Discovering behavioral patterns in collective authorship of place-based information

Darren Hardy*
Bren School of Environmental Science and Management
University of California, Santa Barbara, USA

Abstract

While current GIS research has focused on technological issues of visualization and data organization, the emergence of new forms of collective authorship suggest we need new information frameworks and behaviors. How do individuals contribute place-based information to a digital commons? What are the authorship dynamics of such collective effort? For my research, I will use spatial data mining methods to characterize authorship behavior on a corpus of 1 million geotagged articles across 20 languages from Wikipedia.

Keywords: geotagging, peer production, Wikipedia, bots.

1 Introduction

The so-called “GeoWeb” – exemplified by the meteoric rise in use of geobrowsers like Google Earth™ – extends the functionality of the Web by providing a geographic worldview and lightweight geospatial services to publicly shared data. Concurrently, Web 2.0 is increasingly emphasizing distributed collaboration and highly interactive media while the rapid adoption and diverse use of social computing technologies is enabling new forms of collective authorship. Does this mixture of collective effort, high interactivity, and location-awareness induce new information behaviors? How do individuals contribute place-based information to a digital commons? What are the authorship dynamics of such collective effort? To address these questions, I search for common patterns of authorship behavior using data mining methods on a corpus of 1 million geotagged articles across 20 languages from Wikipedia.

Collective authorship is generally the process of information or knowledge production by multiple individuals, but here I refer to it as a mass collective effort by individuals to produce information artifacts within a digital commons. In Web 2.0 research, for example, terms like wnikinomics (Tapscott and Williams, 2006), collective intelligence (O’Reilly, 2005), crowdsourcing (Brabham, 2008), online collectivism (Lanier, 2006), and peer production (Benkler, 2005) also describe various aspects of this large-scale collective action phenomenon. In fact, Web 2.0 systems fundamentally recognize the utility of user-contributed content, and their common thread is the basic approach to harness social behaviors of large online user communities.

*Email: dhardy@bren.ucsb.edu. Submitted to 9th International Conference of the Association of Internet Researchers (Internet Research 9.0: Rethinking Community, Rethinking Place. Copenhagen, Denmark, 15–18 October 2008). This draft includes minor corrections to the original.
Today, users can share and pool geographic information via the Internet at lower costs, in effect creating a global digital commons of geographic knowledge (Goodchild, 2007). Geobrowser is a recent term – driven by the immense popularity of Google Earth™ – that highlights how the geobrowser transforms traditionally heavyweight geographic information systems (GIS) into a simple yet compelling Web browser-like interface. While current GIS research has focused on technological issues of visualization and data organization, new forms of collective authorship have emerged during the Web 2.0 era. For example, recent research on collaborative mapping, or so-called neogeography – “geography without geographers” (Sui, 2008, p. 4) – suggests needs for new information frameworks and behaviors. Neogeography expands the notion of the “public” from prior work in public participation GIS (Craig et al., 2002) to include much larger, distributed participation, but at its core is the notion of collective authorship.

In my research, I examine the subset of Wikipedia articles and contributions that are about a particular place – that is, geotagged Wikipedia articles that have a specific geographic location. My methodology uses data extraction and web mining methods for collecting corpus data, and spatial clustering methods for analyzing authorship behavior. Due to its scale and complexity, the Web follows power laws in its link structure (Barabási and Albert, 1999) and surfing behavior (Huberman et al., 1998). So does Wikipedia for both readership (Priedhorsky et al., 2007) and editing (Almeida et al., 2007). That is, a small number of articles receive the majority of contributions, and the vast majority of articles receive a small number of contributions (i.e., the “long tail”). But when Wikipedia authors write about place, do the same authorship dynamics apply?

2 Background

In this section, I review disciplinary literature, and discuss social computing, new forms of collective authorship, and geographic information in Web 2.0.

2.1 Social computing

Social computing is the interdisciplinary study of socially-aware information systems. Computer-mediated communication research studies how interpersonal and group dynamics change when communication is through computerized media like email or instant messaging. This research (e.g., Ackerman, 1997; Hancock et al., 2004) primarily focuses on technology (computer-supported cooperative work) or human factors (human-computer interaction), and on sociological aspects (e.g., Katz and Rice, 2002).

