Chapter 11
The Geographic Nature of Wikipedia Authorship

Darren Hardy

Abstract The efficacy and use of volunteered geographic information (VGI) is an active research area, but the geography of VGI authorship is largely unknown. Wikipedia is an online collaborative encyclopedia where anyone can edit articles, including those about place. Moreover, Wikipedia's editorial transparency facilitates *in situ* observations of collective authorship. The empirical study described in this chapter collects 32 million contributions to Wikipedia's geographic articles over 7 years. It finds exponential decay in the spatial patterns of Wikipedia's authorship processes, which is consistent with other sociospatial phenomena, like innovation diffusion. As global information infrastructures continue to reduce communication and coordination costs, this study may provide insight into whether geographic distance ultimately matters in information peer production. This chapter begins by discussing core concepts behind collective authorship; then provides an overview of Wikipedia, its contributors, and their production processes; discusses the results and implications from spatial modeling of geotagged Wikipedia article contributions; and concludes with future research issues.

11.1 Introduction

A notable example of a widely popular system with volunteered geographic information (VGI) capabilities is Wikipedia, an online collaborative encyclopedia. Wiki technology provides simple methods for Web-based collective authorship where anyone can contribute. Using this technology, Wikipedia provides a large-scale social computing system in which participants collectively author encyclopedic information.

D. Hardy (✉)
Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA
e-mail: dhardy@bren.ucsb.edu

D. Sui et al. (eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, DOI 10.1007/978-94-007-4587-2_11,

Since 2001, Wikipedia has 17.5 million articles in 263 languages. Since March 2007, Alexa has ranked Wikipedia in the top 10 Internet sites. As of 23 February 2010, Wikipedia has 15 million articles in 272 languages with 860 million edits from 22 million contributors (Wikimedia 2010). During 2009 alone, Wikipedia had 365 million unique visitors that generated 133.6 billion page views (Zachte 2010a). Its impact on the Web's content is significant. Fifty-one percent of its site visits come from link-based search engine referrals (Alexa Internet, Inc. 2009). Of those page views that were referred to Wikipedia by external sites, 42% were referred by Google search, maps, and other services, and 8% were made by Google's "web-crawling" software GoogleBot (Zachte 2009). Over 1.2 million articles are place-based articles (i.e., "geotagged") (as of April 2011). These geotagged articles span dozens of languages and are accessible through geobrowsers and online mapping services.

As the Internet itself grows, many describe it as place*less*—cyberspace without place. Yet sociological researchers find cultural differences in virtual communities that mimic real-world environments, and a shared understanding of a virtual place is a central determinant in such research. But today, any Internet user can get some sense of place through rich interactive geovisualization technologies. "Slippy maps" depict roads and buildings and other geographical features using simple point-and-drag navigational and informational tools and even 3D imagery. Within these online mapping interfaces, users may access a diverse set of VGI, including geotagged Wikipedia articles and photographs.

Yet, despite the advantages of the Internet for collaborative work, authors are fundamentally engaged in knowledge production processes that are grounded in social structures and norms, and in turn, physical place. Geographic distance, in particular, should be a significant factor in online knowledge production. But the nature of the Internet in a globalized world has led to debate on whether geographic distance matters (cf. Cairncross 1997; Friedman 2005; Goodchild 2004; Marston et al. 2005). That is, the Internet may redefine the role of physical place in our lives due to reduced communication costs and increased ubiquity. Zook (2005, p. 54) summarizes this debate as a new "geography of electronic spaces," as the Internet becomes "a recombinant space for political, cultural, and economic interaction."

This chapter focuses on information production methods and processes behind geographic Wikipedia articles and discusses the nature of these production processes. For example, are contribution patterns similar between VGI and non-VGI content? How do authors geotag articles? What is the geography of Wikipedia's authorship? What is the spatial distribution of articles and contributors, and how does physical proximity influence contributions, either by article topic or language?

11.2 Collective Authorship Processes

Collective authorship is one type of information production process—a mass collective effort by individuals to produce information artifacts within a digital commons. The term "information production" itself has different semantics across

disciplines. In the humanities, the term may represent the authoring of a written work or book; in economics, market resources, or commodities, or perception, or even a constitutive force in society (Browne 1997, p. 266); in library science, how we communicate collaborative work to public scientific knowledge (Cronin 2001); and, in social computing, collaborative filtering or recommendation systems (Beenen et al. 2004), blogging as community forums (Nardi et al. 2004), and user-generated tag clouds (Golder and Huberman 2006). For Wikipedia, the terms wikinomics (Tapscott and Williams 2006), collective intelligence (O'Reilly 2005), and crowdsourcing (Brabham 2008) all reflect the user-centric processes that drive information production.

And user-centric it is. Each month, over ten million authors contribute to Wikipedia articles, roughly divided into two classes of contributors—a small, highly productive set, then everyone else. The Web itself has a scale-free, power law distribution in its link structure (Broder et al. 2000) and surfing behavior (Huberman et al. 1998), and Wikipedia has them for both readership (Priedhorsky et al. 2007) and editing (Almeida et al. 2007; Kittur et al. 2007; Voss 2005). For example, the intensity of authorship shows that a small number of Wikipedia articles receive the majority of edits, and the vast majority of articles receive a small number of edits (i.e., the long tail).[1]

Wikipedia's production processes are nontrivial, despite its perception in the popular media as a loose or chaotic system. Wikipedia has many policies and mechanisms to govern contributions, including rule-making, monitoring, conflict resolution, and norms (Forte and Bruckman 2008; Lih 2009; Viégas et al. 2007a, b). Its most well known policy is that contributors must write articles using a neutral point of view, and this is a key discussion point between authors (e.g., Bryant et al. 2005; Viégas et al. 2004). As described by Wikipedia, *neutral point of view* (NPOV) is "a fundamental Wikimedia principle and a cornerstone of Wikipedia," requiring that "all content [be] written from a neutral point of view, representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources" (http://en.wikipedia.org/wiki/NPOV).

The term *Wikipedian* does not have a strict definition, other than being a contributor to a Wikipedia article generally.[2] Registered, anonymous, administrative Wikipedians and bots are the four basic types of contributor. *Registered* Wikipedians create an account on Wikipedia, and their contributions are explicitly tagged in the article history using their account. *Anonymous* Wikipedians do not provide any

[1] In scientific authorship, Lotska's law predicts an inverse power relationship (e.g., $w \sim n^{-\beta}$) between the number of authors n, the size of their contributions w, and a constant β. Zipf's law is a reformulation of this principle, generalized to individual contributions among group effort — i.e., the rank r of an individual is proportional to the inverse of her contributions n (e.g., $n \sim r^{-\beta}$) (Almeida et al. 2007).

[2] Some have a more narrow definition of highly or consistently active contributors (Zachte 2010b), but in this chapter, "Wikipedian" refers to any contributor to a Wikipedia article, regardless of activity level.

registration information, and their computer's IP address is used in lieu of an account. "Bots" and other *administrative* Wikipedians are both special cases of registered accounts, but they have additional access or permissions to edit articles. The overwhelming majority of Wikipedians do not collaborate with each other in a traditional sense. They do not often discuss their contributions with others (Viégas et al. 2007a) and as such form a loosely collaborative, online collective authorship. The most active segments of the Wikipedian population are 91,817 Wikipedians with at least five contributions *per month* and 1,076,908 Wikipedians with at least ten contributions total (Zachte 2010b). The "long tail" has 21.1 million Wikipedians, each of whom have less than ten contributions total.

Although authorship processes are largely invisible to readers, the authors themselves struggle to control article content around information types, responsibility, perspectives, organization, or provenance and creation (Miller 2005; Sundin and Haider 2007). Wikipedia provides complete article histories for those readers wanting detailed authorship information. *WikiScanner*, for example, is a data-mining tool that extrapolates from article edit histories the location or affiliation of anonymous authors (Griffith 2007). But the utility of explicit authorship information is debatable. As summarized by Viégas (2005, p. 61), on the one hand explicit authorship information may be "an important part of social collaboration in the sense that it adds context to interactions," and on the other hand it may be "irrelevant and sometimes even detrimental to the creation of truly communal repositories of knowledge."

