstein
LKA	Sri Lanka
LSO	Lesotho
LTU	Lithuania
LUX	Luxembourg
LVA	Latvia
MAC	Macao
MAF	Saint Martin (French part)
MAR	Morocco
MCO	Monaco
MDA	Moldova, Republic of
MDG	Madagascar
MDV	Maldives
MEX	Mexico
MHL	Marshall Islands
MKD	Macedonia, the former
SEN	Senegal
SGP	Singapore
SGS	South Georgia and the South Sandwich Islands
SHN	Saint Helena, Ascension and Tristan da Cunha
SJM	Svalbard and Jan Mayen
SLB	Solomon Islands
SLE	Sierra Leone
SLV	El Salvador
SMR	San Marino
SOM	Somalia
SPM	Saint Pierre and Miquelon
SRB	Serbia
SSD	South Sudan
STP	Sao Tome and Principe
SUR	Suriname
SVK	Slovakia
SVN	Slovenia
SWE	Sweden
SWZ	Swaziland
SXM	Sint Maarten (Dutch part)
SYC	Seychelles
SYR	Syrian Arab Republic
TCA	Turks and Caicos Islands
TCD	Chad
TGO	Togo
THA	Thailand
TJK	Tajikistan
TKL	Tokelau
TKM	Turkmenistan
TLS	Timor-Leste
TON	Tonga
TTO	Trinidad and Tobago
TUN	Tunisia
TUR	Turkey

DEU	Germany
DJI	Djibouti
DMA	Dominica
DNK	Denmark
DOM	Dominican Republic
DZA	Algeria
ECU	Ecuador
EGY	Egypt
ERI	Eritrea
ESH	Western Sahara
ESP	Spain
EST	Estonia
ETH	Ethiopia
FIN	Finland
FJI	Fiji
FLK	Falkland Islands (Malvinas)
FRA	France
FRO	Faroe Islands
FSM	Micronesia, Federated States of
GAB	Gabon
GBR	United Kingdom
GEO	Georgia
GGY	Guernsey
GHA	Ghana
	Yugoslav Republic of
MLI	Mali
MLT	Malta
MMR	Myanmar
MNE	Montenegro
MNG	Mongolia
MNP	Northern Mariana Islands
MOZ	Mozambique
MRT	Mauritania
MSR	Montserrat
MTQ	Martinique
MUS	Mauritius
MWI	Malawi
MYS	Malaysia
MYT	Mayotte
NAM	Namibia
NCL	New Caledonia
NER	Niger
NFK	Norfolk Island
NGA	Nigeria
NIC	Nicaragua
NIU	Niue
TUV	Tuvalu
TWN	Taiwan, Province of China
TZA	Tanzania, United Republic of
UGA	Uganda
UKR	Ukraine
UMI	United States Minor Outlying Islands
URY	Uruguay
USA	United States
UZB	Uzbekistan
VAT	Holy See (Vatican City State)
VCT	Saint Vincent and the Grenadines
VEN	Venezuela, Bolivarian Republic of
VGB	Virgin Islands, British
VIR	Virgin Islands, U.S.
VNM	Viet Nam
VUT	Vanuatu
WLF	Wallis and Futuna
WSM	Samoa
YEM	Yemen
ZAF	South Africa
ZMB	Zambia
ZWE	Zimbabwe

Appendix III – Terms used for the location identification parser

To identify where people are from, we had to infer from self descriptions. The methods are discussed more fully in the main document. Here we present the terms used to identify an association.

	Preceding Statements	Succeeding Statements
Works in	working in worked in work in	
Lives in	wikipedians in wikipedians in the lived in live in living in lives in live at based in reside in resides in residing in resided in resident of returned to return to moving to move to moved to based in graduated from graduate of user in	
Born/Nationality	user proud {demonym} user {place/demonym} user from {place/demonym} user is {demonym} user templates{demonym} template user {demonym} user ancestry {demonym} i'm a {demonym} i am {demonym}	{demonym} wikipedians {demonym} wikipedian {demonym} user {demonym} citizen {demonym} citizens {demonym} born

i’m {demonym}
am an {demonym}
am a {demonym}
wikipedians of {demonym}
of native {demonym}
user citizen {place}
am from
i'm from
raised in
is from
come from
coming from
born in
wikipedians from
originally from
home in

