Getting to the Source: Where does Wikipedia Get Its Information From?

Heather Ford
Oxford Internet Institute
Oxford, United Kingdom
heather.ford@oii.ox.ac.uk

Shilad Sen
Macalester College
St. Paul, Minnesota
ssen@macalester.edu

David R. Musicant
Carleton College
Northfield, Minnesota
dmusican@carleton.edu

Nathaniel Miller
Amazon, Inc.
Seattle, Washington
nmiller@alumni.macalester.edu

ABSTRACT

We ask what kinds of sources Wikipedians value most and compare Wikipedia's stated policy on sources to what we observe in practice. We find that primary data sources developed by alternative publishers are both popular and persistent, despite policies that present such sources as inferior to scholarly secondary sources. We also find that Wikipedians make almost equal use of information produced by associations such as nonprofits as from scholarly publishers, with a significant portion coming from government information sources. Our findings suggest the rise of new influential sources of information on the Web but also reinforce the traditional geographic patterns of scholarly publication. This has a significant effect on the goal of Wikipedians to represent "the sum of all human knowledge."

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and SoftwareInformation Networks; H.5.3 [**Information Systems**]: Group and Organization Interfaces—*computer-supported collaborative work*

General Terms

Human Factors, Measurement

Keywords

Wikipedia, citations, sources, policy

1. INTRODUCTION

Wikipedia is one of the most influential sources of information about the world. Ranked as the 6th most popular website by Alexa, Wikipedia now reaches 365 million readers worldwide every day [1] making it the largest and most popular general reference works on the Internet. Although Wikipedia is available in 286 languages, Wikipedia English dominates the userbase for Wikimedia projects at almost 11 million views per hour, with the next highest number of views belonging to Spanish Wikipedia at only about 2 million [22].

How do Wikipedians, collaborating often across countries, disciplines and areas of expertise, determine what should be included in this powerful information resource? One of Wikipedia's core content policies is that it works towards "verifiability, not truth" [21], claiming that what is reflected on Wikipedia is only what "reliable sources" believe to be true, rather than the beliefs of individual Wikipedians. As Nathaniel Tkatz [17] notes, however, Wikipedia's policies *do* effectively define truth. They are just one of many competing knowledge systems that could theoretically be used on the site. For example, the so-called "Neutral Point of View" defines a host of policies and practices that draw a line around what should exist on Wikipedia and what should not. By following procedures such as these, Wikipedians socially construct a very particular version of the truth using Wikipedia's specific rules of inclusion and exclusion.

A key element of the construction of a Wikipedia "truth" happens through the choice of sources. This process fulfills two key functions within Wikipedia work. First, the practice of citing sources enables Wikipedians to determine the notability of an article's subject. According to policy, "If no reliable third-party sources can be found on a topic, Wikipedia should not have an article on it." [20]. In other words, if a reliable source writes about a topic, Wikipedians are more likely to decide that the topic warrants a Wikipedia article. Second, citations allow readers to verify information contained within the article. This helps resolve content debates among Wikipedians and lends credibility to an article [8].

A significant proportion of Wikipedia's hundreds of policies, guidelines and essays [2] refer to sources and provide indications on how to choose them. Although Wikipedia policies state that context is important for deciding whether a source should be used or not, they also define a clear hierarchy of sources according to specific informational characteristics. Wikipedia's "Core Content Policy" on "Verifiability" [20], for example, notes that "where available, academic and peer-

WikiSym '13, August 5-7, 2013, Hong Kong, China.

reviewed publications are usually the most reliable sources, such as in history, medicine, and science". In addition, "self-published sources" are included under the headline: "Sources that are usually not reliable" [20]. This trend continues in the "Identifying Reliable Sources" [18] guideline that states: "Articles should rely on secondary sources whenever possible" and that "extreme caution is advised" when relying on primary sources. Thus, even though guidelines provide examples where primary sources can be used, this usage must be argued within a framework outlining a very clear hierarchy of sources.

Although Wikipedia policy states that secondary sources, particularly those from reputable scholarly and news organisations, should be primarily employed in Wikipedia articles, there are a vast number of topics about which scholarly articles do not exist. These constitute articles about entertainment and breaking news which make up the majority of articles on English Wikipedia [15]. These topics are generally out of the remit of scholarly and traditionally reliable news publications since they are either considered unworthy of scholarship or too current to have elicited interpretation by secondary source authors. Scholarly and reliable news sources are also largely unavailable for articles concerning the developing world where the relative proportion of scholarly and media publishers is low and digitisation lags behind developed countries [7].

The question then becomes: How do Wikipedians balance the need to provide up-to-date, comprehensive information about a particular topic within the framework of what is a clear bias towards scholarly journals? We explore this topic on English Wikipedia through the lens of two foundational research questions:

RQ1: What types of sources are most common?

RQ2: Which sources are most valued?

Given Wikipedia's policies on sources and citations, we would expect the majority of citations to come from traditional scholarship, from secondary sources rather than primary or tertiary sources and, since we're focusing in this first study on English Wikipedia, we expect that they should come from English sources since Wikipedia's policy on verifiability requires that readers are able to look up the sources of material in order to check its accuracy.

As sections 4 and 5 will show, however, the picture of citations on the ground is a lot more complex. By looking at both policy and practice on Wikipedia, we gain a more accurate perspective on where Wikipedia is, in fact, getting its information from. Since patterns in source use help reveal the lens through which Wikipedians view the world and since Wikipedia is the most widely read global reference, our analyses may help to explain which voices are represented in what has become a prominent source of information for millions of people around the world. Answering these questions, we believe, holds a key to understanding whether peer produced information environments like Wikipedia broaden our access to alternative points of view or whether they merely reinforce the perspectives of dominant global information regimes.