In fact, the success of Wikipedia and other "user-generated content" Web services (O'Reilly 2005) has challenged academic theories of production. Benkler (2002) argues that in terms of economic models of production, when the efficiency gains of "peering" exceeds the costs of organizing human capital into a firm or market, a commons-based peer production system will emerge. Its advantage is based not only on reduced costs of human capital and communications but also on the nonrival aspects of Web-based information artifacts—i.e., many people can read (consume) a webpage simultaneously without degrading its value. This effectively eliminates allocation costs to consumers and increases the pool of potential contributors, which mitigates effects from free riders.

When applied to geographic information production, these factors will likely challenge the "knowledge politics" of spatial data infrastructures (Elwood 2010). For example, they may weaken traditional notions of authoritative sources as the collective social production of spatial information increases (Budhathoki et al. 2008; Coleman et al. 2009; Sieber and Rahemtulla 2010). As Sui (2008, p. 4) argues, the "wikification of GIS is perhaps one of the most exciting, and indeed revolutionary developments since the invention of [GIS] technology in the early 1960s." Moreover, Wikipedia's editorial patterns in the production of VGI content are similar to those for nongeographic content. That is, each of the four types of contributors exhibits editorial patterns that are systematic when contributing to geographic articles, but idiosyncratic across languages (Hardy 2008).

Table 11.1 Example geotag formats for University of California, Santa Barbara (UCSB; approx. 34.41°N, 119.85°W)

(a) *Template:Coord* and *Template:Infobox* in Wikipedia

```
UC Santa Barbara {{coord|34|24|35|N|119|50|59|W}}
UC Santa Barbara {{coord|34.41254|-119.84813|display=title|type:edu}}
{{Infobox_University |name=UC Santa Barbara
                     |latd=34 |latm=24 |lats=35 |latNS=N
                     |longd=119 |longm=50 |longs=59 |longEW=W
...}}
```

(b) Geo microformat for HTML (Çelik 2005)

```
<DIV CLASS="geo">UC Santa Barbara
    <SPAN CLASS="latitude">34.41</SPAN>,
    <SPAN CLASS="longitude">-119.85</SPAN>
</DIV>
```

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

```
<META NAME="DC.title"       CONTENT="UC Santa Barbara" />
<META NAME="DC.coverage.x" CONTENT="-119.85"/>
<META NAME="DC.coverage.y" CONTENT="34.41"/>
```

(d) Geo metadata for HTML (Daviel and Kaegi 2007)

```
<META NAME="geo.position"  CONTENT="34.41;-119.85"/>
<META NAME="geo.placename" CONTENT="UC Santa Barbara" />
<META NAME="geo.region"    CONTENT="US-CA" />
```

11.3 Volunteered Geographic Information in Wikipedia

Now, we turn to the specific types of geographic information produced through collective authorship in Wikipedia. Geographic information, in general, informs us about the *where* of things. It is spatial information about a phenomenon's distribution in our geographic world (Goodchild 2000). *Georeferencing* is the set of methods for defining a geographic location on the globe (Hill 2006), and *geotagging* assigns geographic locations to content (Amitay et al. 2004), referring to "tagging" georeferenced metadata to a document or other content. A geotag may contain geographic coordinates, extent, shape, or feature type information. A useful geometry for cataloguing georeferenced content is the *minimum-bounding rectangle*, which is the smallest rectangle aligned with the coordinate axes that spans all coordinates for a given location.

Wikipedia primarily uses single points and bounding rectangles rather than fine-resolution polygons in its geotagging processes. In this case, a geotag contains simple geographic coordinates for latitude and longitude, and this georeferenced information is embedded into articles using one of many microformats and extensions to *Wikitext*, Wikipedia's content markup language. For example, the *Template:Coord* and *Infobox* Wikitext templates accept point coordinates (Wikipedia.org 2008). In fact, there are dozens of ways to include geographic coordinates in an article. There is not a single "geotag" standard or format for Wikipedia, or the Web for that matter (Table 11.1).

The geotagging process itself in Wikipedia is haphazard. Wikipedia started explicitly using structured geotagging in February 2005 when geotags were introduced into

Wikipedia in 2005 by Egil Kvaleberg's *gis* extension to *MediaWiki*. Some authors create geotags manually using a reference digital or paper map to estimate coordinates, while others resolve toponyms based on existing online gazetteers. Alternatively, bots perform a bulk of the automated geotagging based on *GEOnet Names Server*, an online gazetteer, and run periodically. This process also adds geographic feature type (i.e., city, river, mountain, etc.) when it is available from the gazetteer.

The vast majority of geotagging is reportedly done by a variety of bots (Kühn and Alder 2008), and their ad hoc nature ultimately makes it more difficult to extract geotags from articles. For example, a semiautomated bot *Anomebot2* runs periodically to geotag articles or mark those that *may* need a geotag. It cross-references named entities in over 100,000 article titles with online gazetteer services.[3] These bots provide a structural mechanism to integrate existing geographic data sources into articles. But they are not semantic in nature, nor do they generate standardized markup (Table 11.1). In fact, they increase the complexity of extracting structured geographic information from articles because of their chaotic, ad hoc nature and that of the Wikitext markup and templates themselves (Sauer et al. 2007). The end result is that geotag extraction requires ad hoc or data-mining approaches to deal with the nondeterministic, semistructured nature of article templates and ad hoc inclusion of geotags. But, anecdotally, some claim the majority of geotags were created manually and not via automated processes (T. Alder, 22 April 2008, personal communication). This further obscures the lineage of these geographic coordinate data.

To index place-based articles, the *Wikipedia-World* project creates a catalog of geotagged articles (Kühn and Alder 2008). Since geotagging in Wikipedia is chaotic, this process relies on data-mining methods and is largely heuristic (Fig. 11.1). In May 2008, this process found 1,163,797 geotagged articles across 230 languages and 234,474 unique locations (at 1 km resolution). Wikipedia-World uses these data to provide various online mapping services and exports the underlying geographic data as database tables. And the index of place-based articles is growing rapidly. In May 2011, the same process found 1.7 million geotagged articles across 273 languages and 1.1 million unique locations (at 1 km resolution, Fig. 11.2).

11.4 Geography of Authorship

In systems like Flickr and Wikipedia, VGI content itself is spatially clustered (Hecht and Gergle 2010), and Wikipedia articles are also more likely to link to articles about places nearby (Hecht and Moxley 2009). But the literature does not directly address whether

[3] These services include *GEOnet Names Server* (GNS) and *Geographic Names Information System* (GNIS) (http://en.wikipedia.org/wiki/User:The_Anomebot2). Using gazetteers as data sources is common for these automated processes, but there are other data sources in use. *Rambot*, for example, uses its own database of 3,141 counties and 33,832 cities to create geographic articles (http://en.wikipedia.org/wiki/User_talk:Rambot).

