The Pennsylvania State University

The Graduate School

College of Earth and Mineral Sciences

A GEOVISUAL ANALYSIS OF SOCIAL INFLUENCE IN OPENSTREETMAP CONSTRUCTION

A Dissertation in

Geography

by

Sterling Daniel Quinn

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2016

The dissertation of Sterling Daniel Quinn was reviewed and approved* by the following:

Alan M. MacEachren
Professor of Geography
Dissertation Adviser
Chair of Committee

Anthony C. Robinson
Assistant Professor of Geography

Clio Andris
Assistant Professor of Geography

Guoray Cai
Associate Professor of Information Science and Technology

Cynthia Brewer
Professor of Geography
Head of the Department of Geography

*Signatures are on file in the Graduate School

ABSTRACT

With the Web 2.0-induced rise of online user-generated content, digital databases of "volunteered geographic information" (VGI) are beginning to emerge, ranging from locational coordinates attached to social media posts, to full-fledged cartographic databases. One of the most well-known examples of VGI is OpenStreetMap (OSM), wherein anyone with an Internet connection is invited to contribute to an open source digital map of the world. Although OSM allows a new bottom-up approach to geographic data collection and a possible way out of "cartographies of silence" identified by critical human geographers, the project suffers from participation divides observed in other media sources wherein the Global North produces the majority of the content. In this dissertation, I explore spatial questions about how OSM participation divides play themselves out at different scales, how they affect map contents, and how individual influences in OSM construction can be better understood using geovisualization and visual analytics. In Chapter 1 I lay out the context for the overall work and describe the research goals and objectives. In Chapter 2 I employ automated language identification software to map the geolinguistic footprints of OSM contributors, revealing which areas are mapped by local-language speakers and which are likely mapped from afar. I further investigate varying priorities that users of each language exhibit toward mapping different items. In Chapter 3 I undertake a qualitative and quantitative analysis of the OSM history in small cities, constructing a picture of each contributor's work on the map and developing a typology of contributor motivations. I show how OSM contributors approach the map at different scales with widely varying motives, including strong love of place, business incentive, and an affinity for particular feature types. Chapter 4 describes the design, implementation, and application of an interactive visual analytics tool to help users compare and make sense of the history of OSM construction in small cities across the world. The concluding Chapter 5 highlights salient results, challenges, and

avenues for future work. Such investigations are helpful for institutions examining OSM's fitness-of-use for their projects, as well as researchers studying crowdsourced online databases in general.

TABLE OF CONTENTS

List of Figures viii

List of Tables x

List of Abbreviations xi

Acknowledgements xii

Preface xiii

Chapter 1 Introduction and justification of the work 1

- Impacts of OSM 2
- Where OSM falls short of the ideal 4
- The lure of OSM and the influence of contributors 5
- Goals and objectives of this research 6
  - A geolinguistic approach for comprehending local influence in OpenStreetMap 8
  - Using small cities to understand the crowd behind OpenStreetMap 9
  - Exploring a crowdsourced map using geovisual analytics 9
  - Concluding chapter 11
- Benefits of the research 11
  - Unique contributions of this research 12
- Chapter 1 references 14

Chapter 2 A geolinguistic approach for comprehending local influence in OpenStreetMap 18

- Abstract 18
- Introduction 19
- Background and related theory 22
  - Who is mapping whom in OSM? 24
- Language identification of OSM contributor comments as a means of assessing local influence 26
  - Parameters and overview of the study dataset 28
  - Validating the language identification 32
  - Summary of spatial patterns among languages 33
  - Evaluating languages against confirmed locations from user wiki and profile pages 35
- Geographies of English usage 38
- What do the users of the different languages prioritize? 44
- Conclusions and directions for future research 46
- Chapter 2 references 51

Chapter 3 Using small cities to understand the crowd behind OpenStreetMap 56

Abstract 56
Introduction 57
Theoretical background and related work 60
Why study small cities? 62
Small cities as a lab for studying contributor motivation and behavior 64
Methods 66
Results 71
Size of the crowd 71
Geographic origins of the crowd 75
Retention of the crowd 78
Scale of edits 82
Contributor motives and activities 84
Conclusions 92
Chapter 3 references 95

Chapter 4 Exploring social influence in a crowdsourced map using geovisual analytics 102

Abstract 102
Introduction 103
Relevant literature: What visual analytics offers OpenStreetMap studies, and vice versa 107
Visual analytics tools for crowdsourced data 110
Design and development of Crowd Lens 113
Contributor list 116
Crowd filters 118
Map... 120
Individual contributor panel 120
Using the components together 122
Evaluation of Crowd Lens 123
Early usability testing 124
Scenarios of use and claims analysis 126
Testing by geospatial technology professionals 136
Discussion of OSM contributor characteristics observed using Crowd Lens 144
Conclusion 147
Chapter 4 references 149

Chapter 5 Conclusion 162

Challenges 166
Avenues for further study 169
Chapter 5 references 174

Appendix A Data from Chapter 2 analysis 177

Appendix B Online user evaluation of Crowd Lens, focused toward geospatial technology professionals 181

Appendix C Introductory documentation presented to Crowd Lens testers in Chapter 4 185
What is Crowd Lens? 185
Getting started 186
Learning about the crowd 188
Learning about individual contributors 189
More information 191

LIST OF FIGURES

Figure 2-1. XML metadata from an OSM changeset 27

Figure 2-2. Number and percentage of changeset comments in the study dataset, by language. 31

Figure 2-3. Number of contributors using a language at least once in a changeset comment. 31

Figure 2-4. Spatial distribution of language use in changeset comments. 34

Figure 2-5. Locations of contributors maintaining a profile or wiki page where a location could be detected, grouped by language used in changeset comments 37

Figure 2-6. Median changesets per contributor when considering language used and determined location 38

Figure 2-7. Dominant language of OSM contributor comments. 39

Figure 2-8. Percent of OSM changesets commented in English. 40

Figure 2-9. GDP (PPP) per capita plotted against percent of OSM changesets commented in English. 41

Figure 2-10. Rank of OSM tag values by language. Note that in OSM tagging, the value "yes" is used when more specific information is not available. 46

Figure 3-1. South American cities whose OSM data was studied in this analysis. 67

Figure 3-2. North American cities whose OSM data was studied in this analysis. 68

Figure 3-3. Small multiple maps of each contributor's work within Tres Arroyos over the history of the OpenStreetMap project. Edited ways are green lines and edited nodes are gray dots. 74