2. RELATED WORK

Although several studies examine Wikipedia sources and citations, they have generally focused on a small proportion of articles or analyze how sources reflect external definitions of quality.

Nielsen conducted research on both scientific journal and news citations on English Wikipedia and found that astrophysics and biology journals turn out to be the most frequently cited. Nielsen also found that Wikipedia citation patterns tend to reinforce patterns of source inequality in related fields. His research on usage of the "cite journal" template [13] found that the number of Wikipedia citations for a journal is highly correlated with the product of the Journal Citation Reports' impact factor and total citations with the addition of a few outliers. In another study analysing usage of the "cite news" template for a Wikipedia dump from 2008, Nielsen finds the BBC, New York Times, and Washington Post to be the most cited news organizations, with the BBC far ahead other organisations [12]. He also found that the majority of the 20 most-cited news sources are American, with an additional four each being Australian and British.

Other research has looked at the entire Wikipedia English corpus using visualization techniques. Summers visualized links contained in an External Links SQL Wikipedia dump from 2010 [16]. His graphs showed large numbers of academic journal archives (JSTOR, WorldCat), news organizations (BBC, NY Times), and government websites (ncbi.nlm.nih.gov, geonames.usgs.gov) among the top 100 most linked-to domains from articles. Also with strong presences were entertainment websites (imdb.com, allmusic.com, billboard.com) and sports databases (baseball-reference.com, sports-reference.com). The domain that was most linked-to was toolserver.org, a website hosting software tools for Wikipedia users.

More critical research analysing the sources of Wikipedia's articles relating to particular history articles has been conducted by Brendan Luyt and others. In a study of a sample of Wikipedia world history articles, Luyt and Tan [11] found that they "suffer" from the choice of "a few US government and online media news" sites rather than academic journals or other scholarly publications. They argued that this overreliance on foreign government sources means that "the nature of the institutions producing these documents makes it difficult for certain points of view to be included" and that the reader might conclude that this overreliance on such sources is "a warning that other short-cuts may have been taken in the course of the article's preparation" [11].

In a study comparing Wikipedia accounts of Singaporean and Philippine history, Luyt went on to show that, despite the potential of new media for making visible previously marginalized voices, a more likely outcome is a mapping of the status quo in historical representation onto new media [9]. Luyt found that the account of Singapore history follows the dominant historiographical tradition much more closely than the Philippines because of the greater visibility of alterative historical narratives in the Philippines.

More recently [10], Luyt investigated articles relating to Philippine history and found that the most common cita-

<ref>[http://www.wikisym.org/ws2008 2008 WikiSym website]</ref>
<ref>{{Citation | year=2007 | editor-last=Désilets | editor-first=Alain | editor2-last=Biddle | editor2-first=Robert | title=Proceedings of the 2007 International Symposium on Wikis | publisher=ACM Press | url=http://portal.acm.org/toc.cfm?id=1296951}}</ref>
<ref>{{cite news | title=NIH to Begin Enforcing Open-Access Policy on Research It Supports | first=Paul | last=Basken |url=http://chronicle.com/article/NIH-to-Begin-Enforcing/135852/ | newspaper=[[The Chronicle of Higher Education]] |date=19 November 2012 | accessdate=26 November 2012}}</ref>

Table 1: The three most common methods for adding citations to articles. Editors include <ref> tag citations alongside the sentences they support and place {{Reflist}} at the end of the article to automatically generate the references list.

tions are short texts that summarize historical events of periods that have the advantage of being both easily found and easily read, providing facts about Philippine history that can be easily "mined". Luyt argues that Wikipedia introduces an even "worse" situation than in the time of printed encyclopedias because these summaries that are relied on so heavily by Wikipedia editors, "already represent the congealed consensus of the institutions hosting them." [10]

This study attempts to build from these earlier projects in order to gain a baseline for citation popularity and persistence over time on the English Wikipedia.

3. METHODOLOGY

We chose English Wikipedia in this first study because of its influence and its global nature. Unlike many language Wikipedias, Wikipedians from all parts of the world, regardless of their home language, edit English Wikipedia [3]. English Wikipedia's global body of authors support our analysis of source diversity.

Mediawiki supports markup syntax for citations called "<ref> tags". However, since the software does not require editors to use this syntax, humans add references in many different ways. We catalogued the Wikipedia syntax used for citations across a random selection of 30 articles, and analysed how sources were cited. Most pages used <ref> tags (described below), but others used embedded URLs in the text of a page, or included a References section created by hand that did not use the formal citation syntax. Given the wide variety of formats we observed, it was impossible to exhaustively identify all references. We focused on <ref> tags for three reasons. First, they show clear citation intent (as opposed to a random URL). Second, they are displayed in traditional citation format (as numbered links inline with the text they are sourcing that refer to citations at the end of the article). Third, they are the most common citation format. Figure 1 shows three examples of <ref> tags. Users insert these tags alongside the information they want to source and add the "{{RefList}}" template to the end of the page. After these two components are added to a page, Wikipedia generates a references section.