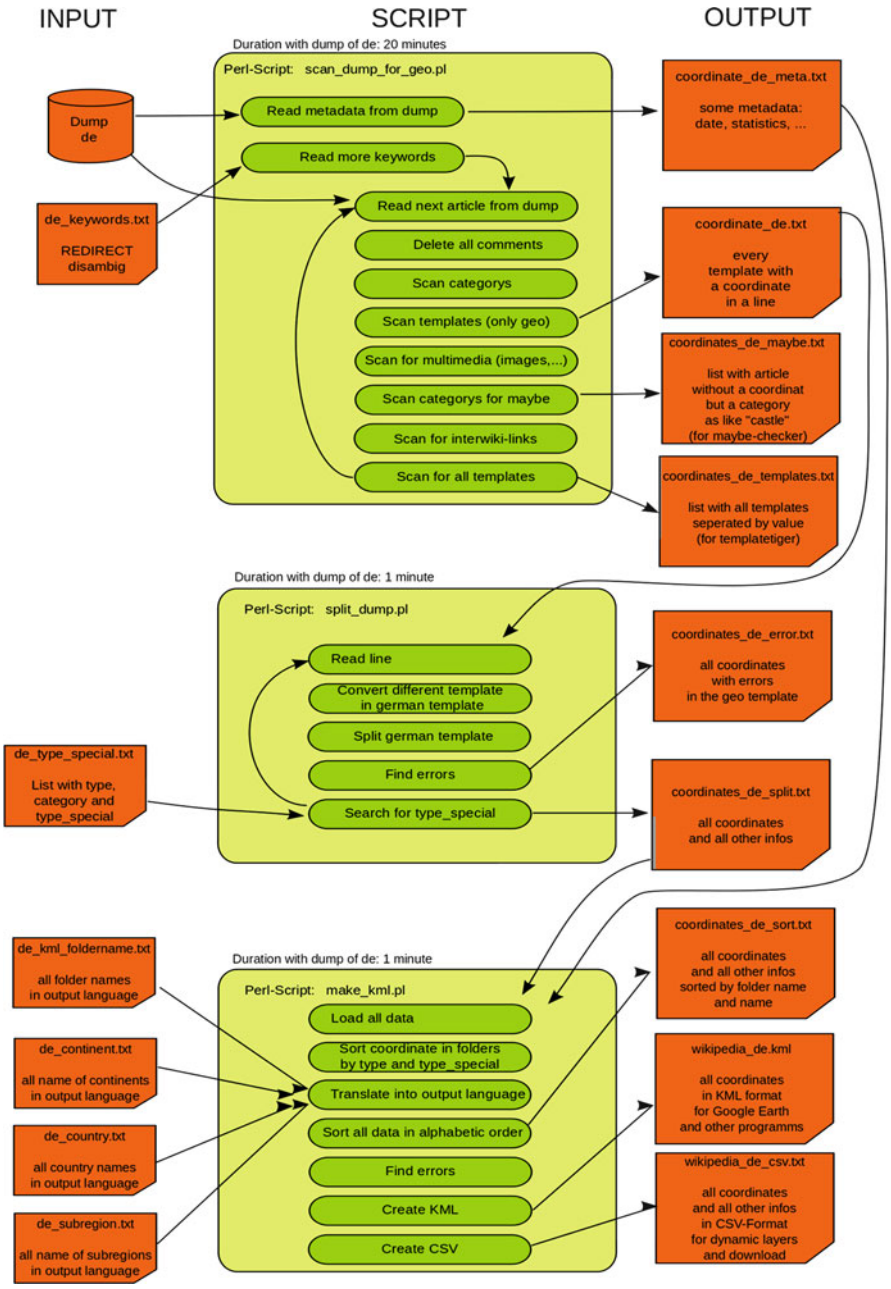

Fig. 11.1 Detailed workflow for geotag data-mining software (Reprinted from Kühn 2008)

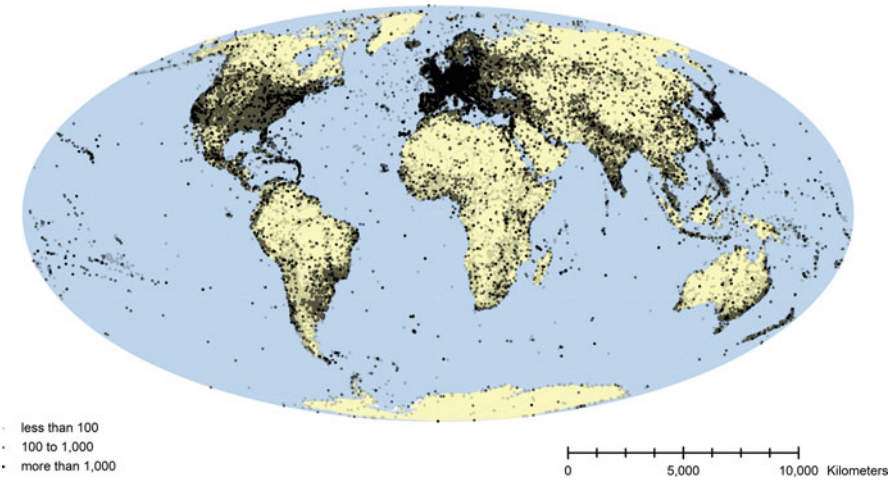

Fig. 11.2 Spatial distribution of geotagged Wikipedia articles, visualized using log-scale density for number of article contributions at 10 km resolution

VGI production processes themselves exhibit regular spatial patterns. This section will discuss a spatial model for contributions, and results that show anonymous contributors exhibit geographic effects that fit an exponential distance decay function.

11.4.1 Data Collection

Wikipedia manages hundreds of individual language-specific databases across three data centers in the United States, Netherlands, and South Korea. Their services use open-source MediaWiki software and data models (MediaWiki 2006). Wikipedia provides article and metadata via periodic dumps of their database and as static HTML files (http://meta.wikimedia.org/wiki/Data_dumps), but historically, these data do not always include complete article contribution records due to their volume and limited operational resources (e.g., the August 2008 dump of the English Wikipedia had 2.5 million articles and 250 million contributions—http://en.wikipedia.org/wiki/Special:Statistics).

The openness of their data lends itself to empirical study by researchers (e.g., Almeida et al. 2007; Priedhorsky et al. 2007; Voss 2005). This study collects data directly via SQL from near real-time replicas of Wikipedia databases, provided by Wikimedia Deutschland's *Toolserver* (http://toolserver.org). These databases use MySQL and the MediaWiki database schema, which organizes articles by revision. Briefly, 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

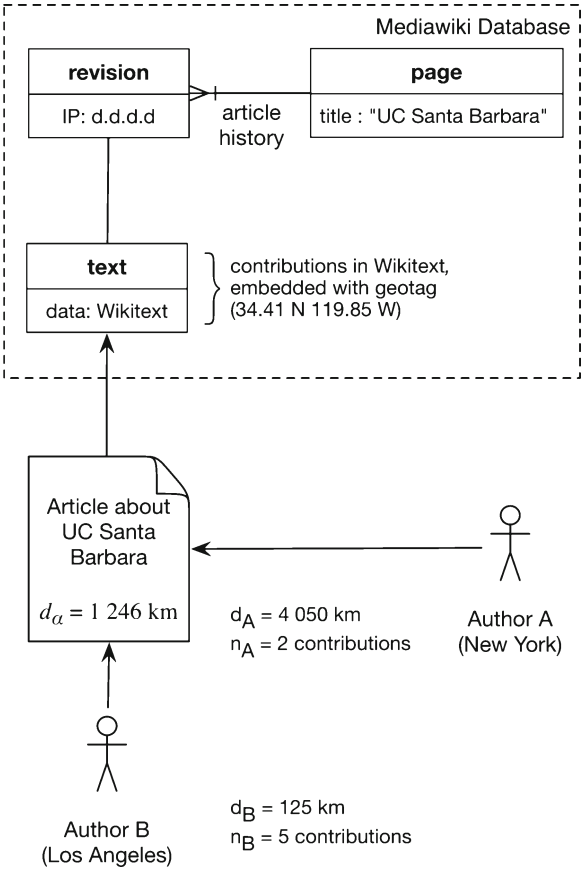

Fig. 11.3 VGI production process in Wikipedia. Authors contribute to place-based articles using Wikitext and embedded geotags that are stored in database tables, including a full history of revisions. For anonymous authors, each revision includes their IP address. In the example, two authors contribute to an article about UC Santa Barbara whose signature distance d_α is 1,246 km, defined as the average distance weighted by contributions, for example, $(2{\cdot}4050 + 5{\cdot}125) / (2 + 5) = 1246$

page table contains a unique identifier and the language-specific title for the article, and the *text* table stores the article's contents. Wikipedians write articles using *Wikitext*, a loosely structured markup language (http://en.wikipedia.org/wiki/Wikipedia:MARKUP), and they embed semistructured metadata *within* the article (Fig. 11.3). The nondeterministic nature of Wikitext's grammar and conventions causes problems for structured data extraction (cf. Sauer et al. 2007). The *WP:GEO* project in Wikipedia governs an infrastructure for adding geographic information to articles (http://en.wikipedia.org/wiki/Wikipedia:GEO). They provide an array of "wiki templates" that have a semistructured syntax for embedding geographic coordinates.

Wikipedia-World's database (Kühn and Alder 2008) from 10 May 2008 uses an extensive data-mining process to extract geotags embedded in Wikitext articles (Fig. 11.1).[4] For each geotagged article, we extract all the authoring history and the most recent version from the replica databases.

To simplify computation across language-specific databases, we migrate the authoring histories into a single shared database, where we modify MediaWiki tables to associate a source language for each record (e.g., *page_id* and a new *page_lang* column comprise the primary key instead of only *page_id*) and to remove data incidental to analysis. This data model provides a multilingual abstraction layer to Wikipedia articles, authors, and their contributions. It has tables for *article*, *author*, and *geotag* data, and *author_article* and *geotag_article* association tuples. It also provides fast access to summary statistics per article and per author. The data extraction from the MediaWiki tables results in *page* and *text* with 990,315 articles, *revision* with 32,141,334 author contributions between 2001 and 2008, and *user* with 578,448 registered author accounts. Since the *user* table contains records only for registered authors, the analysis extracts and parses data from the *revision.rev_user_text* column to identify IP addresses for anonymous users and to integrate them into the data model.