In the 1990s, the emergence of the Web drove this research to further examine sociological aspects, such as the dynamics of online communities (Jones, 1998; Kolbitsch and Maurer, 2006). Today, the so-called “Web 2.0” technologies provide further Web functionality through its emphasis on distributed collaboration, highly interactive media, and user-contributed content (O’Reilly, 2005). These technologies have roots in prior research on computer supported cooperative work (e.g., Grudin, 1994). In fact, Web 2.0 systems fundamentally recognize the utility of user-contributed content. Examples include (a) distributed social bookmarking sites like del.icio.us™; (b)

---

1The term social computing is also used in an unrelated context for computational methods for social network analysis in sociology. Wang et al. (2007) propose the following integrative definition for social computing (emphasis added): “Computational facilitation of social studies and human social dynamics as well as the design and use of information and communication technologies that consider social context.”
image file sharing sites like Flickr™; (c) knowledge production like Wikipedia®; (d) video sharing sites like YouTube™; and, (e) social networking sites like Facebook®, which tracks relationships between individuals and provides communication and sharing services. The common thread in these distributed, socially-aware Web information systems is the basic approach of designing systems to harness social behaviors of large online user communities. In particular, these systems sustain their viability only when they enjoy massive participation.

2.2 Collective authorship

Theories of information or knowledge production behavior are typically researched distinct from information seeking behavior. Knowledge production research focuses on epistemological issues as well as communication and psychological issues. Schema theory from cognitive psychology, for example, suggests mental processes by which individuals integrate and understand knowledge. Collective authorship is generally the process of information or knowledge production by multiple individuals. The term “information production” itself has different semantics across disciplines. In humanities, the term may represent the authoring of a written work or book. In economics, the term may represent market resources, or commodities, or perception, or even a constitutive force in society (Browne, 1997, p. 266). In library science, the term may represent how we communicate collaborative work to public scientific knowledge (Cronin, 2001; Galison, 2003; Rowlands and Bawden, 1999). I refer to collective authorship, however, as a mass collective effort by individuals to produce information artifacts within a digital commons. In Web research, terms like wikinomics (Tapscott and Williams, 2006), collective intelligence (O’Reilly, 2005), crowdsourcing (Brabham, 2008), and online collectivism (Lanier, 2006) also describe various aspects of this large-scale collective action phenomenon.

Thus far, many new forms of collective authorship that have arisen from social computing technologies, such as (a) email discourse in group decision-making (Sproull and Kiesler, 1991) (b) collaborative filtering or recommendation systems (Beenen et al., 2004); (c) blogging as community forums (Nardi et al., 2004); (d) wiki systems (Cunningham and Leuf, 2001); and (e) user-generated tag clouds (Golder and Huberman, 2006). Can we use social and behavioral science theories to study these systems? For example, due to the scale and nature of this collective effort, we might apply research methodology from social complexity theory (e.g., Mathews et al., 1999) or social psychology theories (e.g., Beenen et al., 2004; Cosley et al., 2006, 2007; Kreijns et al., 2003; Latane et al., 1979), or socioeconomic theories on commons problems might also apply in large-scale online communities (e.g., Adar and Huberman, 2000).

Wikipedia

Wiki technology provides simple methods for collective authorship on the Web. Wikimedia Foundation, Inc.’s Wikipedia® is a premier example of a large-scale social computing system in which participants collectively author encyclopedic information. The most common finding is that Wikipedia dynamics follow power law distributions in both readership (Priedhorsky et al., 2007) and editing (Almeida et al., 2007; Kittur et al., 2007; Voss, 2005). The Web also has a scale-free, power law distribution in its link structure (Barabási and Albert, 1999; Broder et al., 2000) and surfing behavior (Huberman et al., 1998). In Wikipedia, the intensity of authorship shows that a tiny number of articles receive the majority of edits, and the vast majority of articles receive a small number of edits
(i.e., the “long tail”). The mathematical theory of Lotska’s law predicts that in scientific authorship, there is an inverse power relationship (e.g., \( w \sim n^{-a} \)) between the number of authors \( n \) and the size of their contributions \( w \). Zipf’s law is a reformulation of this principle generalized to individual contributions among group effort—that is, the rank \( r \) of an individual is proportional to the inverse of her contributions \( n \) (e.g., \( n \sim r^{-k} \)) (Almeida et al., 2007).