We used the wikipedia-map-reduce Java software package[1] to extract citations from English Wikipedia[2]. To support the longitudinal analyses in section 5, we analyzed the full revision history for all main namespace articles until May 2nd, 2012. We used Apache's Map Reduce framework on Amazon's Elastic Map Reduce (EMR) cloud computing infrastructure[3] to efficiently extract the history of references to all articles. We ignored revisions corresponding to Preidhorsky et al's definition of reverts and vandalism [14]. The result of this process was a list of source postings. Each posting consisted of <article, citation, start, end>, where start and end refer to the revision number and timestamp where a citation first appeared and was ultimately removed. In total, the software extracted 67,026,537 source postings for 3,482,541 distinct articles.

4. RQ1: WHAT TYPES OF SOURCES ARE MOST COMMON?

We created a taxonomy of sources to investigate the distribution of source types in Wikipedia according to Wikipedia's "Reliable Sources" policy [18]:

> The word "source" in Wikipedia has three meanings: the **work itself** (a document, article, paper, or book), the **creator** of the work (for example, the writer), and the **publisher** of the work (for example, Oxford University Press). All three can affect reliability.

Our taxonomy captured these three dimensions described above, plus three other dimensions that represented the geographic and cultural diversity of sources: Language (English or non-English), Country of Publisher (based on top-level-domains such as ".au" or the whois administrative contact), and Media (print or web). Geographic diversity is an important feature to evaluate both because we can evaluate the extent to which a range of publishers from different regions of the world are being reflected in the encyclopedia that aims to reflect "the sum of all human knowledge" [4]. Different publishers reflect different local concerns; what is considered notable and thus becomes available on the Kenyan Herald Tribune may not be covered by the New York Times.

We extracted a random sample of 500 citations that appeared in articles on May 2nd 2012. Two of the authors began by coding two sets of 50 citations, checking the context of the article where it was unclear what function the particular information was playing. The first set was used to refine the taxonomy, and the second set was used to validate intercoder reliability on the finalized taxonomy. Kappa values on the second set indicated strong agreement (0.7 or higher across all dimensions). Each coder then individually completed an additional 200 citations, for a total of 500 codings (50 + 50 + 200 + 200).

27% of citations referenced URLs that no longer existed, and were removed from our analysis. Of the remaining citations, 80% referenced material intended for viewing in a web browser — almost exclusively web pages (78%), supplemented by a small number of videos (1%) and images (1%). The remainder (20%) referenced material intended for print — primarily PDFs but also digitized books.

[1]https://code.google.com/p/wikipedia-map-reduce/

[2]http://en.wikipedia.org/wiki/Wikipedia:Database_download

[3]http://aws.amazon.com/elasticmapreduce/

Type: We categorized the type of information characterized by citations according to Wikipedia's definitions of primary, secondary and tertiary sources in its "Identifying reliable sources" core content policy [18]. Wikipedia generally distinguishes between primary, secondary and tertiary sources as the distance the author is from the subject. Primary sources are defined as "original materials that are close to an event, and are often accounts written by people who are directly involved", secondary sources are at least one step removed from an event and contain "an author's interpretation, analysis, or evaluation of the facts, evidence, concepts, and ideas taken from primary sources", and tertiary sources are compendia, encyclopedias, textbooks, obituaries, and other publications that summarize primary and secondary sources [19].

Table 2 describes these categories in detail and provides examples of them.

Wikipedia prefers that the majority of articles should be based on "reliable, published secondary sources and, to a lesser extent, on tertiary sources", advising that "primary sources are permitted if used carefully" [19]. We found that the majority of citations are not, in fact, from secondary sources and that the "scholarship" category only constitutes 16% of citations, behind "opinion and analysis" (21%) and "data/statistics" (18%). We also noted that some primary source citations seem to coexist with Wikipedia policy. For example, citations to homepages and brochures (12%) often identified specific entities; for example the text "WikiSym" may cite the WikiSym homepage.

Citations to data (considered as a primary source according to Wikipedia's definition) formed the second largest source type (18%). Many of these web pages were published by enthusiasts who had collected and published tables and statistics about sports, music, or historical events that were collated from a variety of sources. Such sources included published texts collated by ad hoc groups and the published sources were generally not cited on the page itself but rather cited more generally on an 'about' page. Some of these sites published data from government or academic entities. Others collected user generated content.

Creator: The creator categorization focused on possible source authorship differences between Wikipedia and traditional media (Table 3).

Half of all citations were created by named individuals, most of whom authored the content as part of their jobs. 45% of citations were cited to an organization - usually the same organization hosting the content - without stating a specific author. It is interesting to note that organizations such as religious organizations, non-profits, or NGOs that have clear political agendas make up such a large proportion of sources. These may conflict with Wikipedia policy discouraging sources that advocate for "political, financial, religious, philosophical, or other beliefs" [18].

The remaining 5% of citations pointed to collaborative kinds of sources, primarily user generated content (UGC). Wikipedia explicitly discourages UGC stating that "self-published media... are largely not acceptable. This includes any website whose content is largely user-generated, including the Internet Movie Database (IMDB), CBDB.com, collaboratively created websites such as wikis, and so forth... [18]." The policy does, however, state that there are occasions where self-published sources can be used, such as when they are used as sources of information about the subject of the article (sources about themselves) — thus, the low proportion here.

Publisher: The publisher is the organisation or group that hosts, maintains, edits, or reviews the content pointed to by a citation. We considered seven different types of publishers, separating out traditional from nontraditional media and academia in order to understand how citation practices on the ground relate to Wikipedia's policy on what constitutes reliable publishers of information. According to Wikipedia's guidelines on "Identifying reliable sources", reliable publishers exhibit "editorial control and a reputation for fact-checking" whereas "anyone" could publish a webpage or a book and claim to be an expert in the field and therefore "self-published" sources are "largely unacceptable" [18]. Table 4 describes these categories in more detail.