Appendix IV – Blog Posts from this Project

Blog Post Title	URL
Mapping Wikipedia Article Quality in North America	http://www.floatingsheep.org/2011/12/mapping-wikipedia-article-quality-in.html
Open invitation to a workshop in Amman: Middle Eastern Participation and Presence in Wikipedia	http://www.floatingsheep.org/2012/02/open-invitation-to-workshop-in-amman.html
A new tool to explore the geography of Wikipedia	http://www.floatingsheep.org/2012/04/new-tool-to-explore-geography-of.html
O Mundo Pela Wikipédia	http://www.floatingsheep.org/2012/04/o-mundo-pela-wikipedia.html
What percentage of edits to English-language Wikipedia articles are from local people?	http://www.floatingsheep.org/2013/03/what-percentage-of-edits-to-english.html
Mapping Controversy in Wikipedia	http://www.floatingsheep.org/2013/06/mapping-controversy-in-wikipedia.html
Geographies of the World's Knowledge	http://www.zerogeography.net/2011/09/geographies-of-worlds-knowledge.html
Mapping Arabic Wikipedia	http://www.zerogeography.net/2011/09/mapping-arabic-wikipedia.html
Mapping Wikipedia at the global-scale in Arabic, English, French, Hebrew and Persian	http://www.zerogeography.net/2011/11/mapping-wikipedia-at-global-scale-in.html
Mapping Wikipedia's augmentations of our planet	http://www.zerogeography.net/2011/11/mapping-wikipedias-augmentations-of-our.html
Article Quality in English Wikipedia	http://www.zerogeography.net/2011/12/article-quality-in-english-wikipedia.html
Wikipedia Article Quality in East Asia	http://www.zerogeography.net/2011/12/wikipedia-

	article-quality-in-east-asia.html
Wikipedia Article Quality in Africa	http://www.zerogeography.net/2012/01/wikipedia-article-quality-in-africa.html
Where do Wikipedia edits come from?	http://www.zerogeography.net/2012/02/where-do-wikipedia-edits-come-from.html
Mapping Edits to Wikipedia from Africa	http://www.zerogeography.net/2012/03/few-days-ago-i-blogged-about-map-that-i.html
Mapping Edits to Wikipedia from the Middle East and North Africa	http://www.zerogeography.net/2012/03/mapping-edits-to-wikipedia-from-middle.html
Interactive Wikipedia mapping tool	http://www.zerogeography.net/2012/04/interactive-wikipedia-mapping-tool.html
Mapping Wikipedia edits from South America	http://www.zerogeography.net/2012/04/mapping-wikipedia-edits-from-south.html
Mapping Wikipedia edits from Europe	http://www.zerogeography.net/2012/05/mapping-wikipedia-edits-from-europe.html
Dominant Wikipedia language by country	http://www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html
The most visible country in Europe (on Wikipedia) is...	http://www.zerogeography.net/2012/11/the-most-visible-country-in-europe-on.html
Virtuous Visible Circles: mapping views to place-based Wikipedia articles	http://www.zerogeography.net/2012/11/virtuous-visible-circles-mapping-views.html
Mapping Wikipedia Views in the Middle East and North Africa	http://www.zerogeography.net/2012/12/mapping-wikipedia-views-in-middle-east.html
Short reflection on our Wikipedia workshop in Amman	http://www.zerogeography.net/2013/01/short-reflection-on-our-wikipedia.html
Wikipedia is where there is all of the information' - defining histories in Peru	http://www.zerogeography.net/2013/02/wikipedia-is-where-there-is-all-of.html
Die Welt in der Wikipedia als Politik der Exklusion	http://www.zerogeography.net/2013/03/die-welt-in-der-wikipedia-als-politik.html

What percentage of edits to English-language Wikipedia articles are from local people?	http://www.zerogeography.net/2013/03/what-percentage-of-edits-to-english.html
Who edits Wikipedia? A map of edits to articles about Egypt	http://www.zerogeography.net/2013/03/who-edits-wikipedia-map-of-edits-to.html
Mapping Controversy in Wikipedia	http://www.zerogeography.net/2013/05/mapping-controversy-in-wikipedia.html
Controversy in Wikipedia in Africa	http://www.zerogeography.net/2013/06/controversy-in-wikipedia-in-africa.html
Controversy in Wikipedia in Australia	http://www.zerogeography.net/2013/06/controversy-in-wikipedia-in-australia.html
Controversy in Wikipedia in the UK and Ireland	http://www.zerogeography.net/2013/06/controversy-in-wikipedia-in-uk-and.html

Appendix V: Principle Wikipedias included in this analysis

Wikipedia	Language	Wikipedia	Language
ar	Arabic	ja	Japanese
bg	Bulgarian	ko	Korean
ca	Catalan	lt	Lithuanian
cs	Czech	ms	Malay
da	Danish	nl	Dutch
de	German	nn	Norwegian (Nynorsk)
el	Greek	no	Norwegian (Bokmål)
en	English	pl	Polish
eo	Esperanto	pt	Portuguese
es	Spanish	ro	Romanian
et	Estonian	ru	Russian
eu	Basque	simple	Simple English
fa	Persian	sk	Slovak
fi	Finnish	sl	Slovene
fr	French	sr	Serbian
gl	Galician	sv	Swedish
he	Hebrew	te	Telugu
hi	Hindi	tr	Turkish
hr	Croatian	uk	Ukrainian
hu	Hungarian	vi	Vietnamese
id	Indonesian	vo	Volapük
it	Italian	zh	Chinese