Another key finding is that Wikipedia articles have uncertain quality. Most research for computing trust metrics neglect communication theory on how people judge information as trustworthiness, and instead focus on available data from provenance, content, and links\(^2\). Others have experimented with content analysis to detect or highlight possible information quality problems (e.g., Luyt et al., 2008). The authors themselves struggle to control article content around information types, responsibility, perspectives, organization, or provenance and creation (Miller, 2005; Sundin and Haider, 2007). But these processes are largely invisible to readers. Is authorship transparency important to readers of Wikipedia articles? WikiScanner is a data mining tool that collects article edit histories from anonymous unregistered authors (Griffith, 2007). Its purpose is to increase transparency behind authorship by estimating the affiliation of anonymous authors using its database of 34 million edits from 0.2 million organizations between 7 February 2002 and 4 August 2007. Google, Inc.’s “Knol Project” distinguishes itself from the Wikipedia\(^{®} \) approach by requiring article attribution (Manber, 2007) to provide more transparency of authorship to the reader. No further research, however, has been done to measure the utility of this additional transparency from WikiScanner and Knol. Indeed, the utility of explicit authorship information is debatable (Viégas, 2005, p. 61):

> Explicit authorship of contributions on wiki pages is an issue of some contention among wiki users; whereas some feel that authorship is an important part of social collaboration in the sense that it adds context to interactions, others feel that authorship data is irrelevant and sometimes even detrimental to the creation of truly communal repositories of knowledge.

**Tagging**

*Collaborative tagging* or “folksonomy” is a categorization method via collective authorship. Bookmarking services (e.g., Keller et al., 1997; Lee, 2006) popularized this method, and rely on massive distributed participation in which individuals tag content with keywords from an uncontrolled or pseudo-controlled vocabulary. Users then navigate the categorization scheme via a “tag cloud” or via search (Shilad et al., 2006; Voss, 2007). Some recent research has found empirical evidence that collaborative tagging systems exhibit a power law distribution, commonly found in complex systems (Halpin et al., 2007). Some argue that tags are content themselves and not only metadata (Berendt and Hanser, 2007). Geotagging is another tagging process to assign geographic location to content (Amitay et al., 2004), and it is a popular method for sites like Flickr\(^{™} \).

\(^2\)A notable exception is Chesney (2006) who conducted an experiment of expert opinion on the credibility of Wikipedia articles through a bifurcated survey. The first group \((N = 30)\) evaluated an article in their own expertise, and the second group \((N = 24)\) evaluated a randomly selected article. He measured credibility for author, article, and site using metrics based on prior media effects research (e.g., Flanagin and Metzger, 2000; Flanagin, 2005). He found that only article credibility metrics, not author or site credibility, differed between the expert and non-expert groups.
Mashups

Mashups are simply web pages that combine information from other web pages (Ankolekar et al., 2007), but their common use is geographic. Mashups provide the “ability to superimpose geographic information from sources distributed over the Web” (Goodchild, 2007, p. 36). The adoption of simplified data markup (e.g., microformats) promotes the sharing of geographic information by laypersons (Kuhn, 2007). Today, mashups use a geographic map as the metaphor for information integration (Maguire, 2007), where several distinct information sources conflated to present new information. Google Earth™ is a such an example that conflates multiple heterogeneous information sources into a seemingly unique authorship.

2.3 Geographic information in Web 2.0

Digital geographic information has traditionally been authored by centralized authorities, and now so-called “Digital Earth” systems use visualization and Internet technologies to share geographic information. Gore (1999) first defined Digital Earth as “a multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data.” Grossner et al. (2008) later define a digital earth system as “a comprehensive, distributed geographic information and knowledge organization system.” We have many examples of digital earth systems now, such as (a) Google Earth™ with its meteoric growth to 200 million downloads since it launched in June 2005 (Butler, 2006; Google, Inc., 2007); (b) Microsoft’s TerraServer® (Barclay et al., 2000); (c) NASA’s Digital Earth Testbed (de la Beaujardire et al., 2000); and, (d) UCSB’s Alexandria Digital Library (Frew et al., 2000). What these visions have not anticipated adequately is the immense utility of user-generated content, or collective authorship such as Wikipedia and mashups.