Figure 3-4. Languages favored in OSM changeset comments by top contributors in each city. Places of origin, if known, are listed in parentheses. 76

Figure 3-5. Yearly person-days of mapping 79

Figure 3-6. Percentage of contributors active in 2014. 80

Figure 3-7. Activity levels of top 10 contributors in Rosario, Argentina (large city). 81

Figure 3-8. Activity levels of top 10 contributors in Tacuarembó, Uruguay (small city). 82

Figure 3-9. Changeset bounding boxes for General Pico, colored by language of the changeset comment. 84

Figure 4-1. Crowd Lens user interface. 113

Figure 4-2. Contributor list and interactive map in Crowd Lens 117

Figure 4-3. Crowd filters in Crowd Lens 119

Figure 4-4. Interaction between contributor list and individual contributor panel 121

Figure 4-5. Responses to the pre-assessment question "How confident are you in the quality of OpenStreetMap data for your work purposes?" 139

Figure 4-6. Responses to the post-assessment question "To what degree has this tool affected your perception of the quality of OpenStreetMap data for your work purposes?" 142

Figure 4-7. Number of unique contributors in OSM history files for years 2007–2015. 146

LIST OF TABLES

Table 3-1. Contributor metrics showing the size and activity of the crowd in the cities studied. 72

Table 3-2. Countries of origin determined from public OSM user profile and wiki pages, in the format <number of contributors: country>. 77

Table 3-3 . Roles describing OpenStreetMap contributor activities, with changeset comments representative of each. * indicates a translation from Spanish by the author. 86

Table 4-1. Claims analysis related to the Crowd Lens scenarios of use. 135

Table A-1. 2014 GDP (PPP) per capita and percent of OSM changesets commented in English at the country level 177

Table A-2. 2014 monthly income in Brazilian states and percent of OSM changesets commented in English. 178

Table A-3. Percent of population lacking at least one basic human need in Argentine provinces and percent of OSM changesets commented in English. 179

LIST OF ABBREVIATIONS

The following is a list of common abbreviations or acronyms used in this dissertation.

Acronyms are also spelled out on first use in each chapter.

GB	gigabyte
GDP	gross domestic product
GIS	geographic information system
GPS	global positioning system
GUI	graphical user interface
HCI	human-computer interaction
IBGE	Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics
INDEC	Instituto Nacional de Estadística y Censos (National Institute of Statistics and Census, Argentina)
IRB	institutional review board
NGO	non-governmental organization
OSM	OpenStreetMap
OSMF	OpenStreetMap Foundation
PPGIS	public participation GIS
PPP	purchasing power parity
RAM	random access memory
UGC	user-generated content
VGI	volunteered geographic information
XML	extensible markup language

ACKNOWLEDGEMENTS

This dissertation would not exist without the help and encouragement of my wife Rachel. I thank her and our children for coming to Pennsylvania at this time in their lives to support me in graduate school.

I thank Alan MacEachren for agreeing to oversee this work and being everything that a student could want in an advisor in thoroughness, responsiveness, and insightfulness. I also thank Anthony Robinson for his selfless mentoring over the years, as well as Clio Andris and Guoray Cai for their constant support and always "keeping an open door" as part of my committee. I thank Lakshman Yapa for his friendship and encouragement to study the critical cartography literature, which directly influenced the trajectory of this work.

I am grateful to Greg Milbourne for help with data processing and scripting, and Mark Graham for offering ideas that improved the analysis in Chapter 3. I also thank the anonymous peer reviewers who contributed to the improvement of Chapters 2 and 3. The GeoVISTA Center at Penn State provided computing resources for some of the work described in this dissertation.

Numerous other academics and professionals have influenced my journey toward earning a PhD and I particularly thank Mark Gahegan, David DiBiase, Jim Detwiler, Perry Hardin, and Brandon Plewe. I also appreciate Anne Reuland at Esri for giving me a chance at a job that provided wonderful opportunities for technical training and growth during the formative years of my career.

PREFACE

Chapters 1–3 and Chapter 5 of this dissertation were authored by Sterling Quinn. Chapter 4 was authored by Sterling Quinn as the first author and Alan MacEachren as the second author. In Chapter 4, Sterling Quinn designed and implemented the tool described in the chapter and was the primary author of the chapter text. Alan MacEachren contributed to the design of the user studies, reviewed the tool interface, and suggested revisions to portions of the chapter text.

Chapter 1

Introduction and justification of the work

When we look at an online map, whose influence do we see? In the following dissertation, I address this question with reference to OpenStreetMap (OSM), a project wherein people across the Internet work in a loosely-coordinated "crowdsourced" environment to produce an online map. I explore new ways of understanding who creates geographic content online in a given place, why they choose to map the places they do, and how individual mappers exercise their influences on crowdsourced geographic data.

The past decade has seen rapid improvement in the interactivity of the world wide web, leading to an Internet wherein casual users not only seek information, but also create and share information. This user-generated content (UGC) is manifested in blogs, social media posts, consumer product reviews, video sharing sites, wikis, and even maps. In this environment, projects such as Wikipedia and OSM are built using a crowdsourcing model, seeking to harness the work and knowledge of millions of volunteers to create an information product that would be unfeasible to assemble in any other way.

Founded in 2004 by Steve Coast, OSM is a community-owned project offering the ability for anyone to draw features on the map and "tag" them with attributes. OSM began with a focus in the United Kingdom, partially in resistance to the prices exacted for geographic map data by the government-run Ordnance Survey. Within a few years, the map began to see more rapid

development in other areas of the world, and in 2006 the not-for-profit OpenStreetMap Foundation (OSMF) was created as a fund-raising and governance mechanism for the project. [1]

The schema of OSM attributes is loosely organized and is formalized through a community proposal and voting process. Contributors can use their own personal knowledge, uploaded global positioning system (GPS) tracks, high-resolution imagery, and other open source geographic data to guide their contributions to the project. Incoming data is not always screened for accuracy, although OSMF "working groups" monitor incoming anomalies, imports, and vandalism, sometimes using automated programs or "bots" to detect them. Unlike with Wikipedia, anonymous edits are forbidden in OSM and contributor IP addresses are not publicly revealed.

Because OSM allows for the storage of both spatial geometries and attributes, it is much more than a map. Rather, it is a free and open source geographic database that can be downloaded, restyled, and applied in a variety of cartographic and geographic information system (GIS) contexts including base mapping, thematic cartography, geographic search, geocoding, and turn-by-turn routing. Thus, OSM can be thought of as an enormous collection of voluntarily-offered codified knowledge about the world, not built through any institutional specification, but wholly consisting of what the body of contributors wanted to put in the map.