Our findings suggest that Wikipedians cite many publishers who do not, in fact, fit with its own standards of identifying reliability. Governments and associations are not traditionally recognised as publishers of scholarly information, are not mentioned explicitly in sources policies on Wikipedia. Additionally, according to Wikipedia's own standards, sources from such organisations do not have a well-accepted method of fact-checking and could, in fact, possess motives that work in opposition with the provision of neutral information. Governments and associations make up a surprisingly large percentage of the publishers of the citation sample at 9 and 14% respectively. Although Wikipedia prefers scholarly content, such material makes up a relatively small portion of overall citations (16%). Self-published sources (from individuals) make up a surprisingly large percentage of citations (6%), given Wikipedia's discouragement of them. Many of these self-published sources appeared to be reference sites created by fans or enthusiasts of some topic. Though traditional media (27%) and non-traditional media (26%) appear in roughly equal numbers, traditional academic sources (16%) vastly outnumber non-traditional academic sources (2%), indicating either a low number of academic blogs and other alternative forms of publication or the possibility that Wikipedians do not count such sources as reliable.

Geography and Language: Figure 1 shows the distribution of publisher by country and continent. Most publishers are located in countries whose primary language is English (80%) but a number of countries where English is the official language are not included in this list. India (2%) is the only developing country that accounts for a significant number of sources.

5. RQ2: WHICH SOURCES ARE MOST VALUED?

The previous section considers how sources in Wikipedia classify into different categories. In this section, we will consider which sources are most valued by Wikipedia editors. "Most valued" can mean any number of things. For purposes of this paper, we consider two definitions of this

	type	%	description	examples
primary	**data / statistics**	18%	Information such as numbers and short phrases that would suit storage in a database or table	Statistics of baseball players presented in tabular format http://www.baseball-reference.com/players/n/nashji01.shtml
	art / pop culture	1%	Artefact that is being referred to in the article	Music video directed by the person who is the subject of the Wikipedia article where the citation occurs http://www.cmt.com/videos/ty-herndon/71920/i-have-to-surrender.jhtml
	homepage / brochure	12%	Artefact that is promoting a particular product/service/person (all URLs that point to the whole site rather than a particular page)	The homepage of the St. Mary Catholic Church in Wilmington, North Carolina is cited in an article with the same title (http://www.thestmaryparish.org)
	conversation / announcements	3%	Information that is aimed at generating conversation or at alerting others to a particular topic	A US Federal Register PDF announcing the completion of an inventory of human remains cited in an article about the geographical site http://www.gpo.gov/fdsys/pkg/FR-2005-02-22/pdf/05-3322.pdf
secondary	**scholarship**	16%	Books, journal articles, papers written by academics and/or researchers	A journal about 'Attachment and Human Development' cited in an article about 'Attachment therapy' http://www.tandfonline.com/toc/rahd20/5/3
	news	16%	Information describing recent events or happenings	A news story about the recent setup of a Turkish-Maltese Business Council cited in an article about Malta-Turkey relations http://www.maltamedia.com/artman2/publish/financial/article_3364.shtml
	opinion / analysis	21%	Information that analyzes (not just describes) or provides opinion about a particular topic	A blog post by the European Film Academy (EFA) protesting the arrest of Iranian filmmaker Jafar Panahi in an article abou Panahi http://efareviews.cineuropa.org/2010/03/jafar-panahi-arrest-european-film.html
tertiary	**reference**	13%	Encyclopedic or encyclopedic-like entries used to describe particular topics in the authoritative voice	This page on Eldena, an American Steam Merchant that was sunk in WWII contains numerative information about ship as well as a description of the sinking http://uboat.net/allies/merchants/ships/2994.html
	directories / archives	0%	Link directories or archives of information published elsewhere	None found.

Table 2: Details of the taxonomy for the "type" of an article. The percentages indicate the percent a type accounted for in our 500 citation sample.

type	%	description	examples
individuals	50%	Attribution of authorship to named individuals.	Article by Sally Hofmeister http://articles.latimes.com/1996-08-23/business/fi-36983_1_set-top-boxes
organization	45%	Information attributed to the organization/company/government dept rather than to specific individuals.	Public records statement of the County of Los Angeles Public Library http://www.colapublib.org/aboutus/publicinfo.html
collaborative	5%	Authors working together outside the structure of a company or organization.	List of matches played by cricket player on a site that asks for help in transcribing http://cricketarchive.com/howtohelp.html

Table 3: Details of the taxonomy for the "creator" of an article.