Nevertheless, current digital earth systems allows users to create, share, and pool geographic information via the Internet, creating a global digital commons of geographic knowledge. Perhaps phenomenal success of these systems, and Google Earth™ in particular, is driving an increased interest in digital earth research. Goodchild (2008) argues that Digital Earth qualifies as a “grand challenge” for researchers, and claims that some of its most problematic issues are institutional in nature. Yet, thus far digital earth research has been mostly technological in nature (Foresman, 2008; Goodchild, 1999, 2000; Grossner, 2006; Grossner et al., 2008; Maguire, 2007; Yongxiang, 1999), with little attention to sociological or institutional behaviors which are critical to understanding how we use technology in our lives3. Moreover, Digital Earth technology is enabling a new media itself (Sui and Goodchild, 2001, 2003). While current research has focused on technological issues of visualization and data organization, new forms of collective authorship have emerged. For example, collaborative mapping, or so-called “neogeography” – geography without geographers (Sui, 2008), and public participation GIS have advanced from small-group, face-to-face applications to large-scale, highly-distributed communities. Recent research on neogeography (e.g., Balram and Dragičević, 2006) suggests needs for new information frameworks and behaviors that account for these types of communities.

3Science and technology studies have argued for a deeper understanding of how technology impacts society and how social factors influence technology development and diffusion (e.g., Jasanoff et al., 1995; Sawyer and Rosenbaum, 2000; Williams and Edge, 1996).
3 Data and Methods

In this section, I describe how Wikipedia articles are geotagged and my analysis of authorship behavior. My methodology uses data extraction and mining methods for collecting corpus data, and spatial clustering methods for analyzing authorship footprints.

3.1 Geotagging

The geotagging process itself is haphazard. In general, a geotag may contain geographic coordinates, extent, shape, and feature type information. But most commonly, a geotag contains simple geographic coordinates for latitude and longitude (e.g., 34° 24’ 35" N, 119° 50’ 59" W). Wikipedia started explicitly using structured geotagging in February 2005 with the introduction of the Template:Coor Wikipedia template (wikipedia.org, 2008), and now many Wikipedia templates accept geographic coordinates, such as Template:Infobox_Mountain. In fact, there are dozens of ways to include geographic coordinates in an article or webpage – there is not a single “geotag” standard or format (see Table 2 for examples).

Many Wikipedia articles are about a place, and thus geotagging the article will locate the place on a map. Some authors create geotags manually using a reference digital or paper map to estimate coordinates, while others resolve toponyms based on existing online gazetteers. An example of the latter is The Anomebot2\(^4\), a semi-automated bot that has geotagged over 100,000 Wikipedia articles. It runs periodically to create geotags based on pattern matching against online gazetteers, including GEOnet Names Server, Geographic Names Information System, and Ordnance Survey Great Britain, and with other data mined from Wikipedia articles. But, anecdotally, it appears as the majority of geotags were created manually and not via automated processes (T. Alder, personal communication, 22 April 2008). This further obscures the lineage of these coordinate data.

The Wikipedia-World project\(^5\) creates a catalog geotagged articles. Since geotagging in Wikipedia is chaotic, this process relies on data mining methods and is largely heuristic. In May 2008, this process found 264,288 unique geotags across 230 Wikipedia databases used by 1,163,797 articles. Wikipedia-World use these data to provide various mapping services for Wikipedia users, such as WikiMiniAtlas\(^6\) and Google Earth views of geotagged Wikipedia articles. The underlying data are exported as a database tables. The first table\(^7\) has one record per unique geotag. The second table\(^8\) has an association record for each geotagged article across all Wikipedia’s languages.

3.2 Data extraction

Wikipedia provides article and metadata via periodic dumps\(^9\) of their database and as static HTML files, but these data do not always include complete article contribution records due to their volume and limited operational resources. For example, as of August 2008, the English Wikipedia had 2.5 million articles and 250 million contributions\(^10\). Instead of relying on these data dumps, I

\(^4\)http://en.wikipedia.org/wiki/User:The_Anomebot2
\(^6\)http://meta.wikimedia.org/wiki/WikiMiniAtlas
\(^7\)http://tools.wikimedia.de/~kolossos/wp-world/pub_C_geo_id.sql.gz
\(^8\)http://tools.wikimedia.de/~kolossos/wp-world/pub_C_lang.sql.gz
\(^9\)http://tools.wikimedia.de/~kolossos/wp-world/pub_C_geo.sql.gz
perform data extractions using SQL directly from live replica databases provided by the Wikimedia Toolserver project\(^{11}\). These databases use MySQL and the Mediawiki database schema\(^{12}\). In brief, the Mediawiki schema organizes the Wikipedia articles by revision. The revision table provides metadata for author contributions and links to the page and text table for details on the article’s contents. For every article, the page table contains a unique identifier and the title for the article, and the text table stores the article’s contents as a blob.