Impacts of OSM

What happens when the door is thrown open for the contribution of any kind of geographic data into an online map? The process of OSM creation may seem anarchical (recalling the resistance by early OSM mappers to governments that charged fees for the use of geographic

[1] See http://wiki.openstreetmap.org/wiki/History_of_OpenStreetMap for a history of OSM, maintained on the project's wiki page.

data); however, in various places throughout Europe, OSM has steadily grown so that road coverage now meets or exceeds that of the government-produced maps (Kounadi 2009, Haklay 2010, Graser et al. 2014). Interestingly, some government-related organizations such as regional transportation authorities are now turning to OSM due to the flexibility of the data for restyling, search, and rapid updates (McHugh 2014, Perelli 2014).

OSM has further seen uptake in the domains of crisis management and natural hazard planning (Chapman 2014). The project experienced a boost of publicity when hundreds of volunteers flocked to the map-building environment following the 2010 Haiti earthquake to trace and import whatever data they could find from the public domain, creating the most detailed map of Port au Prince ever available to the general populace (Zook et al. 2010). Subsequent crises such as Typhoon Yolanda, the West Africa Ebola outbreak, and the Nepal earthquake have seen similar assistance, with the activities now more formally coordinated by the Humanitarian OpenStreetMap Team. Longer-term efforts such as the US State Department's MapGive and the Red Cross-sponsored Missing Maps project draw on volunteer altruism to improve OSM in traditionally undermapped areas by tracing buildings and infrastructure, with the intent that the data can contribute to long-term mitigation and response efforts for natural hazards (Maron 2015, Patel 2015).

At the more personal scale, OSM offers a way to express one's own geography on the public map. No one is prevented from mapping his or her house, school, or favorite restaurant on OSM, thereby imparting higher visibility and attention to these places while increasing the ability of other individuals to know about and interact with them. While some traditionally marginalized or isolated communities may exhibit suspicion and even hostility at having aspects of their lifestyles recorded on maps (Bryan and Wood 2015), others have welcomed the visibility as a way to assert their identity and presence. The Map Kibera effort to put the largest slum in Nairobi into OSM helped reaffirm the community's existence and needs, while also publicizing resources

about clean water points, health clinics, schools, and so forth that could benefit local residents (Hagen 2010). Some resources related to community health and well-being such as urban agriculture sites may be traditionally absent from institutionally-produced maps, but readily available for representation in OSM, allowing a way of rethinking the place (Quinn and Yapa 2016). The OSM community has the flexibility to decide which features are represented and how.

Where OSM falls short of the ideal

The above OSM-enabled interventions help to fill "cartographies of silence" identified by Harley (1988) in traditional paper maps wherein individuals with power decided (either intentionally or unintentionally) who and what was represented on the map. And yet it may be argued in some ways that OSM is more susceptible to those silences than maps produced by Google, the UK Ordnance Survey, or other institutions who have a business incentive or public mandate to provide full coverage within their domains. Each editor of OSM wields the power to add or erase, and those very editors are not evenly distributed throughout the map, do not often match gender and racial distributions in the mapped population, and fail to apply an equal level of focus to all places and features.

OSM belongs to a class of geographic data commonly called volunteered geographic information (VGI). When Goodchild (2007) popularized this term, he described VGI as being created by "citizens as sensors". Human sensors are less predictable than mechanical sensors and it is difficult to "calibrate" them to focus on providing comprehensive geographic coverage in a consistent fashion (although this has been the focus of tools such as the Humanitarian OpenStreetMap Task Manager). Graham (2010) has thus noted the potential of VGI to suffer from "virtual black holes", which I liken to the cloud cover, engineering failures, or atmospheric interference that sometimes causes gaps in the data returned by mechanical sensors.

The potential for places to be consistently neglected in OSM due to human bias was readily admitted by the project's founder Steve Coast (GISPro 2007) and has been investigated by Haklay (2010), Neis et al. (2013), and others. These omissions are sometimes manifested in empty map screens in places where there is disproportionately low connectivity or participation. Traditionally unbalanced geographies of information production in which the Global North produces most codified information (Graham et al. 2014) might explain many of these situations where entire cities and suburbs in Africa, Asia, and elsewhere are drawn using no more than a few lines representing the main routes through town. In other cases the map seems to be "full" of data, but is missing the perspective that could be included if more mappers, or at least a more representative sample of mappers, were influencing the project (Wood 2014). For example, Stephens (2013) described ways that a more balanced gender makeup of contributors might enrich the ontology of features approved by the OSM community.

The lure of OSM and the influence of contributors

Despite these shortfalls, OSM continues to enjoy rapid uptake in a various business, government, and non-governmental organization (NGO) domains. One place to see this manifested is the myriad presentation topics at the annual "State of the Map" conferences around the world. At the 2014 conference in Buenos Aires (the first held in South America) I interviewed several attendees from Argentina and Uruguay and asked them how they discovered OSM and why they were using it. They mentioned feeling compelled to seek out alternatives because Google Maps coverage was not acceptable in their local work areas. OSM was particularly attractive because its coverage was better, it was free to use, and it could be modified if needed. Still, the interviewees confessed little understanding of the pedigree of the OSM data, including who created it, how old it was, and how accurate they could expect it to be. When forced to select

between a map of OSM data or no map at all, the way forward was clear; and yet other institutions that have a choice between commercial, government, and open source data may desire to understand how often the open source data is being maintained, and how many contributors are actively augmenting, evaluating, and fixing the data. This information could help them make a more informed decision about whether the benefits of a switch to open source data will outweigh some of the risks associated with allowing anybody to edit the data, a particular concern for governments who fear liability for showing potentially inaccurate or vandalized maps (Johnson and Sieber 2013, Ballatore 2014).

There are other reasons it would be helpful to understand who is putting data into OSM and why. In 2015 at the State of the Map US conference, I met three employees of a large well-known social media corporation. They were beginning to explore OSM data as a way of augmenting their location services. Outside in the conference exhibit space, vendors were promoting OSM as a source of base maps for business applications, and as a back end data source for vehicle routing and navigation. All of these applications of OSM can potentially affect the user's perceptions of the physical world and their interactions with it (Graham et al. 2013). Our choice of a grocery store, transportation mode, driving route, or child care facility, can rest in the set of items returned from a search of a geographic database, initiated by a few taps on a smartphone. In this scenario OSM contributors exercise a direct influence on what is known or can be known about the world.