type	%	description	examples
government	9%	Any .gov	County of Los Angeles Public Library site http://www.colapublib.org/aboutus/publicinfo.html
association	14%	Any .org and/or any NGO/ association/religious/political organisation	Press release from 'Horn Relief' NGO http://www.hornrelief.org/goldman-prize-2002.htm
trad media	27%	Mass media company (includes publishers of newspapers, magazines, television, music„ popular books) or publisher of traditional formats of media including news, opinion and analysis.	NYTimes article http://www.nytimes.com/1989/06/25/world/a-us-soldier-who-defected-is-given-30-years-for-spying.html
trad academic	16%	Any .edu and/or publisher of research by academics/professional researchers and/or publisher of reference works	Journal article published by Bentham Science publishers http://www.benthamdirect.org/pages/content.php?cpd/2007/00000013/00000018/0006b.sgm
non-trad media	26%	Organisation/company/group where the editing/review process is unclear and/or 'new media' enterprises including platforms like YouTube, WordPress, Flickr and other companies that are considered new entrants into the media business and/or where the primary business of the group is not in the provision of media but that this is a secondary supporting function	Article on band's fan site http://onesecondbush.com/bush/tour/_2000/
non-trad academic	2%	Publishers hosting content outside of traditional formats (e.g. academic blogs rather than academic journals) or outside of traditional peer review (e.g. academic online archiving services like arxiv.org)	Blog post in academics' group blog http://www.e-ir.info/2009/09/09/international-law-and-the-bush-doctrine/
individuals	6%	A single individual acts as the entity hosting, maintaining, editing and reviewing content	Information about political constituencies maintained by an individual http://www.leighrayment.com/commons/Ncommons1.htm

Table 4: Details of the taxonomy for the "publisher" of an article.

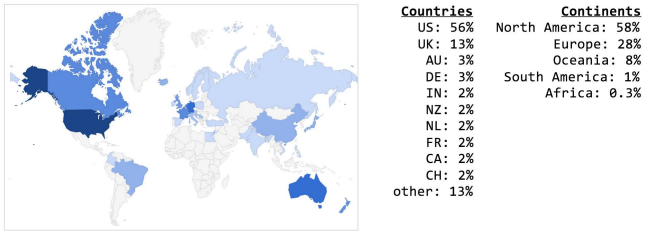

Figure 1: Geographic distribution of English Wikipedia's sources, grouped by country and continent.

phrase: those sources that are cited most often, and those sources that persist over time without being removed.

Because we wanted citations with a variety of different formats, we analyzed URLs — a common element in many different citation formats. Specifically, we extracted a URL's domain name, the logical equivalent of a publisher. Our analysis below thus focuses on citations to a particular domain name.

Popularity: We first analyze an obvious measure of value: the number of citations to each domain:

RQ2a: Which sources are cited most often?

In order to measure the frequency of cited sources, we extracted all 11M citations that appeared in articles on May 2nd 2012. This analysis is then limited to the 77% of those citations that referenced a URL. Table 5 lists the 20 most cited domain names. Google.com tops the list with 1.97% of citations, mostly references to books.google.com (1.37%).

Several patterns emerge in these top results. US sources dominate the list. The domain names correspond to widely-recognized organizations who provide data, archives, and news content. 10 of the top 20 publishers are traditional media companies with origins in print or television. Four of the remaining domains primarily publish data: stat.gov.pl (census and location data about Poland), imdb.com (movie data), allmusic.com (music data), and census.gov. Youtube (#8 with 0.44% of all citations) stands out as a unique content type (video) among the top domains.

Persistence: The previous results showed that the New York Times is the second most cited source in Wikipedia. Does it appear as #2 because it publishes a massive amount of information, or because the Wikipedia community views it as reliable? This section considers an alternative measure of value that sheds light on this distinction:

RQ2b: Which sources are most persistent?

The answer to this question should indicate what sorts of citations the Wikipedia community deems most worthy of keeping, not merely adding. As with the previous results, we grouped our persistence findings by the web domain referenced in a citation. To measure persistence, we used two different approaches, described below.

The first approach, which we call ***deletion-persistence***, calculates the percent of citations for a domain that had been removed. This was done over the entire history of the Wikipedia until May 2nd, 2012. Values close to 0 indicate that few citations are removed. One challenge in measuring a fraction such as this one is in handling domains with a very few citations. For instance, domain a, where 2 of 1000 had been removed, and domain b, where 0 of 2 citations were removed, would have "naive" deletion-persistence of 0.2% and 0.0% respectively. However, intuition suggests that domain a is more persistent. We used beta-binomial bayesian smoothing to control for these small sample sizes, where the alpha and beta parameters were estimated from the entire population of citations [5]. The results are shown in Table 6, which presents the top 20 most deletion-persistent domains in English Wikipedia. Note that many of these "domains" are actually Wikipedia templates, which in turn contain references to external citations.

The most deletion-persistent domains, shown in Table 6, are almost all associated with government or academic sources who publish data such as census information, geographic information, government structure, economic data, and so on. This suggests that the Wikipedia editing process evaluates sources of data as authoritative, credible, and relatively non-controversial. Some of these citations are templates or semi-automated sorts of entries that turn out to be deletion-persistent *despite* the fact that humans have not spent considerable time crafting or defending them. This is particularly interesting when contrasted with the Wikipedia policy "No original research" (NOR) [19]. The NOR policy generally prefers the use of secondary sources over primary ones. The opening sentence of the section on sourcing states "Wikipedia articles should be based on reliable, published secondary sources and, to a lesser extent, on tertiary sources." [19]. Later on, it does indicate with a caveat that "...primary sources that have been reliably published may be used in Wikipedia; but only with care, because it is easy to misuse them. Any interpretation of primary source material requires a reliable secondary source for that interpretation." A list of admonitions then follows, describing a variety of ways that primary sources should be not used. Though the caveats exist, the main language of the policy leads contributors away from primary sources. The fact that data-based primary sources avoid deletion more than any other kind of source provides insight into how Wikipedians value sources in practice.