As described above, the Wikipedia World project performs periodic data mining of geotagged articles (Kühn and Alder, 2008). I use their data from 10 May 2008 to identify geotagged articles from 21 different languages (see Table 4 for a complete list of languages). I then extract all the authoring history and the most recent version for each geotagged article from the replica databases. I migrate the authoring histories into a single shared database with Mediawiki tables modified to associate a source language for each record (e.g., for the page table, page_id and a new page_lang column comprise the primary key instead of only page_id), and to remove data incidental to my analysis.

The corpus of geotagged articles from 21 Wikipedia databases (see Table 1) contains 990,315 articles, 3,429,940 authors, and 32,141,334 author contributions between 2001 and 2008. The data extraction from the Wikipedia tables results in page and text with 990,315 records, revision with 32,141,334 records, and user with 578,448 records. The latter table contains records only for registered authors, so I mine the revision.rev_user_text column to identify IP addresses for anonymous users, and integrate them into my analysis schema. My schema provides a consistent framework for the Mediawiki data across any number of languages, and has tables for article, author, and geotag data, and author_article and geotag_article association tuples. It also provides optimized access to summary statistics for both per article, and per author, such as the number of contributions by an author or the number of authors for an article.

### 3.3 Authorship behavior

Authors are of three types or “groups”: registered, anonymous, and bots. Registered authors have an account with Wikipedia and also a username. Anonymous authors do not have accounts, and the author’s IP address is used in lieu of a username. Bots are a special type of registered author to denote that the author is actually an automated agent. There are also other special types of registered authors\(^ {13}\) for various editorial and administrative functions, which I combine into a single “other” category in my analysis.

For each author, I define a “footprint” as the set of contributions made to any geotagged article by that author. Every author has a single footprint, and every article may be included in one or more footprints. Since each article is geotagged, footprints also have a spatial dimension. Figures 3, 4, & 5 show the spatial dimension for example footprints of registered, anonymous, and bot authors.

\(^{11}\) <http://toolserver.org>

\(^{12}\) <http://www.mediawiki.org/wiki/Manual:Database_layout>

\(^{13}\) <http://en.wikipedia.org/wiki/Wikipedia:User_access_levels>
Power laws

As with Wikipedia in general, authorship in the geotagged subset of Wikipedia follows power law distributions (see Figure 1). As discussed earlier, other research finds that author contributions follow power laws, that is, the “long tail” where most of the vast majority of authors contribute less overall than a small active minority who contribute the most (Almeida et al., 2007; Priedhorsky et al., 2007; Kittur et al., 2007; Voss, 2005). In my corpus, anonymous and registered authorship both follow power law distributions better ($R^2 = 0.90$ and $0.83$, respectively) than the “other” and bot authors do ($R^2 = 0.75$ and $0.44$, respectively). The distributions have increasingly strong fits as the authorship base increases, from 2.3 thousand bots up to 2.8 million anonymous authors. But even with a small base, bots perform 39% of all Wikipedia contributions in general\footnote{http://stats.wikimedia.org/EN/BotActivityMatrix.htm} and 26% for the geotagged corpus.

Table 3 shows summary authorship statistics organized by groups. These data show that even though anonymous users outnumber registered users four-to-one, registered users rank highest in contributions. This difference is reflected in a steeper slope of the anonymous contributions; that is, the anonymous authors have a heavier “long tail” than registered authors. Table 4 has summary authorship statistics organized by language. In my corpus, the authors and contributions both follow power law distributions across languages better ($R^2 = 0.92$ and $0.86$, respectively) than articles or spatial coverage do ($R^2 = 0.60$ and $0.63$, respectively). Table 5 shows the correlation matrix between variables for the group and language summary statistics. Between the languages, authors and contributions are correlated as expected, but so are articles and places. That is, as the number of geotagged articles increases, so does the number of distinct places. This demonstrates that articles are not significantly clustered spatially across languages. None of the variables are correlated when organized by authorship group. This demonstrates that each group exhibits significantly different authorship behavior.

Spatial distribution

Figure 2 shows the density of geotagged Wikipedia articles, dominating in the USA and Europe. The spatial clustering is a simple global grid technique. In my analysis, I use a one-tenth degree global grid containing 6,480,000 cells, the average area of which is 78 square kilometers, roughly the size of a big city. I define a “place” as one grid cell. Authoring activity is present only in a small fraction of the global space. For the entire corpus, only 103,310 cells, or only 1.6% of the global space contains one or more articles. This is likely an underestimate since Wikipedia’s geotags contain point coordinates only, and thus complete extent information is not available\footnote{For an example of extent-based geotagging, see Wikimapia.org which uses polygons to define their geotags.}.