Goals and objectives of this research

Each place in OSM is created by a finite set of contributors with different focuses, motives, and backgrounds. The goal of this dissertation is to develop and evaluate methods and tools for better understanding these social influences behind OSM production and their

geographic variations, in particular: (1) the characteristics of the OSM contributor set in a given place, and (2) the unique influence on OSM data exercised by individual contributors to the OSM project. I have limited this goal to OSM (rather than addressing VGI more broadly) due to the unique global nature of the OSM project and its rapid adoption as an alternative to government and commercially produced geographic data; however, I comment throughout on ways that findings from this work might apply to VGI in general.

In support of the above goal, the following specific research objectives are pursued in this dissertation:

- Detect and map the languages used by OSM contributors as they comment on their work, while developing and testing a method to understand how much these geolinguistic footprints reveal about the amount and nature of locally contributed data (as opposed to remotely contributed "armchair mapping" efforts).
- Use historical OSM metadata and publicly-available user profile pages to make a qualitative and quantitative evaluation of how OSM takes shape in small cities, outside of the attention of large urban mapping communities.
- Develop a geovisual analytics tool for exploring the characteristics of the OSM contributor crowd in different places, while allowing interactive inquiry of how the individual actions of each contributor have combined to construct the present map.

The body of this dissertation explains how these objectives were addressed. It consists of three studies produced for submission to scholarly research journals. Each study is a self-contained chapter offering its own review of relevant literature and theory. Although any chapter can stand independently, they each contribute toward the goal of understanding the set of people behind OSM and the effect each person has on the map. The three objectives were addressed in the order listed above, therefore methods from Chapter 2 (such as automated detection of languages) were applied where appropriate in the later chapters, and the OSM metadata

calculations and visualizations introduced in Chapter 3 were often re-invoked and expanded in Chapter 4.

The focus of each study is detailed briefly below.

A geolinguistic approach for comprehending local influence in OpenStreetMap

The first study (Chapter 2) proposes and applies a new method for comprehending the degree of local vs. nonlocal influence in the map by detecting the mix of languages used by OSM contributors. Some OSM contributors map faraway places by tracing remotely sensed photographs, bulk importing datasets, or fixing logical errors in the database. Other contributors have access to the place being mapped and can perform local data collection. I use automated language detection software to identify languages used by editors of the South American map in OSM. These languages are then compared with home locations reported in the profiles of OSM contributors to provide an idea of how much South American influence can be inferred when looking at Spanish-commented edits, English commented edits, and so forth. I analyze how the use of English varies in prevalence across the urban-rural fabric and among household income levels. I also reveal ways that different contributor language mixes tend to affect the types of things that get added to the map. In particular, I focus on the propensity of Spanish and Portuguese-speaking contributors to map items related to everyday livelihoods and routines in South America, when compared with English speaking contributors who are more likely to map sites of tourism and large features that can be traced from remotely sensed imagery.

Using small cities to understand the crowd behind OpenStreetMap

The second study (Chapter 3) examines the size, origin, and motives of the "crowd" behind OSM in small cities. Existing studies of OSM have focused on urban areas or entire countries. They neglect small cities and rural places, and yet many of the applications where OSM is being invoked require comprehensive coverage. Furthermore, businesses and governments in smaller cities may be attracted to OSM as a way to achieve effective and flexible local maps without being tethered to a commercial product. I therefore take advantage of the finite and bounded contributor set in small cities to drill into a deeper qualitative analysis of the habits and motives of OSM contributors. I select three small cities in South America and compare them with other large and small cities to understand the typical size of the contributor set in these places. I use sets of small multiple maps to show each contributor's piece of the city, while also mapping bounding boxes of edits to the city in an analysis of the scales at which contributors approach the map. Other factors such as each contributor's number of days active and preferred language used in the metadata (applying methods from Chapter 2) are also taken into account. These analyses reveal that OSM is created by a mix of data importers, bots, paid mappers, and people who have a topophilic connection to the place, an understanding somewhat more complicated than the original "citizen sensor" conceptualization of VGI. I therefore conclude by introducing a typology of OSM contributor roles that I derived from a study of the open-ended contributor comments included in the project metadata.

Exploring a crowdsourced map using geovisual analytics

The diverse human motives and practices behind OSM are difficult to discern from the flattened map on openstreetmap.org, therefore the third study (Chapter 4) describes the design,

implementation, and testing of a geovisual analytics tool for examining the collective and individual influences of the OSM contributor crowd in a given place. The tool reveals the size of the crowd behind OSM and how this fluctuates across different map extents. It allows the user to filter the OSM contributor crowd based on particular criteria such as language used, types of entities added, or time frame of activity in the project, while also allowing a detailed look at the publicly available attributes of any one contributor. The tool includes a map that links OSM features with the free-form commit messages supplied by their contributors.

The main purpose of the tool is to assist professionals in the geospatial technology industry with understanding how confident they can be in the attention paid to OSM data in their respective regions of focus; however, it is also expected to be of interest to social scientists studying Internet crowdsourcing patterns in general In many ways this tool is the culminating work of the dissertation, drawing on methods that were introduced and initially evaluated in the other two studies. Static visualizations and statistics from Chapter 3 are enhanced into an interactive format in this tool, allowing faster and more comprehensive analysis. The language identification method described in Chapter 2 is again applied in this study, such that users of the tool can view the language used most often by any individual contributor and filter the crowd by this characteristic as well. I focus this initial implementation of the tool itself on small cities in multiple continents to further investigate the findings from Chapter 3: that outside major urban areas, OSM is still often influenced by just a handful of local contributors whose work is supplemented by a much larger crowd of digital "passers by" submitting fixes and imports from faraway locations.

Concluding chapter

A brief concluding chapter (Chapter 5) discusses how the results of the above three sections link together. It also describes challenges encountered along the way, and lays out potential directions for future study.

Benefits of the research

The above pieces of this dissertation contribute to at least three broader domains of study:

This research benefits scientists who are studying Internet-based crowdsourcing and UGC. It helps address the questions, “What do we get when we open the doors to anybody contributing to an information product?” How close are we really coming to reaching "the sum of all human knowledge" as Wikipedia founder Jimmy Wales aspired[2] , and where might we be perpetually falling short? In what ways do projects like Wikipedia and OSM democratize information production, or reinforce existing disparities?