11.4.2 Spatial Model of Authorship

Each author in Wikipedia has a "spatial footprint" comprised of all of the articles to which they have contributed. For anonymous authors, we can estimate their location using IP geolocation (Fig. 11.4). For registered authors and bots (Figs. 11.5 and 11.6), we have no direct estimate of their location, although an indirect estimate based on their spatial footprint is possible (Lieberman and Lin 2009). But are there spatial patterns in these interactions between the authors and the places about which they write?

11.4.2.1 Gravity Models

In regional geography and related disciplines, spatial interaction models form the basis of social theories (Haynes and Fotheringham 1984). These models pertain to flows (interactions) between two or more geographic regions. They have a decades-long history in geography dating back to "social physics" in the early twentieth century (Fotheringham 1981; Wilson 1969, 1971). Distance decay or "gravity" models are one

[4] Their software targets a predetermined set of 21 languages: Catalan (ca), Chinese (zh), Czech (cs), Danish (da), Dutch (nl), English (en), Esperanto (eo), Finnish (fi), French (fr), German (de), Icelandic (is), Italian (it), Japanese (ja), Norwegian (no), Polish (pl), Portuguese (pt), Russian (ru), Slovak (sk), Spanish (es), Swedish (sv), and Turkish (tr).

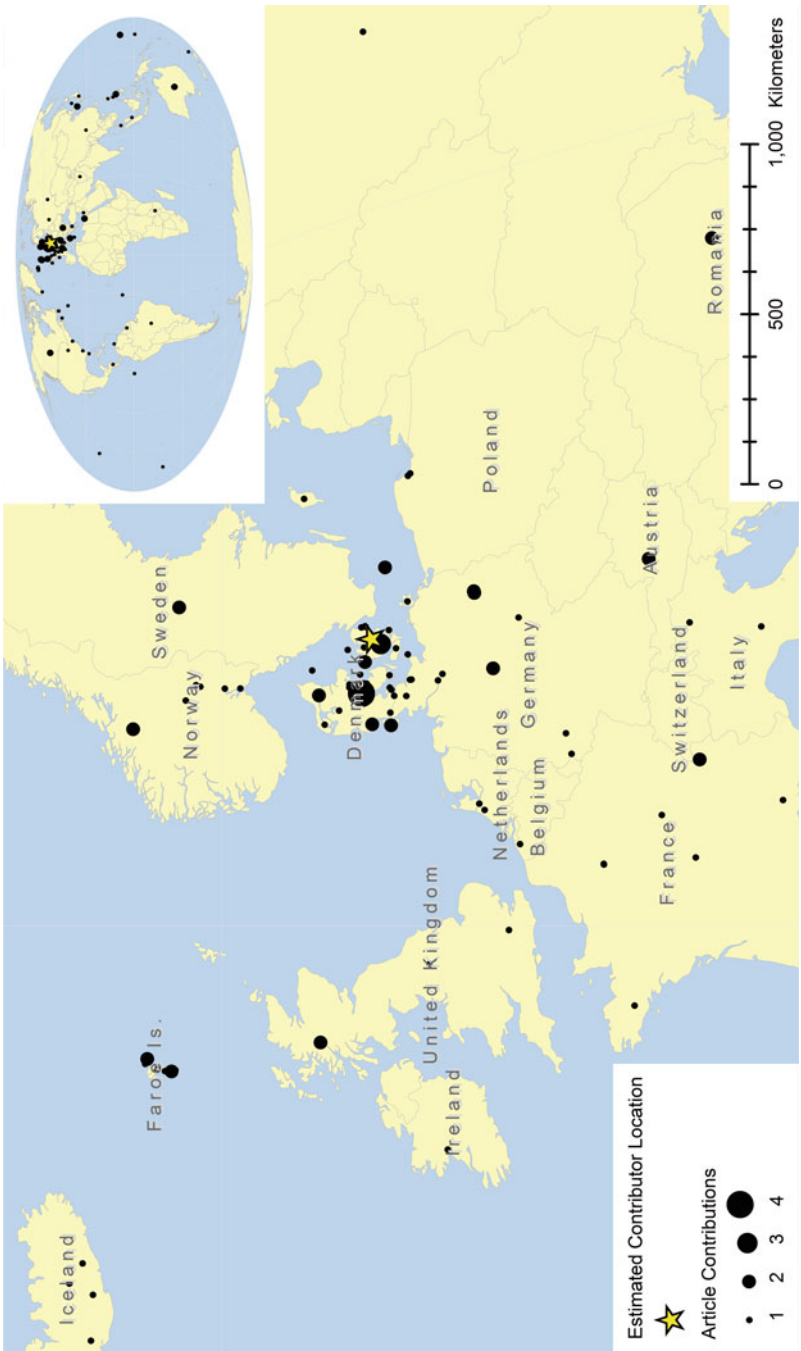

Fig. 11.4 Spatial footprint of an *anonymous* author with 172 contributions to 143 articles in the Danish Wikipedia. The *yellow icon* represents an estimate of the author's location, based in IP geolocation

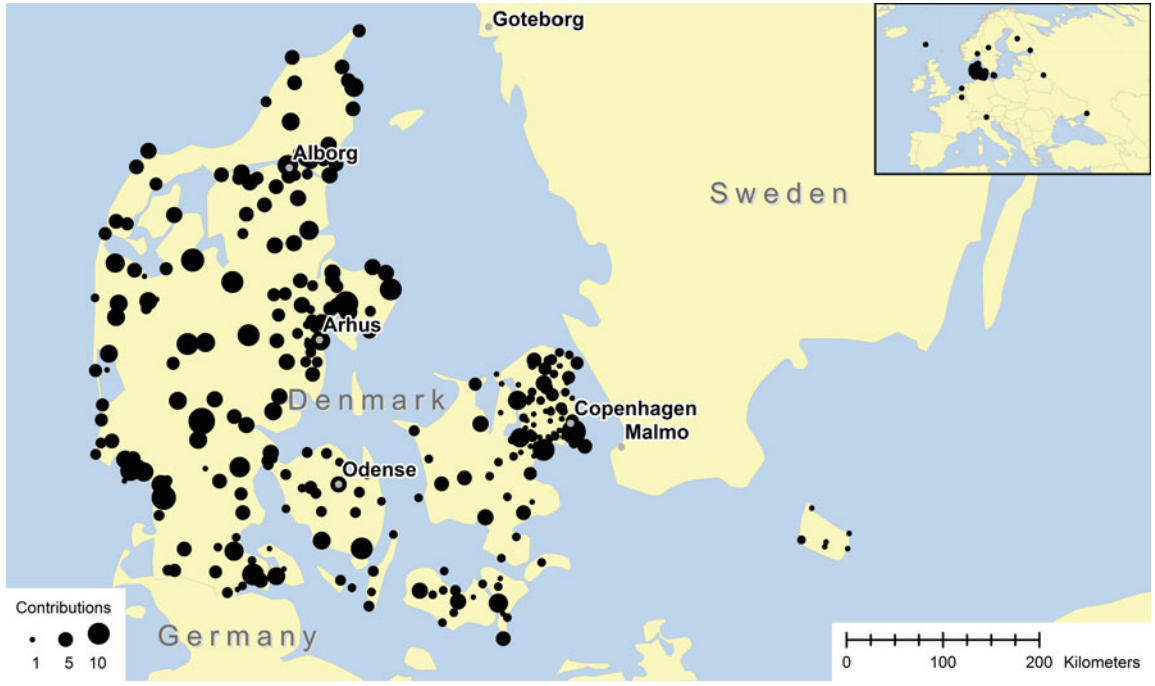

Fig. 11.5 Spatial footprint of a *registered* author with 1,099 contributions to 296 articles in the Danish Wikipedia. *Markers* represent geotagged location of each article edited by author, the vast majority of which are clustered inside Denmark