Table 3 & 4 also measure spatial locality, a normalized index for spatial coverage where a higher value denotes more narrow spatial coverage. The locality between authorship groups shows that anonymous authors tend to contribute to fewer distinct places than other groups. The locality between languages, however, appears idiosyncratic.
4 Future work

Data quality

Some bots may not be identified correctly in the analysis. Using a simple pattern match in the username, I found 856 registered authors who were not identified as bots but may actually be bots. These authors had an order-of-magnitude more contributions than the average registered author, a combined 369,625 contributions. Some of this discrepancy might be from the fact that some Wikibots are (mis)classified as registered users rather than as bots.

Author’s Location

Since 80% of the authors are anonymous, I have access to IP addresses. MaxMind’s GeoIP product has an IP to geolocation database down to the city-level. Their database is primarily based on IANA registration records, and thus the location information is generally administrative in nature, such as a corporate headquarters. However, a country-level resolution may provide more accurate spatial locations for authors, and thus be useful to add to the footprint data. See Figure 4 for an example.

Classification analysis

To discover behavioral patterns in the footprints dataset, I will use multi-dimensional clustering and classification Berkhin (2006), where the clustering dimensions will be geographic coordinates, shape, feature and geotag types, authorship effort, and time. My results would consider spatial factors and provide empirical evidence on information behavior in collective authorship.

Socio-spatial network analysis

The editorial process in Wikipedia is highly formalized despite the informal nature of authorship (Viégas et al., 2007). Do spatial factors influence these informal ties? Using the spatial clustering of a one-tenth degree resolution grid we can derive a social network based on spatial relationships. That is, authors who contribute to articles about similar locations (i.e., in the same grid cell) would have a connection in the social network. Edges would be weighted by their relative contributions to the articles in that grid cell. The scale of this network would require some clustering or thresholding methods, such as a Statistical Information Grid (STING) approach – a grid-based multiresolution analysis (Wang et al., 1997). A version of this network is available online16.

Gazetteer-based geotagging

Currently, most geotags in Wikipedia are not automatically derived from gazetteer information. Instead the location information is derived by hand from an individual author using Google Maps or another georeferencing tool. Moreover, the geotagging methods used in Wikipedia lack of extent or shape information, and thus rely solely on point coordinates. Perhaps authors should create geotags based on a place from a gazetteer to preserve the provenance and expand the geographic information beyond simple point data.

16 http://toolserver.org/~drh08
5 Conclusion

While current GIS research has focused on technological issues of visualization and data organization, new forms of collective authorship such as collaborative mapping, or so-called “neogeography,” suggests needs for new information frameworks and behaviors. I use data mining methods to discover existing authorship behavior patterns in a subset of geotagged articles from Wikipedia.

My data show authoring behavior on geotagged Wikipedia articles differed between groups of anonymous and registered authors, and bots, but not between languages. As a group, registered authors contribute the most overall and they have a near complete spatial coverage over all authored places, but they are small minority of the author population. The bots group are the next in contributions, but are only a tiny faction (<1%) of the author population. Finally, the anonymous group makes less contributions than bots but they are the vast majority of the user population, and they have the most narrow spatial coverage over authored places of all the groups. I find no correlation between the groups on either spatial locality or contribution work metrics. Between languages, I find correlation between spatial coverage and articles, and between authors and contributions.

6 Acknowledgements

I would like to thank Wikimedia Toolserver\(^{17}\), a service of Wikimedia Deutschland e.V., for providing computing resources and access to Wikipedia databases, and Wikipedia’s WP-World project\(^{18}\) (Tim Alder and Stefan Kühn, in particular), for their help with the geotag data mining processes.

| Table 1: Summary of geotagged corpus extracted from 21 Wikipedia databases. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | 21              | 103,310         | 990,315         | 3,429,940       |
| languages       | places          | geotagged articles | authors       |
| 22 February 2001 | 17,560,994     | 32,141,334      | earliest author contribution |
| 13 August 2008  | latest author contribution |
| 10 May 2008     | geotag data mining completion |

\(^{17}\) http://toolserver.org

Table 2: Example geotag formats using the coordinates of the University of California, Santa Barbara (approx. 34° 24’ 35” N, 119° 50’ 59” W).