The second measure of persistence, which we refer to as ***revision-persistence***, calculates the number of revisions that each citation survived, and averaged the count for each domain. This was done over for the same time period as described above for deletion-persistence. This measure is motivated by the idea that a citation peristing through many edits for a period of time might be considered more persistent than a citation persisting through a small number of edits for the same period. In order to deal with domains with low numbers of actual citations, we performed a Gaussian Bayesian smoothing process [5].

Table 7 shows the 20 most revision-persistent domains in English Wikipedia. There are clear similarities to the the most deletion-persistent revisions (Table 7). The domains

Row	Domain	Information Type	Country	% Citations
1	google.com	search engine and archive	US	1.97%
2	nytimes.com	news and analysis/opinion	US	1.18%
3	bbc.co.uk	news and analysis/opinion	UK	1.04%
4	stat.gov.pl	data/statistics	Poland	0.49%
5	guardian.co.uk	news and analysis/opinion	UK	0.49%
6	imdb.com	directory/archive	US	0.47%
7	archive.org	directory/archive	US	0.45%
8	youtube.com	video	US	0.44%
9	allmusic.com	directory/archive	US	0.41%
10	cnn.com	news and analysis/opinion	US	0.36%
11	yahoo.com	search engine	US	0.36%
12	nih.gov	analysis/opinion	US	0.26%
13	latimes.com	news and analysis/opinion	US	0.26%
14	telegraph.co.uk	news and analysis/opinion	UK	0.26%
15	census.gov	data/statistics	US	0.25%
16	washingtonpost.com	news and analysis/opinion	US	0.24%
17	espn.go.com	news and analysis/opinion	US	0.23%
18	independent.co.uk	news and analysis/opinion	UK	0.22%
19	amazon.com	directory/archive	US	0.21%
20	time.com	news and analysis/opinion	US	0.21%

Table 5: Most cited domains on May 2, 2012.

Row	Domain (+template, if applicable)	% Removed	Description
1	amar.org.ir (irancensus2006)	0.24%	census data
2	geonames.nga.mil (geonet3)	0.29%	geographic name data
3	basketball-reference.com (cite basketball-reference)	0.37%	sports data
4	citation needed cheap	0.41%	Wikipedia template for citation needed
5	statistik.tg.ch	0.42%	government statistics
6	wiki:digital elevation model	0.48%	Wikipedia article reference (geography related)
7	pxweb.bfs.admin.ch	0.48%	government statistics
8	so.ch	0.51%	government information
9	wbprd.gov.in	0.53%	government information
10	biographi.ca (cite dcb)	0.58%	biographical data
11	insae-bj.org	0.58%	economic data
12	wiki:Instituto Nacional de Estadística e Informática	0.60%	Wikipedia article for government statistics site
13	stat.gov.pl	0.60%	government statistics
14	fr.ch	0.60%	government information
15	media-stat.admin.ch	0.61%	government information
16	wahlen.rlp.de	0.67%	government information
17	deutschebahn.com (dbcatsurl)	0.67%	rail prices
18	deldot.gov (delaware road map)	0.67%	maps
19	rfspro.ru	0.70%	soccer tournament data
20	ilo.cornell.edu	0.72%	economic data

Table 6: Most deletion-persistent domains over entire history until May 2, 2012.

Row	Domain	Avg # revisions	Description
1	vfxworld.com	176.7	video news and information
2	aci.aero	163.3	airport data
3	filmforce.ign.com	155.3	film news and information
4	scoringsessions.com	154.9	music news and information
5	inogolo.com	149.1	pronunciation guide
6	nielsenmedia.com	143.0	entertainment news and commercial service
7	cre.gov.uk	143.0	policy news and advocacy
8	wiki:Larco Museum	141.1	Wikipedia article for museum
9	host17.hrwebservices.net	135.8	music data
10	metoffice.com	135.5	climate information and data
11	gamecriticsawards.com	135.2	game data
12	premiere.com	134.9	entertainment news
13	eng.gov.spb.ru	134.7	government information
14	devdata.worldbank.org	132.7	government information
15	airports.org	131.8	airport data
16	footballfanscensus.com	131.0	sports data
17	worldairlineawards.com	129.6	airline data
18	morganquitno.com	125.7	government data
19	grb.uk.com	124.5	university data
20	weather.yahoo.com	121.8	weather data

Table 7: Most revision-persistent domains over entire history until May 2, 2012.

that are most revision-persistent again include data intensive sources. Interestingly, some top revision-persistent domains are entertainment-related sources of information. This observation is consistent with entertainment-related articles receiving more contributions than any other category [15]. Citations to reliable sources in these revision-heavy entertainment articles have the opportunity to persist through more revisions than citations in rarely-edited articles.

TLD Diversity: Tables 5, 6, and 7 appear to have differences in geographic diversity. We studied this more closely by looking at the diversity of top level domains (TLDs) among citations. TLDs constitute the last portion of a domain name (e.g. ".org" or ".uk"). Although some TLDs such as ".com" do not conventionally represent a specific geography, many do. Empirically analyzing the distribution of all TLDs provides a glimpse into geographic diversity.

Table 8 shows the distribution of TLDs for four groups of citations. The first pair of columns shows the TLD distribution for all domains (a baseline). The second, third, and fourth pair of columns shows the TLD distribution for citations to the 1000 most cited, most deletion-persistent, and most revision-persistent domains respecitively. The table contains the top 20 TLDs from each list, with the remainder grouped into a row called "other."