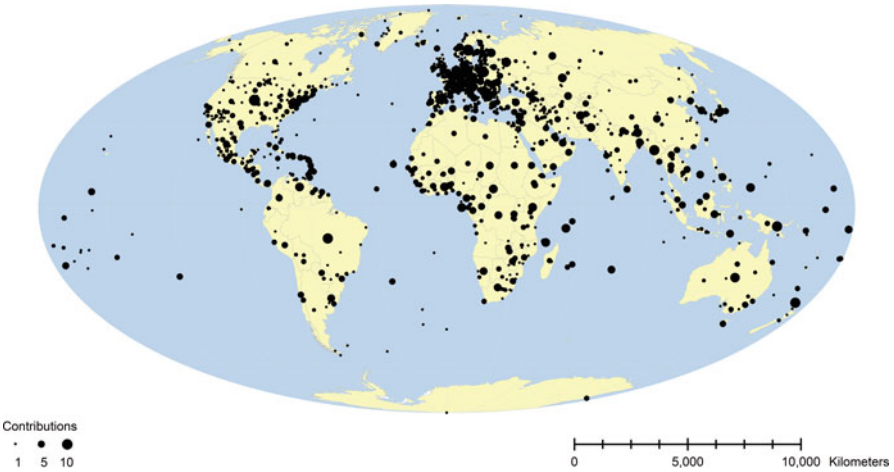

Fig. 11.6 Spatial footprint of a *bot* with 3,006 contributions to 1,601 articles in the Danish Wikipedia

type of spatial interaction model. They use "mass" functions to deal with scale and distance effects. The general gravity model (Sen and Smith 1995, p. 3) is

$$T_{ij} = A_i \bullet B_j \bullet F(d_{ij}), \tag{11.1}$$

where T_{ij} is the interaction between population centers i and j; A_i and B_j are unspecified origin and destination weight (mass) functions; d_{ij} is the spatial factor, or distance between regions i and j; and $F(d_{ij})$ is an unspecified distance decay function, which is commonly a power, exponential, or gamma (combined) function (Sen and Smith 1995, pp. 93–99).

In spatial information theory, an individual's *information field* is the spatial distribution of the "knowledge an individual has of the world" (Morrill and Pitts 1967, p. 406) and is a factor when modeling sociospatial behaviors, like diffusion of innovation or migration (Hägerstrand 1967). An individual's information field decays as the distance from the individual increases. In quantitative geography, gravity models formalize spatial interaction analysis by using this type of distance decay function (Fotheringham and O'Kelly 1989; Sen and Smith 1995). When Wikipedians choose to write about a place, their mean information fields should exhibit distance decay effects found in other sociospatial phenomena, like innovation diffusion. When Wikipedians as a group write more articles, for example, they expand the overall spatial coverage of Wikipedia articles. But when an individual Wikipedian writes an article about a place, that place is likely to be nearby. Thus, our hypotheses for this study are (a) Wikipedians write articles about nearby places more often than distant ones and (b) this likelihood follows an exponential distance decay function.

Fig. 11.7 The UCSB article in English has a signature distance of 533 km based on 135 anonymous authors with 719 revisions. Each contribution is shown as a *white line*, with *thicker lines* denoting more contributions

11.4.2.2 Gravity Model for VGI Production

To model VGI production as a spatial process, we define a probabilistic model where the dependent variable is a likelihood for interaction, based on a spatial factor. Specifically, we use a *probabilistic invariant exponential gravity model* (Sen and Smith 1995, p. 102). In terms of Eq. 11.1, T_{ij} is converted to the probability of an interaction based on a spatial factor. The mass terms A_i and B_j are combined into a single invariant constant K to allow for uneven distributions of authors and articles over the Earth's surface. Finally, $F(d_{ij}) = exp(-\beta d_{ij})$, an exponential distance decay function.

$$\Pr(d = d_\alpha) = K \cdot exp(-\beta d_\alpha), \quad \text{where} \quad d = d' \pm \varepsilon. \tag{11.2}$$

Equation 11.2 shows the model using the probability $\Pr(d = d_\alpha)$ as the likelihood that a given article has a signature distance d_α equal to a distance d within a range of $d' \pm \varepsilon$ (K and β are empirically derived constants). For this spatial model, we use a "signature distance" d_α metric to measure the proximity effect for a given article (Hardy et al. 2012). The metric is the average distance between an article and its n authors, weighted by relative number of contributions from each author (Figs. 11.3 and 11.7). That is, each anonymous author has a spatial footprint that is the set of contributions

made to any geotagged article by that author. Every author has a single footprint, and every article belongs to its authors' footprints. This model requires a known location for both articles and authors, so we use MaxMind's GeoLite City database, which uses proprietary methods to convert IP addresses into geographic coordinates, to estimate the locations of anonymous Wikipedians whose IP addresses are embedded into their contributions.[5] Location-based services have driven the development of methods to convert IP addresses into geographic coordinates (Muir and Oorschot 2009; Stanger 2008) and to evaluate positional accuracy (Gueye et al. 2006; 2007).

11.4.2.3 Model Results by Article

To fit the model in Eq. 11.2 to the study data, we use an ordinary least squares regression method with a logarithmic transformation to a linear model:

$$\ln\left[\Pr\left(d = d_{\alpha}\right)\right] = \ln K - \beta d_{\alpha}. \tag{11.3}$$

All geographic calculations use ~10-km resolution and great circle distances (where $1' = 1.852$ km). We selected the sample from available data to satisfy the methodological requirements that articles have at least one anonymous contribution (for author location estimates) and that articles have one and only one geotag (for signature distance metric). We convert the units of d_{α} from km to 10^3 km, and use observed relative frequency for $Pr(d = d_{\alpha})$. The model fits at $K = 0.0022$ and $\beta = 0.2842$ ($n = 438{,}077$; $R^2 = 0.9005$; $p < 0.01$; $f = 17{,}480$; $DF = 1{,}930$). When signature distances are relatively low ($d_{\alpha} < 2$), there is no correlation across language databases, suggesting spatial behavior is idiosyncratic across languages.

11.4.2.4 Model Results by Article Category

To test whether signature distances vary by category, we collected categorical data for English articles. Contributors may categorize Wikipedia articles into one or more categories. These categories are not strictly tags but rather registered categories, although anyone may create a new category. These categories are often descriptive of a topic such as "14th-century architecture" or "Art museums and galleries in Paris." They may be editorial, however, and denote workflow items such as "Tokyo railway station stubs," or "All articles needing style editing," or "Articles lacking sources from December 2009." The category space is flat with no consistent nomenclature. Each article's category is displayed at the bottom of the article, and each category has an "article" that lists all articles belonging to that category. From our

[5] Wikipedia provides access to IP addresses for anonymous, but not registered, Wikipedians. Reportedly, Wikipedia logs IP addresses for all contributions—from anonymous and registered Wikipedians alike—but they restrict access to those data to authorized administrators.

Table 11.2 Popular topic keywords in English articles, sorted by distance

1,000 km	2,000 km	3,000 km	4,000 km	5,000 km	6,000 km
Carolina	Area	Airports	Areas	Paris	Islands
Channel	California	Architecture	Articles		
County	Census-designated	Building	Australia		
Illinois	Communities	Buildings	Communes		
Indiana	England	Cities	Containing		
Metropolitan	Established	Cleanup	Districts		
Michigan	Establishments	District	Former		
Micropolitan	London	Lacking	Geography		
Missouri	Museums	Municipalities	Language		
Ohio	New	National	Mountains		
Pennsylvania	North	Needing	Prefecture		
Television	Opened	Places	Province		
Texas	Railway	Populated	Region		
TV	States	References	Sites		
Washington	Stations	State	South		
York	United	Structure	Statements		
	Venues	Structures	Stubs		
	Villages	Towns	Text		
		West	Wikipedia		

study, we collected 8,474 unique categories with at least ten English articles, comprising 372,793 articles.[6] We then extracted 4,512 unique keywords (minus common words) from the category title to create an inverted index of category keywords. Each index entry has a unique category keyword, the number of articles that belong to the category, and a mean signature distance for those articles.

For topic keywords with at least 50 articles, Table 11.2 shows the popular topic keywords in English articles by the mean signature distance d_α (n=372,793; mean = 3,049 km). While not conclusive, there is some evidence that signature distances do vary by topic. Topic keywords with lower mean distances are "local" in scope such as cities ("[New] York"), state names, administrative boundary terms ("County" or "Metropolitan"), and buildings ("Museums"). Those with higher mean distances were "regional" in scope such as non-English speaking cities ("Paris"), country names ("Australia"), and regional boundaries ("Islands" or "Province" or "Region").

11.5 Discussion

This section presents some further research issues on architectural, social, and methodological factors, beginning with how both geotagging and geolocation could better support VGI production processes.

[6] Other languages also have categories, but this content analysis is restricted to English.