(a) Template:Coor (Wikipedia)

```
{{coor dd|34.41|-119.85}}
{{coor dms|34|24|35|N|119|50|59|W}}
```

(b) Geo microformat for HTML (Çelik, 2005)

```html
<DIV CLASS='geo'>University of California, Santa Barbara:
    <SPAN CLASS='latitude'>34.41</SPAN>,
    <SPAN CLASS='longitude'>-119.85</SPAN>
</DIV>
```

(c) Dublin Core for HTML (Kunze, 1999)

```html
<META NAME='DC.coverage.x' CONTENT='-119.85'>
<META NAME='DC.coverage.y' CONTENT='34.41'>
```

(d) GEO for HTML (Daviel and Kaegi, 2007)

```html
<META NAME='geo.position' CONTENT='34.41;-119.85'>
<META NAME='geo.placename' CONTENT='University of California, Santa Barbara'>
<META NAME='geo.region' CONTENT='US-CA'>
```

Table 3: Authoring behavior based on author type or “group.” Authors and contribs are distinct counts, articles and places are percent coverage out of a possible 990,315 articles and 103,310 places, and work and locality are both normalized indexes of contributions per article and places per article, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Authors</th>
<th>Articles</th>
<th>Contribs</th>
<th>Work</th>
<th>Places</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>anonymous</td>
<td>2,850,813</td>
<td>44.3%</td>
<td>7,304,171</td>
<td>0.5</td>
<td>82.7%</td>
<td>1.9</td>
</tr>
<tr>
<td>registered</td>
<td>567,244</td>
<td>73.8%</td>
<td>11,333,151</td>
<td>0.5</td>
<td>97.6%</td>
<td>1.3</td>
</tr>
<tr>
<td>other</td>
<td>8,821</td>
<td>57.8%</td>
<td>5,207,218</td>
<td>0.3</td>
<td>94.5%</td>
<td>1.6</td>
</tr>
<tr>
<td>bot</td>
<td>2,298</td>
<td>93.0%</td>
<td>8,268,549</td>
<td>0.3</td>
<td>93.8%</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>3,429,176</td>
<td>990,315</td>
<td>32,113,089</td>
<td>103,310</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Authoring statistics for each Wikipedia database by language. The variables are defined the same as Table 3, except articles is a distinct count.

<table>
<thead>
<tr>
<th>Language</th>
<th>Authors</th>
<th>Articles</th>
<th>Contribs</th>
<th>Work</th>
<th>Places</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1,564,815</td>
<td>198,077</td>
<td>10,731,813</td>
<td>1.7</td>
<td>91.0%</td>
<td>4.5</td>
</tr>
<tr>
<td>German</td>
<td>620,898</td>
<td>103,418</td>
<td>5,090,612</td>
<td>1.5</td>
<td>40.0%</td>
<td>3.8</td>
</tr>
<tr>
<td>French</td>
<td>250,718</td>
<td>78,408</td>
<td>3,203,403</td>
<td>1.3</td>
<td>30.2%</td>
<td>3.8</td>
</tr>
<tr>
<td>Spanish</td>
<td>208,036</td>
<td>43,077</td>
<td>1,685,382</td>
<td>1.2</td>
<td>23.4%</td>
<td>5.4</td>
</tr>
<tr>
<td>Italian</td>
<td>130,363</td>
<td>79,904</td>
<td>2,007,312</td>
<td>0.8</td>
<td>30.3%</td>
<td>3.8</td>
</tr>
<tr>
<td>Japanese</td>
<td>122,492</td>
<td>22,616</td>
<td>917,430</td>
<td>1.3</td>
<td>12.5%</td>
<td>5.5</td>
</tr>
<tr>
<td>Portuguese</td>
<td>91,363</td>
<td>77,804</td>
<td>1,205,596</td>
<td>0.5</td>
<td>40.7%</td>
<td>5.2</td>
</tr>
<tr>
<td>Poland</td>
<td>88,145</td>
<td>70,303</td>
<td>1,357,792</td>
<td>0.6</td>
<td>27.3%</td>
<td>3.8</td>
</tr>
<tr>
<td>Dutch</td>
<td>84,490</td>
<td>127,653</td>
<td>1,893,587</td>
<td>0.5</td>
<td>50.5%</td>
<td>3.9</td>
</tr>
<tr>
<td>Russian</td>
<td>49,509</td>
<td>38,478</td>
<td>726,508</td>
<td>0.6</td>
<td>18.4%</td>
<td>4.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>35,878</td>
<td>21,610</td>
<td>480,577</td>
<td>0.7</td>
<td>14.4%</td>
<td>6.6</td>
</tr>
<tr>
<td>Chinese</td>
<td>32,246</td>
<td>12,425</td>
<td>426,208</td>
<td>1.1</td>
<td>9.3%</td>
<td>7.4</td>
</tr>
<tr>
<td>Finnish</td>
<td>30,816</td>
<td>11,687</td>
<td>383,654</td>
<td>1.0</td>
<td>8.3%</td>
<td>7.0</td>
</tr>
<tr>
<td>Turkish</td>
<td>28,340</td>
<td>7,500</td>
<td>200,168</td>
<td>0.8</td>
<td>4.7%</td>
<td>6.2</td>
</tr>
<tr>
<td>Norwegian</td>
<td>24,130</td>
<td>18,669</td>
<td>388,181</td>
<td>0.6</td>
<td>13.5%</td>
<td>7.1</td>
</tr>
<tr>
<td>Czech</td>
<td>18,123</td>
<td>15,210</td>
<td>328,631</td>
<td>0.7</td>
<td>6.4%</td>
<td>4.1</td>
</tr>
<tr>
<td>Danish</td>
<td>14,225</td>
<td>6,404</td>
<td>188,376</td>
<td>0.9</td>
<td>4.7%</td>
<td>7.3</td>
</tr>
<tr>
<td>Catalan</td>
<td>13,675</td>
<td>12,080</td>
<td>286,678</td>
<td>0.7</td>
<td>6.8%</td>
<td>5.6</td>
</tr>
<tr>
<td>Esperanto</td>
<td>9,861</td>
<td>28,730</td>
<td>376,285</td>
<td>0.4</td>
<td>14.0%</td>
<td>4.8</td>
</tr>
<tr>
<td>Slovak</td>
<td>9,683</td>
<td>14,603</td>
<td>182,894</td>
<td>0.4</td>
<td>5.4%</td>
<td>3.7</td>
</tr>
<tr>
<td>Icelandic</td>
<td>2,134</td>
<td>1,659</td>
<td>52,002</td>
<td>1.0</td>
<td>1.4%</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>3,429,940</td>
<td>990,315</td>
<td>32,113,089</td>
<td></td>
<td>103,310</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Correlation matrix for authoring statistics from Tables 3 & 4.