Furthermore, in its use of geovisual analytics to explore OSM, this dissertation contributes to research on visual analysis and sensemaking of large messy datasets, particularly crowdsourced ones. The open tagging format of OSM, the multiple versions of features, the temporal dimension of the data, the geometries of mapped features, and the nuances of focus exhibited by individual contributors all present challenges to comprehending trends in crowdsourced data contribution. These are problems that visual analytics and its spatially-focused subdomain of geovisual analytics (Andrienko et al. 2007) are well-suited to address.

Finally, this dissertation will be of value to social scientists and GIScientists interested in the social ramifications of online user-generated map content, often associated with terms such as

[2] http://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds

the "geoweb", volunteered geographic information, and neogeography. OSM certainly incorporates elements from all three of these phenomena in its reliance on geographic content submitted over the Internet by people of potentially any skill level. In fact, Haklay (2014) has proposed that OSM is a unique enough project that it may qualify for its own realm of studies within the literature of VGI. This dissertation contributes to such "OpenStreetMap studies" in several unique ways described below.

Unique contributions of this research

This research goes beyond what I call the "node counting" approach of various other OSM studies that have primarily concentrated on the number of features added to the database. By paying special attention to contributors' free-form comments and profile pages, it adds a qualitative and personal dimension to a body of research that has often seemed most concerned with the quantity and positional accuracy of the data. In this regard, it helps with understanding how each contributor's mapping pattern came about.

This dissertation is also unique in its focus on places not heavily studied in existing OSM research, offering a geographic emphasis on the Global South and small urban areas. In order to achieve truly comprehensive coverage for applications such as routing and search, the map must also exist in these places, not just large metropolitan areas in the Global North.

Finally, this research proposes ways that visual analytics methods can demystify the contributor crowd behind VGI and other online information products, at both the group and individual levels. Although some Wikipedia-focused visual analytics research has addressed "edit wars" and conflict between contributors, there is less attention to visualizing the overall contributor set and understanding what each person brings to the project. Similarly, existing

visual analytics approaches for interacting with OSM have focused primarily on the specs and characteristics of the OSM data rather than the individuals and groups who contributed it.

An informed decision about whether to use a dataset often depends on the available metadata and what it says about who created the data and how data collection was approached. OSM offers abundant metadata in its "full history dump" and "full history changeset" files; however, it is presented in a compressed extensible markup language (XML) format that is not easily interpreted by lay users. The methods and tools evaluated in this dissertation are intended to help individuals in the public and private sector to better comprehend the "who" and "how" behind OSM construction, knowledge that can inform decisions about whether to use the data and how reliable OSM may prove for the purpose of any given project.

Chapter 1 references

Andrienko, Gennady, Natalia Andrienko, Piotr Jankowski, Daniel Keim, M.-J. Kraak, Alan MacEachren, and Stefan Wrobel. 2007. "Geovisual Analytics for Spatial Decision Support: Setting the Research Agenda." *International Journal of Geographical Information Science* 21 (8): 839–57.

Ballatore, Andrea. 2014. "Defacing the Map: Cartographic Vandalism in the Digital Commons." *The Cartographic Journal*. http://www.maneyonline.com/doi/abs/10.1179/1743277414Y.0000000085.

Bryan, Joe, and Denis Wood. 2015. *Weaponizing Maps: Indigenous Peoples and Counterinsurgency in the Americas*. New York, NY, USA: Guilford Press.

Chapman, Kate. 2014. "Building a Community: HOT in Indonesia." In *State of the Map US 2014*. Washington, DC. https://vimeo.com/91896398.

GisPro. 2007. "The GiSPro Interview with Steve Coast," October.

Goodchild, Michael F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–21.

Graham, Mark. 2010. "Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth." *Tijdschrift Voor Economische En Sociale Geografie* 101 (4): 422–36.

Graham, Mark, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. 2014. "Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty." *Annals of the Association of American Geographers* 104 (4): 746–64.

Graham, Mark, Matthew Zook, and Andrew Boulton. 2013. "Augmented Reality in Urban Places: Contested Content and the Duplicity of Code." *Transactions of the Institute of British Geographers* 38 (3): 464–79.

Graser, A., M. Straub, and M. Dragaschnig. 2014. "Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph." *Transactions in GIS* 18 (4): 510–26.

Hagen, Erica. 2010. "Putting Nairobi's Slums on the Map." *Development Outreach | World Bank Institute*, July, 41–43.

Haklay, Mordechai. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning. B, Planning & Design* 37 (4): 682.

———. 2014. "OpenStreetMap Studies (and Why VGI Not Equal OSM)." *Po Ve Sham - Muki Haklay's Personal Blog*. August 14. https://povesham.wordpress.com/2014/08/14/openstreetmap-studies-and-why-vgi-not-equal-osm/.

Harley, J. Brian. 1988. "Silences and Secrecy: The Hidden Agenda of Cartography in Early Modern Europe." *Imago Mundi* 40 (1): 57–76.

Johnson, Peter, and Renee Sieber. 2013. "Situating the Adoption of VGI by Government." In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, edited by Daniel Sui, Sarah Elwood, and Michael Goodchild, 65–81.

Kounadi, Ourania. 2009. “Assessing the Quality of OpenStreetMap Data.” *Msc Geographical Information Science, University College of London Department of Civil, Environmental And Geomatic Engineering*. ftp://ftp.cits.nrcan.gc.ca/pub/cartonat/Reference/VGI/Rania_OSM_dissertation.pdf.

Maron, Mikel. 2015. “For Government, OpenStreetMap Is More than Excellent Data. It’s a Transformation.” In *State of the Map US 2015*. New York, NY, USA. http://stateofthemap.us/for-government-openstreetmap-is-more-than-excellent-data-its-a-transformation/.

McHugh, Bibiana. 2014. “Government as a Contributing Member of the OpenStreetMap (OSM) Community.” In *FOSS4G 2014*. Portland, Oregon. https://vimeo.com/album/3606079/video/106226528.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf. 2013. “Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions.” *Future Internet* 5 (2): 282–300.

Patel, Drishtie. 2015. “Missing Maps.” In *State of the Map US 2015*. New York, NY, USA. http://stateofthemap.us/missing-maps/.

Perelli, Julián. 2014. “Transporte público al alcance de todos.” In *State of the Map 2014*. Buenos Aires, Argentina. https://vimeo.com/album/3134207/video/112243714.

Quinn, Sterling, and Lakshman Yapa. 2016. “OpenStreetMap and Food Security: A Case Study in the City of Philadelphia.” *The Professional Geographer*. doi:10.1080/00330124.2015.1065547.