11.5.1 Architectural Factors

The lack of well-structured geotags is problematic. In particular, further research on methods for specifying geotags as *first-class metadata*—rather than as the most basic common denominator of latitude, longitude coordinates—is needed. If collaborative online gazetteers with large-scale global coverage were to emerge, they might serve as a basis for toponym-indexed geotags and thus relieve users from low-level georeferencing tasks. In the meantime, collaborative methods are a possible approach to improving geotag metadata, especially within scientific communities. Currently, geotagging schemes are opaque and inconsistent and are done by automated bots or by users who specify geographic coordinates interactively from a general-purpose mapping service. Neither of these schemes preserve semantic or context information about place and instead leave only precise numerical coordinates of ambiguous intent.

For decades, metadata has been the ever-present, cure-all solution to heterogeneous data integration and use. Yet high-quality, ubiquitous metadata is extremely rare in practice, despite geospatial data infrastructures that are designed to be interoperable and metadata-centric (de By et al. 2009; van Loenen et al. 2009). Current VGI systems may provide insights on how users could produce and manage better metadata for geotags. Metadata is "data about data," intended to facilitate data discovery, integration, and use (or reuse). Practitioners often standardize metadata syntax and semantics, but adherence to metadata standards is extremely rare in distributed systems, especially large or global ones; this is hereafter referred to as the "metadata problem." GIS usually assumes strongly typed spatial data representations, and GIScientists have developed disambiguation methods (e.g., toponym resolution or fuzzy boundaries) for spatial data that do not comply with these structures. These complexities make metadata important for geospatial integration and use. VGI systems, however, successfully integrate heterogeneous data sources on a global scale without solving the metadata problem directly. VGI systems use "best effort" geotagging methods and representations to avoid the complexity of richer GIScience approaches to georeferencing. Moreover, the VGI notion of metadata, and its production and management, is different than in geospatial data infrastructures.

Scientific communities have collaborated on metadata standards and conventions, such as CF (Hankin et al. 2009) and its predecessor (COARDS 1995), but in a study of earth science datasets published via the OPeNDAP protocol (Hardy et al. 2006), they do not accurately follow these conventions. In fact, only a minority of them *claims* their convention (as required), and even of those, only a fraction *accurately* adhere to their stated convention. In practice, scientific data sharing varies by discipline. Ecologists, for example, take idiosyncratic approaches to data sharing and reuse, which depend on disciplinary knowledge and social factors (Zimmerman 2007). This metadata problem forces scientists to use specialized knowledge and manual effort for data reuse.

Wikipedia may provide some lessons for metadata production and management in geospatial data infrastructures (Table 11.3). GIScientists may consider the wiki approach to metadata production and use to address how they might integrate the

Table 11.3 Applying Wikipedia approaches to geotagging

Approach	*Benefit*
Use trivial geotagging	They avoid the complexity (and implications) of GIScience approaches to georeferencing. Typically, they use decimal degree geographic coordinates in an assumed datum (WGS84) and without enforcing numerical precision. For example, in Wikipedia, the geotag for UCSB is (−119.84813°, 34.41254°), where 0.00001° ≈ 1 m precision[a]. Flickr uses a similar approach but saves context for how users select location (e.g., a zoom level on an interactive map) (Jankowski et al. 2010)
Seemingly trivial metadata structures	Wikipedia uses *Wikitext* (Sauer et al. 2007), a lightweight markup language, for its article content and metadata. Flickr uses perhaps the simplest metadata structure of all: tags which are simply any word or phrase in an uncontrolled vocabulary
Use bots extensively	Wikipedia has hundreds of semiautomated programs to perform a wide variety of editorial functions from removing vandalism to extracting metadata to suggesting work
Promote "refined" and flag problematic content	The community identifies content that is exemplary or meets certain quality standards, and promotes this content. They flag any content that needs further work which is then suggested to those looking for content to work on. They also flag content that is subject to controversy or "edit wars" (Viégas et al. 2007a)
Lazy, but rapid integration (most popular first).	Mashups and other rapid prototyping use service-level, rather than data-level, integration. They focus on APIs and "cookbooks" or working examples rather than formalized specifications. This approach enables rapid integration but also lazy integration since they target specific uses with partial APIs
Complete histories of revisions	The complete context for changes is always available (and easily accessible) when issues arise or for tools to utilize
Data mining	They have tools that search through content looking for ways to improve their service, and they provide APIs for anyone who wishes to mine content programmatically

[a]Typical GPS units report coordinates at ≈1 m resolution in WGS84 datum

increasingly voluminous VGI data into metadata-based geospatial data infrastructures. In particular, the novelty and practicalities of VGI production may benefit the scientific community as they confront increasingly large-scale, heterogeneous data integration problems in metadata-poor environments—a recurrent research area (Hardy 2010; Hardy et al. 2006; Lanter 1991; Rodriguez et al. 2009).

Ideally for analysis, all contributions would have explicit geographic information for the author's location. But these data are not available in most VGI

applications, including Wikipedia. Thus, geolocation methods are problematic for VGI contributions due to constraints in data availability and also privacy concerns. This study exploits IP addresses to apply geolocation methods for anonymous contributors. IP geolocation methods, however, are inherently both spatially and temporally dynamic in nature, inaccurate at large scales (i.e., street-level), and relatively easily evaded by savvy users or anonymizing software (Duckham and Kulik 2005; Muir and Oorschot 2009).

Alternatives are similarly constrained. Current survey-based methodologies are limited (e.g., Nov (2007) used email solicitations which yielded about 150 authors) due to the level of anonymity in Wikipedia. Spatial analysis methods based on behavioral patterns, such as the locations of the articles to which an author has contributed, are relatively new in this research area (Lieberman and Lin 2009). Combined approaches (i.e., where quantitative spatial analysis models are calibrated with surveyed locations) may prove useful. Furthermore, VGI is increasingly moving into the mobile domain where users leave (often implicitly) digital traces more conducive to geolocation methods, such as GPS-enabled smart phones, cell phone tower records, or even georeferenced photos (Girardin et al. 2008; González et al. 2008). These trace data can enable spatial data-mining methods for tracking trajectories of individuals or groups (Kisilevich et al. 2010).

Interdisciplinary approaches may also prove useful since geolocation methods are used in other domains. Geographic profiling, for example, is a criminal "investigative methodology that uses the locations of a connected series of crimes to determine the most probable area of offender residence" (Rossmo 2000, p. 1). Geographic profiling systems use spatial distribution and probability distance strategies, such as center of the circle, centroid, median, geometric mean, harmonic mean, and center of minimum distance algorithms (Snook et al. 2005).

11.5.2 Social Factors

How do social factors (such as communication, culture, language, settlement patterns (diaspora), and socioeconomic status) influence VGI contributions? The production and use of VGI will likely shift spatial data infrastructures architecturally to provide for social factors (Budhathoki et al. 2008; Coleman et al. 2009; Elwood 2010; Elwood et al. 2012; Sieber and Rahemtulla 2010). Further modeling of social characteristics in the collaborative authorship process might include spatiotemporal constraints on social networks of Wikipedians or future VGI systems based on increasingly rich social network technologies.

For example, the VGI production model defines work in the signature distance metric in simple terms as an edit count. But the literature has many different definitions for "work," including edit counts (Kittur et al. 2007), edit deltas (Zeng et al. 2006), edit similarity (i.e., information distance) (Voss 2005), edit longevity (i.e., age or survival or persistence) (Adler and de Alfaro 2007; Wöhner and Peters 2009), and edit visibility (Priedhorsky et al. 2007). These

metrics may better model social processes and clarify sociospatial factors in collaborative authorship. In particular, edit longevity and edit visibility more directly reflect social phenomena like "edit wars"[7] and herding behaviors, respectively. Similarly, our study had limited comparison of geographic effects across article categories, but further analysis on content-centric dimensions may help study these social processes.

Another question is whether language and population demographics affect spatial patterns in VGI. Ideally, spatial models for collective authorship would include probabilities for how many potential Internet users who speak a given language are available to make contributions for any given location. This study did not normalize authorship by population or potential speakers due to a lack of available data at the needed resolution. Balk and Yetman (2004) provide relatively large-scale data for population estimations but do not include speaker estimates. Moreover, Internet use is spatially variant (Billón et al. 2008; Zook 2005) where large-scale Internet population estimates are not readily available.

Furthermore, at a global scale of the Internet, the concept of "near" is different than in social science research that conducts studies at smaller scales (Graham 1998). For example, in our study, less than 2,000 km is relatively "near" compared to the full scope of available Wikipedia contributors. Notwithstanding global or even virtual travel (Urry 2002), typical scales for nearness are much smaller than 2,000 km, such as walking in urban centers (Turner and Penn 2002) or commuting distance via transportation networks (Weber 2003).