<table>
<thead>
<tr>
<th>By Language By Group</th>
<th>Authors</th>
<th>Articles</th>
<th>Contribs</th>
<th>Authors</th>
<th>Articles</th>
<th>Contribs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>0.81</td>
<td>1.00</td>
<td>0.98</td>
<td>-0.72</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Contribs</td>
<td></td>
<td>0.89</td>
<td>1.00</td>
<td>-0.02</td>
<td>0.44</td>
<td>1.00</td>
</tr>
<tr>
<td>Places</td>
<td>0.85</td>
<td>0.99</td>
<td>0.90</td>
<td>-0.90</td>
<td>0.67</td>
<td>0.37</td>
</tr>
</tbody>
</table>
**Figure 1:** Power law distributions for author contributions by group. Each line represents a power law regression on a log-log scale with $R^2$ values in parentheses. The anonymous authors best fit a power law distribution with $R^2 = 0.90$, and the distributions are progressively steeper from bots to anonymous.

![Graph of power law distributions for author contributions by group.](image)

**Figure 2:** Spatial distribution of geotagged Wikipedia articles with density on a log-scale for number of articles per location. Source: (Alder, 2007).

![Spatial distribution of geotagged Wikipedia articles.](image)
Figure 3: Example footprint of a registered author with 1,099 contributions to 296 articles in the Danish Wikipedia. Visualization uses KML and Google Earth™. Blue information icons represent geotagged location of each article edited by author, the vast majority of which are clustered inside Denmark.
**Figure 4:** Example footprint of an *anonymous* author with 172 contributions to 143 articles in the Danish Wikipedia. Visualization uses KML and Google Earth™. The yellow house icon represents an estimate of the author’s own location. Since the IP address of the anonymous authors are publicly available, I use IP-based geolocation (MaxMind, Inc., 2008) to estimate author location.
**Figure 5:** Example footprint of a *bot* with 3,006 contributions to 1,601 articles in the Danish Wikipedia. Visualization uses KML and Google Earth™.
References


Berendt, B. and C. Hams (2007). Tags are not metadata, but ‘just more content’ - to some people. In International Conference on Weblogs and Social Media, Boulder, CO.


Foresman, T. W. (2008). Evolution and implementation of the Digital Earth vision, technology and soci-