Stephens, Monica. 2013. “Gender and the Geoweb: Divisions in the Production of User-Generated Cartographic Information.” *GeoJournal* 78 (6): 1–16.

Wood, Harry. 2014. “The Long Tail of OpenStreetMap.” In *State of the Map 2014*. Buenos Aires, Argentina. http://vimeo.com/album/3134207/video/112438218.

Zook, Matthew, Mark Graham, Taylor Shelton, and Sean Gorman. 2010. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake.” *World Medical & Health Policy* 2 (2): 7–33.

Chapter 2

A geolinguistic approach for comprehending local influence in OpenStreetMap[3]

Abstract

OpenStreetMap (OSM) thrives on allowing anyone in the world to contribute features to a free online geographic database, thereby allowing international mixes of contributors to create the map in any given place. Using South America as a test area, I explore the geography of OSM contributors by applying automated language identification to the free-form comments that contributors make when saving their work. By cross-referencing these languages with self-reported hometowns from user profiles, I evaluate the effectiveness of language detection as a method for inferring the percentage of local contributors versus the percentage of "armchair mappers" from elsewhere. I show that most English-speaking contributors to the South American OSM are from outside the continent (rather than multilingual locals). The percentage of English usage is higher in poor areas and rural areas, suggesting that residents of these places exercise less control over their map contents. Finally, I demonstrate that some features related to daily needs of health, education, and transportation are mapped with higher priority by contributors who speak the local language. These findings give researchers and organizations a deeper understanding of the OSM contributor base and potential shortcomings that might affect the data's fitness for use in any given place.

[3] This chapter contains a slightly edited version of a paper that has been accepted for publication as:

Quinn, Sterling. in press. "A Geolinguistic Approach for Identifying Locally Contributed Content in OpenStreetMap." Cartographica. doi:10.3138/CART.3301.

Introduction

Volunteered geographic information (VGI) by one of its earliest definitions relies on citizens as sensors to gather information about the world around them (Goodchild 2007). It is tempting to think of these citizens as a data collecting army, sweeping the globe almost like a satellite in an orchestrated march toward recording and filing away information about everything observed. With such a view, it can be easily forgotten that these citizen sensors exhibit calibrations, capabilities, and geographic distributions that are in no way uniform. Unguided by algorithms or code, each brings their own experiences, motives, and geographies to the data collection process. Factors such as gender, income, technical skill, travel experience, love of a particular place, personal hobbies, and work-related requirements may affect the items that contributors prioritize in the map. This chapter focuses on how VGI contributors' places of residence influence the collected data in a particular region, specifically examining the geographic distribution of local versus nonlocal contributors and the types of map features prioritized by each.

The OpenStreetMap project (OSM) is an especially suitable laboratory for studying the ways that contributor location affects VGI. OpenStreetMap is a digital map of the world wherein any person with an Internet connection is invited to share information about the coordinates and attributes of any place. All contributions are poured into a single online geographic database that can be downloaded and represented in different ways by cartographers and geographic information systems (GIS) analysts. When viewing cartographic depictions of OSM, it is easy to overlook the fact that the data is a patchwork affair of contributions from editors in different places with different resources and motives.

For example, contributors who have never visited a place can still add to OSM by tracing items from satellite and aerial photography, which is often freely included as a background layer

in the OSM editing programs. This "armchair mapping" approach is convenient when no other information is available, and has often been employed in humanitarian situations when a vector map of roads and infrastructure is needed in a hurry (Zook et al. 2010); however, tracing imagery alone limits the categories of entities that can be discerned for addition to the map, and is dependent on the availability of current images. This inevitably results in a shallower product. Long before OSM and the popular uptake of the World Wide Web, Harley (1990) warned of the loss of cartographic information that could result from replacing field surveys with a heavier reliance on remotely sensed imagery. Contributors can perhaps get this "field" knowledge of remote places through vacations, tours, business trips, educational excursions, humanitarian service, or military experience. But beyond even this, I propose that there remain integral aspects of everyday human routines that are still stubbornly difficult to collect without the participation of people who spend long periods of time living in or near the mapped area, and that a greater level of participation by local residents raises the priority given in OSM toward mapping these everyday needs.

But how can it be known whether a contributor is "local", and how does the presence of local contributors vary across space? A challenge of studying the geographies of OSM contribution is that OSM itself does not directly store or depict any information about contributor location. In this chapter I study the degree to which this barrier can be overcome by identifying and mapping patterns of language use among contributors. These geolinguistic contours can reveal subtleties in local participation patterns previously missed by coarser-grained reports. For example, it is known that the Global North dominates in the production of digital information (Graham et al. 2014b, 2015), and that this pattern holds true in OSM in the sense that most contributors are from Europe and North America (Neis and Zipf 2012); however, less attention has been given to the finer-grained geographies of OSM participation occurring in the Global South. My research indicates that some areas in the Global South maintain a substantial

percentage of local OSM contributors, while other areas appear to be created mostly by armchair mappers working from afar.

The objectives of this research are threefold, all working toward the goal of shedding light on the spatial distributions of OSM contributors and the ways that their locations affect map contents. First, I evaluate a method for detecting the relative amount of local influence in OSM, by analyzing the languages contributors use when they make metadata comments about their edits. Second, I map and interpret spatial patterns in language use among OSM contributors, focusing particularly on the use of English and its relationship to other sociodemographic variables such as income and rurality. Third, I determine which types of geographic entities (such as amenities, businesses, etc.) in OSM are highly prioritized by local language contributors versus those prioritized by nonlocal language contributors.

I use South America as a study area for these efforts, due to its manageable number of languages, its small percentage of English speakers, and the dearth of academic research on OSM in the Global South in general. Selecting an entire continent for analysis allows for some comparison between countries and regions, and permits follow-up at finer-grained scales. Finally, South America is a place where wide variations in OSM coverage are visible as one navigates across cities and countries. Some metropolises are missing entire neighborhoods on the outskirts of the urban core (or just have basic features traced from aerial photographs), while in other locations it appears that mappers with an intimate knowledge of local neighborhoods have dedicated much time and energy toward the map.

Beyond evaluating a new method to assess the local character of OSM edits, this study shines a light on the data integrity of OSM, or at least perceived integrity, in a region where the map has not received much scholarly attention. A large percentage of contributions coming from outside the locality may engender concerns of error, bias, or vandalism, whether intentional or unintentional. The results of this research will help local organizations better understand the

fitness-for-use of OSM for their projects, while also guiding OSM communities to the places and contributors that might benefit from further attention. The study also explores the potential for local contributors to generate kinds of information critical to the health of communities that may be missed by top-down institutional mapping efforts.