Finally, the notion of collective action through new media is at the core of VGI. VGI and the related phenomena of *neogeography* expand the notion of the "public" from prior work in public participation GIS to include much larger, distributed civic participation (Elwood 2008; Hall et al. 2010; Sieber 2006; Sui 2008).

11.5.3 Methodological Factors

Finally, what methodological advancements are required for future research? The high-volume, highly distributed, real-time, and social nature of VGI is inherently difficult to analyze with simple computational methods. Rather, as shown in our research, significant computational resources and data-mining methods are better suited for empirical studies of VGI. Data-mining methods with a resolution at sub-article levels, such as sections or paragraphs, would improve the sample size. Also, geographic and network visualization methods may enable a visual analytics approach to studying VGI.

[7] To address these actions, Wikipedia has a policy that states "Wikipedians should interact in a respectful and civil manner" (http://en.wikipedia.org/wiki/Wikipedia:Five_pillars).

In the coming years, wiki-based VGI systems, where the provenance of information is transparent, may no longer apply as the ephemeral and social nature of VGI rises. Specifically, one of the key challenges in methodology will be to effectively cope with data deluge in an environment where data are filtered through social networks (Watts et al. 2002). If information primitives become based on distance or connectivity through fluctuating social networks, then traditional information science methodologies will not be applicable at large scales. Social network methodologies, which are based on graph theory, are now being used to study online collaborative environments, such as massively multiplayer gaming (Szell and Thurner 2010), and blogging (Liben-Nowell et al. 2005).

11.6 Conclusion

Although the underlying technologies of online geographic services have been in development for many years, the behavioral impacts of VGI production are largely unknown. These services require large-scale data interoperability and collaboration, for example, neither of which has a purely technical solution. VGI production will likely create new knowledge politics, and many of the problematic emerging issues are institutional and sociobehavioral in nature, not technological (Elwood 2008, 2010; Goodchild 2008). For example, the capacity of a ubiquitous Internet to reduce communication costs has raised questions of whether geographic distance matters in information and economic production (Cairncross 1997; Castells 2010).

This chapter addresses two basic questions in VGI production, namely, (1) how individuals contribute place-based information to a digital commons and (2) authorship dynamics of such collective effort. Our approach takes a user-centric perspective of spatial behavior in VGI production. Research on VGI production is a nascent area with many unexplored avenues, in architectural, social, and methodological factors. These factors form a basis of a research agenda that asks (a) how to improve the structure and quality of essential geographic metadata, (b) how language and demographics affect VGI, and (c) how social networks change the nature of VGI.

Acknowledgments This research was supported in part by the National Science Foundation (awards #BCS-0849625 “Collaborative Research: A GIScience Approach for Assessing the Quality, Potential Applications, and Impact of Volunteered Geographic Information” and #IIS-0431166 “Collaborative Research: Integrating Digital Libraries and Earth Science Data Systems”) and the US Army Research Office (award #W911NF0910302). Thanks to Wikimedia Deutschland, e.V. in Berlin, Germany, for providing the helpful Toolserver service (http://toolserver.org). They provided database access, Web hosting, and computational resources for this research. Thanks to Tim Alder and Stefan Kühn for comments on geotagging methods in Wikipedia and for sharing their data-mining software and results. Thanks also to reviewer comments and for the many discussions with students and faculty at UCSB’s Center for Information Technology and Society and Center for Spatial Studies.

References

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. *WWW'07*. doi:10.1145/1242572.1242608.

Alexa Internet, Inc. (2009). Alexa traffic rank. http://www.alexa.com/siteinfo/wikipedia.org. Accessed Dec 2009.

Almeida, R., Mozafari, B., & Cho, J. (2007, March 26–28). *On the evolution of Wikipedia.* Paper presented at the 1st international conference on weblogs and social media, Boulder, CO.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. *SIGIR'04*. doi:10.1145/1008992.1009040.

Balk, D., & Yetman, G. (2004). Gridded population of the world (GPWv3). http://sedac.ciesin.columbia.edu/gpw/. Accessed Feb 2010.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. et al. (2004). Using social psychology to motivate contributions to online communities. *CSCW'04*. doi:10.1145/1031607.1031642.

Benkler, Y. (2002). Coase's penguin, or, Linux and the nature of the firm. *The Yale Law Journal, 112*(3), 369–446.

Billón, M., Ezcurra, R., & Lera-López, F. (2008). The spatial distribution of the internet in the European Union: Does geographical proximity matter? *European Planning Studies, 16*(1), 119–142.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies, 14*(1), 75–90.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the web. *Computer Networks, 33*(1–6), 309–320.

Browne, M. (1997). The field of information policy: Fundamental concepts. *Journal of Information Science, 23*(4), 261–275.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. *GROUP'05*. doi:10.1145/1099203.1099205.

Budhathoki, N. R., Bruce, B., & Nedovic-Budic, Z. (2008). Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal, 72*(3), 149–160.

Cairncross, F. (1997). *The death of distance: How the communications revolution will change our lives*. Cambridge, MA: Harvard Business School Press.

Castells, M. (2010). *The rise of the network society* (2nd ed.). West Sussex: Wiley-Blackwell.

Çelik, T. (2005). Geo microformat specification [draft]. http://microformats.org/wiki/geo. Accessed Dec 2009.

COARDS (1995). Conventions for the standardization of NetCDF files. http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html. Accessed July 2009.

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research, 4*, 332–358.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558–569.

Daviel, A., & Kaegi, F. (2007). Geographic registration of HTML documents [draft]. http://tools.ietf.org/pdf/draft-daviel-html-geo-tag-08.pdf. Accessed Dec 2009.

de By, R., Lemmens, R., & Morales, J. (2009). A skeleton design theory for spatial data infrastructure. *Earth Science Informatics, 2*(4), 299–313.

Duckham, M., & Kulik, L. (2005). A formal model of obfuscation and negotiation for location privacy. In H. W. Gellersen et al. (Eds.), *Pervasive 2005* (pp. 152–170, LNCS, Vol. 3468). Berlin: Springer.

Elwood, S. (2008). Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal, 72*(3), 173–183.

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., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers, 102*(3), 571–590. doi:10.1080/00045608.2011.595657.

Forte, A., & Bruckman, A. (2008). Scaling consensus: Increasing decentralization in Wikipedia governance. *HICSS'08*. doi:10.1109/HICSS.2008.383.

Fotheringham, A. S. (1981). Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers, 71*(3), 425–436.

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial interaction models: Formulations and applications*. Dordrecht: Kluwer Academic.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. New York: Farrar, Straus, and Giroux.

Girardin, F., Calabrese, F., Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing, 7*(4), 36–43.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198–208.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779–782.

Goodchild, M. F. (2000). Communicating geographic information in a digital age. *Annals of the Association of American Geographers, 90*(2), 344–355.

Goodchild, M. F. (2004). Scales of cybergeography. In E. Sheppard & R. B. McMaster (Eds.), *Scale and geographic inquiry: Nature, society, and method* (pp. 154–169). Malden: Blackwell.

Goodchild, M. F. (2008). Geographic information science: The grand challenges. In J. P. Wilson & A. S. Fotheringham (Eds.), *The handbook of geographic information science* (pp. 596–608). Malden: Blackwell.

Graham, S. (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography, 22*(2), 165–185.

Griffith, V. (2007). WikiScanner. http://wikiscanner.virgil.gr/. Accessed February 2009.

Gueye, B., Ziviani, A., Crovella, M., & Fdida, S. (2006). Constraint-based geolocation of internet hosts. *IEEE/ACM Transactions on Networking, 14*(6), 1219–1232.

Gueye, B., Uhlig, S., & Fdida, S. (2007). Investigating the imprecision of IP block-based geolocation. In S. Uhlig, K. Papagiannaki, & O. Bonaventure (Eds.), *Passive and active network measurement* (pp. 237–240, LNCS, Vol. 4427). Berlin: Springer.

Hägerstrand, T. (1967). *Innovation diffusion as a spatial process*. Chicago: University of Chicago Press.

Hall, G. B., Chipeniuk, R., Feick, R. D., Leahy, M. G., & Deparday, V. (2010). Community-based production of geographic information using open source software and Web 2.0. *International Journal of Geographical Information Science, 24*(5), 761–781.