All four lists show relatively little geographic diversity, with the first obviously non-English speaking TLD garnishing less than 1% of citations for each group. The distribution of TLDs for the top-1000 most cited domains (the second pair of columns) is similar to the baseline, though the non-English TLD tail is less prominent. The columns associated with the most persistent citations (column pairs three and four) show much stronger representation from the ".gov" domain, reflecting the correlation we observed in earlier sections between primary data sources and persistence. Interestingly, the list of 1000 most cited and revision-persistent shows substantially *less* TLD diversity than Wikipedia overall, with the ".com" TLD accounting for 72.4% of citations. The revision-persistence metric favors lasting citations in articles with higher levels of contribution. These articles may express significant preference for certain TLDs.

6. DISCUSSION

The popularity and persistence of Wikipedia citations help us understand the types of sources that are seen as "reliable" by Wikipedians which, in turn, provides a window into the sources that shape the world's most popular reference work. We explored what types of sources are the most common on Wikipedia and categorized them according to the key method for defining reliability of sources on Wikipedia, namely, whether the source is primary, secondary or tertiary. We also analysed sources according to the institutional and geographic characteristics and according to the publisher and the creator. The second part of our study looked at which sources persist over time. We identify three key themes that we discuss below.

Firstly, our data shows the emergence of new sources that Wikipedians consider reliable beyond traditional scholarship. We see the emergence of "governments," "associations," "collaboratives", and other "non-traditional media sources." Gaming companies provide news on new games and gaming hardware releases, non-profit human rights groups reporting on political activity locally and abroad, and ad hoc volunteer communities provide lists and reference data sourced from a variety of other databases. Organisations such as these whose primary business focus has not been traditionally information-focused and where editing and reviewing processes are opaque are increasingly providing news and information important to Wikipedia.

Given that organisations and groups author similar numbers of sources to named individuals in our sample, we see that groups are acting to shape discourse on Wikipedia. This may signal a significant shift away from the transparency accorded with sourcing information from particular individuals. While scholarship is traditionally seen as a lone enterprise by specific individuals, new media forms are growing that are governed by alternative peer review mechanisms (providing the ability of users to edit, point out flaws or contribute to information). Through Wikipedia's sourcing patterns we may be witnessing a movement away from modernist conceptions of authorship or the growing institutionalisation of what is considered "reliable" content on the Web.

Secondly, we find a significant portion of the citations come from sources that can be considered primary and that many of these sources are also the most persistent, outliving other sources that may be more readily removed by editors. Citations that occur most frequently at any given time in Wikipedia are not necessarily the ones that persist over time. The most frequent citations are often to large and well-known media organizations such as the New York Times or the BBC. Of the most frequent citations, a relatively small subset point to large government-run sources of data, such as stat.gov.pl or census.gov. When we examine the most deletion-persistent sources, we see instead that they nearly all refer to sources of data, many of whom are not nearly as large or well-known as those that are present in the most frequent citations. This could be because of the growing authority of data sources (Big Data), or the important role data plays in supporting facts such as those found in infoboxes. The widespread reliance on data may also be a result of greater availability thanks to individual online publishers such as the enthusiast who collated baseball statistics from a variety of sources. When we examine the most revision-persistent sources, we again see that many of them are data sources; we also see that a number of them relate to entertainment news. This demonstrates that amongst some of Wikipedia's most high-traffic pages, the conception of what is considered "reliable" does not necessarily refer to traditional academic publications.

Finally, we find that information from US sources dominates the spread of Wikipedia citations and that there is a long tail of sources from other countries. The over-reliance on sources from North America (56%) versus Africa (0.3%), for example, where at least 23 countries have English as an official language demonstrates a significant inequality in terms of voices represented on Wikipedia. India has been able to emerge as a source with impact relative to developed countries like Australia, but source use on Wikipedia generally reflects the patterns of media and scholarly publishing around the world [7] This reinforces perspectives [6] that highlight

TLD	% of all (baseline)	TLD	% of top 1000 most-cited	TLD	% of top 1000 deletion-persistent	TLD	% of top 1000 revision-persistent
com	54.2%	com	60.6%	com	30.0%	com	72.4%
org	12.1%	uk	10.2%	gov	27.0%	gov	9.1%
uk	8.4%	org	8.3%	org	14.6%	uk	6.5%
gov	3.2%	gov	5.4%	edu	11.0%	edu	3.2%
net	3.0%	edu	3.3%	ca	4.0%	org	1.9%
edu	2.9%	au	2.3%	us	3.2%	ca	1.5%
au	2.3%	ca	1.6%	au	1.5%	mil	1.2%
ca	1.7%	net	1.4%	mil	1.3%	net	1.2%
de	0.9%	mil	0.9%	ch	1.0%	au	0.9%
us	0.8%	us	0.8%	uk	0.9%	us	0.4%
jp	0.6%	de	0.5%	fr	0.8%	de	0.3%
nz	0.5%	nz	0.5%	nz	0.8%	fr	0.2%
in	0.5%	jp	0.4%	br	0.6%	jp	0.1%
mil	0.5%	in	0.4%	jp	0.3%	fi	0.1%
ru	0.4%	eu	0.3%	pl	0.3%	tv	0.1%
info	0.4%	ie	0.3%	ru	0.3%	nl	0.1%
fr	0.4%	int	0.2%	fi	0.2%	biz	0.1%
ie	0.4%	fr	0.2%	no	0.2%	il	0.1%
nl	0.3%	ch	0.2%	se	0.1%	ie	0.1%
it	0.3%	ph	0.2%	in	0.1%	no	0.0%
other	6.2%	other	2.1%	other	1.5%	other	0.5%

Table 8: Percent of citations associated with each top-level-domain for four groups of citations on May 2, 2012: 1) all citations, 2) citations to the 1000 most popular domains, 3) citations the 1000 most deletion-persistent domains, 4) citations to the 1000 most revision-persistent domains.

the vast territories in the developing world that remain terra igcognita on Wikipedia because they have yet to be covered. This will be one of the most significant challenges for Wikipedians whose ultimate goal is to represent "the sum of all human knowledge" [4].