Background and related theory

Although the terms of use and contribution for an online collaborative project such as OSM might be wide open from a strictly legal standpoint, the project itself might remain inaccessible to many. Graham et al. (2014b) showed that access to broadband is necessary for achieving high levels of user-generated content (UGC), but noted that connectivity alone does not resolve gaps in representation and participation, nor does it immediately change cultural attitudes toward the technology. For OSM contributors, some degree of English language knowledge is required to understand the tagging metadata applied to geometric features to associate them with real-world entities. For example, a tag of amenity=school applied to a point or polygon identifies a school. These tags are not translated into other languages, since they are frequently read and interpreted by automated computer programs.

A mapping project such as OSM may therefore be lacking contributions by local residents in places where English is not widely spoken, perhaps missing some of the features that would be helpful in meeting the needs of residents, such as transportation, civil services, and health care options. These omissions can be hard to notice in cases where the map already appears "full" of data contributed by nonlocal residents tracing aerial imagery or other persons importing bulk datasets from a national or provincial scale. Challenging the notion that neogeography efforts such as OSM reflect a fully "democratized" product, Haklay (2013, p. 67) remarked, " we need to take into account the everyday geography of communities in streets,

villages, and slums and find ways to ensure that the technical codes of neogeography provide the space for the voices from these places to be heard and represented." In this spirit, Elwood (2008) observes that the presence of VGI is a marker of inclusion and empowerment for the people and places it represents.

This reflects a hope that a wider group of citizen sensors will contribute information about the world around them, and in the Global South, this means residents of the Global South. A notable effort of this sort is described by Hagen (2010), wherein local residents of a Nairobi slum mapped clinics, potable water outlets, toilets, places of worship, and other points of interest that would help to meet local basic needs and provide evidence of the community's existence when lobbying for civil services. Also common in OSM is that a single enthusiastic contributor will take it upon his or her-self to map a neighborhood or hometown in great detail, providing information only obtainable on the ground such as street names, house numbers, business names, and so forth. These efforts realize a vision by Crampton (2009) and other critical cartographers of open source tools being employed by the disempowered for the advancement of counter-knowledges and counter-mapping, thereby lending evidence to Harley's (1988) assertion that "there is no such thing as an empty space on a map".

Because of VGI's reliance on self-selected human sensors, projects such as OSM pose special geographic questions for study. Each contributor brings knowledge of a unique set of places. In particular this includes locations where the contributor has lived, but the realm of known places may also extend across the globe for people who have traveled often or who are able to interpret and trace aerial photography. The number of contributors and their collective levels of place-specific knowledge vary across any chosen extent or scale within the map.

This mix of contributors and their places of expertise are often unknown to the end user of the VGI. In many cases, GIS analysts and cartographers are accustomed to using datasets gathered by relatively small groups of highly trained individuals who use strict quality control

procedures and carefully calibrated equipment. Although these individuals may not have personal experience in the places being mapped, they employ systematic data collection techniques to retrieve the needed information. When making a switch to VGI, it can be easy to forget how radically the data collection procedures differ from these more traditional datasets. The approach in VGI is bottom-up, rather than top-down, opening up doors to represent new places and people, but also holding the potential for human omissions and idiosyncrasies to slip into the map.

I have observed that professionals encountering OSM online often do not understand or question the human biases and variations in the data as long as it appears complete on the surface. Yet, a visually "busy" map can mask deficiencies in the data that affect the map's usefulness for certain purposes (Quest 2014). The unstructured nature of the contributors is disguised by the uniform digital symbols and labels used across the extent of the map, and may be easily forgotten or neglected by organizations considering the data for practical applications such as disaster response, routing, urban planning, or scientific research. This is not a trivial issue, because the digital map itself can affect users' perceptions, use, and (re)production of physical space (Zook and Graham 2007, Graham et al. 2013). For example, we can imagine that a restaurant, church, garden, or other amenity placed on the digital map might attract more attention in the physical world than its unmapped counterparts, thus strengthening its role in the creation of the place.

Who is mapping whom in OSM?

By far OSM's largest contributor base is in Europe (Neis and Zipf 2012), where the project was founded as an open source alternative to the fee-laden GIS data offered by government organizations such as the UK Ordnance Survey. A now sizable list of academic investigations including Haklay (2010), Girres and Touya (2012), and Neis, Zielstra, and Zipf (2011) has shown that in numerous European cities, OSM coverage and precisional accuracy

meet or exceed that of institutionally-produced alternatives. At the same time, other regions of the world have seen comparatively little activity in OSM (Latif et al. 2011, Neis et al. 2013). In many cases these stagnant areas appear to be places with less wealth, following a pattern Haklay (2010) observed of deprived regions in the UK receiving less attention in OSM.

When viewing OSM in these regions where the map is still developing, it is natural to ask how much of the contributed information is coming from Europe or elsewhere overseas, and how much is coming from a local audience. Furthermore, it makes sense to investigate the different ways in which local and nonlocal contributors affect the character of the data contributed to OSM. Ultimately this is a question of who is mapping whom, relating to the myriad lived experiences and knowledges of place brought to the map by unique cultures and peoples. These questions are not limited to geographical locations of contributors. Stephens (2013) noted ways that the wide gender imbalance in OSM contributors affects not only the prevalence of certain types of entities in OSM, but also the community voting processes that determine the accepted ontologies of entities. Her conclusion that "In a map or be mapped world, men are mapping and women are being mapped" provoked self-reflection within the VGI community (Wright 2013, Leszczynski and Wilson 2013).

Could similar differences in map content exist when populations in one locale are mapping populations in another locale? How is a map made by locals different from a map that is made largely by external influence? The web in general is affected by geographic disparities in the quantity and focus of UGC, a phenomenon which can sometimes be detected by considering regions of language use. In their study of languages used in Google Maps-indexed content, Graham and Zook (2013) observe that "the digital footprints of languages on the geoweb are readily visualized and in some cases can be particularly sharp" (p. 89), and "not only does the density of linguistic footprints vary over space, but *their potential objects of attention also differ substantially*" (p. 91 – 92, emphasis added). The methods and results below explore the extent to

which geolinguistic patterns can infer the degree of local participation in OSM across space, while attempting to better understand how this geographic distribution of OSM contributors affects the end data product.