Hankin, S. C., & 14 co-authors (2009). NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes. *OceanObs'09: Sustained ocean observations and information for society*. doi:10.5270/OceanObs09.cwp.41.

Hardy, D. (2008, October 15–19). *Discovering behavioral patterns in collective authorship of place-based information*. Paper presented at the 9th international conference of the association of internet researchers, Copenhagen, Denmark.

Hardy, D. (2010, September 14). *"Title not required": The wikification of geospatial metadata*. Paper presented at the GIScience workshop on the role of volunteer geographic information in advancing science, Zurich, Switzerland.

Hardy, D., Janée, G., Gallagher, J., Frew, J., & Cornillon, P. (2006). Metadata in the wild: An empirical survey of OPeNDAP-accessible metadata and its implications for discovery. *Eos Trans. AGU, 87*(52), Fall Meet. Suppl., Abstract IN54A-04.

Hardy, D., Frew, J., & Goodchild, M. F. (2012). Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*. doi:10.1080/13658816.2011.629618.

Haynes, K. E., & Fotheringham, A. S. (1984). *Gravity and spatial interaction models*. Beverly Hills: Sage.

Hecht, B. J., & Gergle, D. (2010). On the "localness" of user-generated content. *CSCW'10*. doi:10.1145/1718918.1718962.

Hecht, B., & Moxley, E. (2009). Terabytes of Tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge. In K. S. Hornsby (Ed.), *Spatial information theory* (pp. 88–105, LNCS, Vol. 5756). Berlin: Springer.

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. Cambridge, MA: MIT Press.

Huberman, B. A., Pirolli, P. L., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science, 280*, 95–97.

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

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2010). Spatio-temporal clustering. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed., pp. 855–874). New York: Springer.

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in Wikipedia. *CHI'07*. doi:10.1145/1240624.1240698.

Kühn, S. (2008). Workflow from Wikipedia-Dump to geodata. http://de.wikipedia.org/wiki/Datei:Wikipedia_Geodata_Workflow.svg. Accessed Oct 2008. Creative Commons license (CC BY-SA 3.0).

Kühn, S., & Alder, T. (2008). Wikipedia-World [in German]. http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World. Accessed Oct 2008.

Kunze, J. (1999). Encoding Dublin core metadata in HTML. http://www.ietf.org/rfc/rfc2731.txt. Accessed Mar 2008.

Lanter, D. P. (1991). Design of a lineage-based meta-data base for GIS. *Cartography and Geographic Information Science, 18*(4), 255–261.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences, 102*(33), 11623–11628.

Lieberman, M., & Lin, J. (2009, May 17–20). *You are where you edit: Locating Wikipedia contributors through edit histories*. Paper presented at the 3rd International AAAI Conference on Weblogs and Social Media, San Jose, CA.

Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. New York: Hyperion.

Marston, S. A., Jones, J. P., & Woodward, K. (2005). Human geography without scale. *Transactions of the Institute of British Geographers, 30*(4), 416–432.

MediaWiki (2006). The technical manual for the MediaWiki software: Database layout. http://www.mediawiki.org/wiki/Manual:Database_layout. Accessed Mar 2008.

Miller, N. (2005). Wikipedia and the disappearing "Author". *ETC: A Review of General Semantics, 62*(1), 37–41.

Morrill, R. L., & Pitts, F. R. (1967). Marriage, migration, and the mean information field: A study in uniqueness and generality. *Annals of the Association of American Geographers, 57*(2), 401–422.

Muir, J. A., & Oorschot, P. C. V. (2009). Internet geolocation: Evasion and counterevasion. *ACM Computing Surveys, 42*(1), 1–23.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41–46.

Nov, O. (2007). What motivates Wikipedians? *Communications of the ACM, 50*(11), 60–64.

O'Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. http://oreilly.com/web2/archive/what-is-web-20.html. Accessed Mar 2008.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. *GROUP'07*. doi:10.1145/1316624.1316663.

Rodriguez, M. A., Bollen, J., & Sompel, H. V. D. (2009). Automatic metadata generation using associative networks. *Transactions on Information Systems, 27*(2), 1–20.

Rossmo, D. K. (2000). *Geographic profiling*. Boca Raton: CRC Press.

Sauer, C., Smith, C., & Benz, T. (2007). WikiCreole: A common wiki markup. *International Symposium on Wikis*. doi:10.1145/1296951.1296966.

Sen, A., & Smith, T. E. (1995). *Gravity models of spatial interaction behavior*. Berlin: Springer.
Sieber, R. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers, 96*(3), 491–507.
Sieber, R. E., & Rahemtulla, H. (2010). *Model of public participation on the geoweb.* Paper presented at the 6th international conference on GIScience, Zurich, Switzerland, September 14–17, 2010.
Snook, B., Zito, M., Bennell, C., & Taylor, P. J. (2005). On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology, 21*(1), 1–26.
Stanger, N. (2008). Scalability of techniques for online geographic visualization of web site hits. In A. Moore & I. Drecki (Eds.), *Geospatial vision: New dimensions in cartography* (pp. 193–217). Berlin: Springer.
Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS [editorial]. *Computers, Environment and Urban Systems, 32*(1), 1–5.
Sundin, O., & Haider, J. (2007). Debating information control in Web 2.0: The case of Wikipedia vs. citizendium. *Proceedings of the American Society for Information Science and Technology, 44*(1), 1–7.
Szell, M., & Thurner, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social Networks, 32*(4), 313–329.
Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything*. New York: Portfolio.
Turner, A., & Penn, A. (2002). Encoding natural movement as an agent-based system: An investigation into human pedestrian behaviour in the built environment. *Environment and Planning B, 29*(4), 473–490.
Urry, J. (2002). Mobility and proximity. *Sociology, 36*(2), 255–274.
van Loenen, B., Besemer, J. W. J., & Zevenbergen, J. A. (Eds.). (2009). *SDI convergence: Research, emerging trends, and critical assessment*. Delft: Netherlands Geodetic Commission.
Viégas, F. B. (2005). *Revealing individual and collective pasts: Visualizations of online social archives*. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *CHI'04*. doi:10.1145/985692.985765.
Viégas, F. B., Wattenberg, M., Kriss, J., & van Ham, F. (2007a). Talk before you type: Coordination in Wikipedia. *HICSS'07*. doi:10.1109/HICSS.2007.511.
Viégas, F. B., Wattenberg, M., & McKeon, M. M. (2007b). The hidden order of Wikipedia. In D. Schuler (Ed.), *Online communities and social computing* (pp. 445–454, LNCS, Vol. 4564). Berlin: Springer.
Voss, J. (2005). *Measuring Wikipedia.* Paper presented at the 10th international conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24–28, 2005.
Watts, D. J., Dodds, P. S., & Newman, M. E. J. (2002). Identity and search in social networks. *Science, 296*, 1302–1305.
Weber, J. (2003). Individual accessibility and distance from major employment centers: An examination using space-time measures. *Journal of Geographical Systems, 5*(1), 51–70.
Wikimedia Foundation (2010). List of Wikipedias. http://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed Sept 2010.
Wikipedia (2008). WikiProject geographical coordinates. http://en.wikipedia.org/wiki/Wikipedia:GEO. Accessed Mar 2008.
Wilson, A. (1969). Notes on some concepts in social physics. *Papers in Regional Science, 22*(1), 159–193.
Wilson, A. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning, 3*(1), 1–32.
Wöhner, T., & Peters, R. (2009). *Assessing the quality of Wikipedia articles with lifecycle based metrics*. 5th international symposium on wikis and open collaboration. doi:10.1145/1641309.1641333.

Zachte, E. (2009). Wikimedia visitor log analysis report: Google requests as daily averages, based on sample period [November 2009]. http://stats.wikimedia.org/wikimedia/squids/SquidReportGoogle.htm. Accessed Feb 2010.

Zachte, E. (2010a). Wikimedia report card [January 2010]. http://stats.wikimedia.org/reportcard/. Accessed Feb 2010.

Zachte, E. (2010b). Wikipedia statistics: Overview of recent months. http://stats.wikimedia.org/EN/Sitemap.htm. Accessed Feb 2010.

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., & McGuinness, D. L. (2006). Computing trust from revision history. *International Conference on Privacy, Security and Trust.*. doi:10.1145/1501434.1501445.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries, 7*(1), 5–16.

Zook, M. (2005). The geographies of the internet. *Annual Review of Information Science and Technology, 40*(1), 53–78.