7. CONCLUSIONS AND FUTURE WORK

Wikipedia's view of the world is ultimately driven by the sources from which its perspectives are derived. In order to better understand the shape and location of the voices behind these sources, we presented a taxonomy for Wikipedia sources, categorized according to both Wikipedia's most prevalent policies indicating which sources are most reliable, as well as chategories relating to geographic voice and representation. We show an emergence of governments and associations as significant producers of sources reflected on Wikipedia, as well as "non-traditional" media publishers and a general reinforcement of traditional patterns of inequitable scholarly publishing around the world. We also show that datasets and statistics are one of the most popular and persistent sources of information on Wikipedia, even though these sources are considered "primary" and thus not as reliable according to Wikipedia policy.

This is a mixed picture of whether Wikipedia provides access to alternative sources of information because, although there is an emergence of non-traditional media publishers, we can also see a rise in institutional sources and a heavy dominance of sources from the United States. More research needs to be done to further investigate this question. We see this study as a baseline for future studies across specific categories of content, for example, news, across different language versions, and longitudinally to understand whether source patterns have changed over time and whether source patterns change according to an article's stage of development. We also note that our definition of "most-valued" sources in this paper is intended as a starting point, and is somewhat narrow. We intend to deepen our understanding of "most-valued" by studying Wikipedian's individual interactions with sources. For example we will study how individual editors draw upon sources in editing disputes, and how editors resolve different perspectives represented by sources, especially when competing information is often rife (e.g. at the beginning of civil uprisings). Finally, this work considers only how citations themselves are characterized in Wikipedia; further work might examine how the *content* of Wikipedia itself is, in turn, influenced by the distribution of citations.

8. ACKNOWLEDGMENTS

We thank Laurel Orr at Carleton College for her considerable work in looking at other measures of citation value. We gratefully acknowledge Amazon for supporting our work through an Amazon Web Services Research Grant for computational time and to Ushahidi, Hivos and the Open Society Institute for funding some of the original thinking around sources on Wikipedia. We also thank John Riedl at the University of Minnesota for his creative input; in particular, the concept of examining persistent sources was suggested by him.

9. REFERENCES

[1] Wikipedia.org site info. http://www.alexa.com/siteinfo/wikipedia.org.

[2] J. E.-. P. J. Butler, B. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1101–1110, 2008.

[3] W. Foundation. Wikipedia editors study: Results from the editor survey, april 2011. `https://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf`, 2011. [Online; accessed 18-March-2013].

[4] W. Foundation. The Wikipedia Foundation Vision.

http://wikimediafoundation.org/wiki/Vision, 2013. [Online; accessed 18-March-2013].

[5] A. Gelmen, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition*.

[6] M. Graham. Wiki space: Palimpsests and the politics of exclusion, 2011.

[7] M. Graham, S. A. Hale, and M. Stephens. The location of academic knowledge, June 2011.

[8] T. Lucassen and J. M. Schraagen. Trust in Wikipedia : How Users Trust Information from an Unknown Source. *Security*, pages 19–26, 2010.

[9] B. Luyt. 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, 2011.

[10] B. Luyt. The inclusivity of wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists. *Journal of the American Society for Information Science and Technology*, 63(9):1868–1878, 2012.

[11] B. Luyt and D. Tan. Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology*, 61(4):715–722, 2010.

[12] F. Å. Nielsen. Top news cites referenced from wikipedia. http://fnielsen.posterous.com/top-news-cites-referenced-from-wikipedia.

[13] F. Å. Nielsen. Scientific citations in wikipedia. *arXiv preprint arXiv:0705.2106*, 2007.

[14] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 259–268, New York, NY, USA, 2007. ACM.

[15] O.-V. J. M.-M. R. . L. C. Reinoso, A. J. Most popular contents requested by users in different Wikipedia editions. 2013.

[16] E. Summers. top hosts referenced in wikipedia (part 2) | inkdroid. http://inkdroid.org/journal/2010/08/25/top-hosts-referenced-in-wikipedia-part-2/.

[17] N. Tkacz. Power, visibility, wikipedia. *Southern Review*, 40(2):5–19, 2007.

[18] Wikipedia. Wikipedia:Identifying reliable sources. `http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources`, 2013. [Online; accessed 18-March-2013].

[19] Wikipedia. Wikipedia:No original research. `http://en.wikipedia.org/wiki/Wikipedia:No_original_research`, 2013. [Online; accessed 18-March-2013].

[20] Wikipedia. Wikipedia:Verifiability. `http://en.wikipedia.org/wiki/Wikipedia:Verifiability`, 2013. [Online; accessed 18-March-2013].

[21] Wikipedia. Wikipedia:Verifiability, not truth. `http://http://en.wikipedia.org/wiki/Wikipedia:Verifiability,_not_truth`, 2013. [Online; accessed 18-March-2013].

[22] E. Zachte. Wikipedia statistics as at April 2013. `http://stats.wikimedia.org/EN/Sitemap.htm`, 2013. [Online; accessed 12-June-2013].