Language identification of OSM contributor comments as a means of assessing local influence

Contributor locations are not systematically reported in any of the OSM data or metadata; therefore, other clues must be exploited in order to gain some kind of picture of the geographic distribution of OSM contributors. The OSM system administrators have access to the IP addresses of contributors (revealed in articles such as Maron et al. 2012) which could be geocoded to map contributor locations; however, these addresses are not available with the same degree of open access as the OSM data itself. Another means of detecting contributor locations would be to examine the thousands of user-created wiki and profile pages created by OSM contributors, some of which contain autobiographical information revealing the user's hometown (eg, "I'm a software engineer in Rio de Janeiro, Brazil"). This technique is not very scalable and it misses the large percentage of users who do not create any profile; however, it does provide some measure of "ground truth" that can be used to evaluate other methods.

As a case in point, whenever OSM contributors save their work (a unit of contribution known as a "changeset"), they are invited to leave a message describing their edits. Contributors use this opportunity to supply the rationale, evidence, or justification behind the set of changes they are saving. These changeset comments are written in a great variety of languages and are linked to the geographic coordinates of the changesets. I propose that the languages used in these comments could be detected by automated software and then cross-matched against self-reported locations in OSM user profiles to understand how much local and nonlocal influence is associated

with each language. This analysis could offer insights about the geographic distribution of OSM contributors in a region, as well as the types of entities favored by these contributors in their volunteer mappings.

Changeset comments range from empty space to abbreviated notes to verbose prose. All of the messages are saved in the full changeset history file made available to the public at planet.osm.org. In addition to the contributor comments, the changeset metadata includes the geographic bounding box of the edits. The bounding box can be used to map the general location of the changes. Each changeset also has a unique ID that can be linked with items in the OSM full history dump files to examine the actual geometry and attribute modifications in greater detail. The methods described below take advantage of each of these pieces of information.

Figure 2-1 shows the metadata for a single changeset taken from the full changeset history. The XML shows that the changeset has an ID number of 21551743 and was made by user Sidromano on April 7, 2014 in southern Brazil. The changeset affected 14 features and the editor used was the iD browser-based editor, version 1.3.8, with Bing Maps imagery in the background.

```
<changeset id="21551743" created_at="2014-04-07T12:43:21Z"
num_changes="14"   closed_at="2014-04-07T12:43:22Z" open="false"
min_lon="-53.0632014" min_lat="-27.3402627" max_lon="-
53.0400835" max_lat="-27.2960595"  user="Sidromano"
uid="1835764">
    <tag k="comment" v="Incluindo vias que não estavam mapeadas
e/ou corrigindo as existentes." />
    <tag k="created_by" v="iD 1.3.8" />
    <tag k="imagery_used" v="Bing" />
  </changeset>
```

Figure 2-1. XML metadata from an OSM changeset

The comment for this changeset is: "Incluindo vias que não estavam mapeadas e/ou corrigindo as existentes". An analyst familiar with Portuguese would know that in this changeset

the contributor is adding previously unmapped tracks and correcting existing ones. However, computer-automated language identification methods are the only feasible way to process and map the large volume of comments and the variety of languages in the OSM changeset history file. Researchers in natural language processing have already tackled the problem of detecting the language of a short piece of text, and these approaches are documented in a growing body of literature. In this analysis, I use the langid.py library developed by Lui (https://github.com/saffsd/langid.py) due to its design for diverse domains, its open source distribution (making it freely accessible), and its ease of integration with the Python scripting language already being used in the project. Details of the language processing algorithm and its packaging into this software library are described in Lui and Baldwin (2011, 2012).

Parameters and overview of the study dataset

To evaluate the changeset comments, a Python script was used to parse all items catalogued in the publicly available full changeset history file, downloaded from the OSM website on August 27, 2014. Because user comments were introduced into the OSM metadata in 2009, this resulted in over five years worth of changesets to analyze. Each changeset was required to meet the following criteria in order to be included in the study dataset:

- The centroid of the changeset must fall within mainland South America or close-lying islands belonging to South-American countries, with the acknowledged limitation that this excludes some remote islands such as the Galapagos and island nations in the Caribbean.
- The bounding box of the changeset must be less than 0.5 degrees of longitude wide and 0.5 degrees of latitude high. This ensures that uncommonly large edits, such as updates of country boundaries, do not disrupt the maps of local patterns of language use.

- The comment must be greater than or equal to 30 characters in length.[4] This ensures that the language identification software has enough characters to evaluate. Although fewer characters could be used, additional research is needed to understand how a lower threshold would affect the accuracy of the language identification with this particular software package.
- The comment must not contain more than two commas. Some contributors just fill the comment with a comma-separated list of places edited, and these lists convey nothing about the contributor's language of choice.
- The language identification must receive a confidence score of greater than 0.99 from the langid.py software. The software has a built-in system of evaluating how confident it is that the identification succeeded. This metric can be used as a threshold for eliminating records whose language was indiscernible. Although a value of 0.99 may seem high, the intent was to start with the records that the software thought it evaluated correctly and evaluate its performance from that point.
- The changeset must not originate from the OpenStreetMap Foundation as part of the 2012 data redaction. This redaction consisted of batch edits related to the implementation of a new OSM license (Wood 2012). All of these messages have the same text and are commented in English. Including these would bias the rest of the analysis.

[4] The script used in this chapter did not count the number of changesets failing to meet the comment length criteria; however, a rough estimate can be derived using the data from the large city of Rosario, Argentina described in Chapter 3. In that dataset, about 3.5% of changesets have a comment at least 30 characters in length. About 41% of changesets in that dataset have no comment at all. Comments were not always possible to add in OSM, so the figures reported in this footnote only consider the time period April 2009 (when the first comment appears in Rosario) to the last week in December 2014. The tendency of some contributors to leave longer comments due to cultural reasons or heavy familiarity with OSM could affect the conclusions extrapolated in this dissertation, and is noted as an important topic for follow-up study.

Out of all changesets meeting the above criteria, I selected those whose comments were identified by the langid.py software as being written in Dutch, English, French, German, Portuguese, or Spanish. Five of these are the major language of at least one South American country, whereas the remaining one (German) is spoken by a sizable body of OpenStreetMap contributors (Neis and Zipf 2012) and was therefore anticipated to have at least some presence on the South American map.

This filtering resulted in a final study dataset consisting of 103,266 changesets for analysis. These were created by 6,502 unique contributors, with a median of 2 changesets per contributor. It should be noted that a total of 1,546 (1.5%) of the changesets that met all other filter criteria were removed because the software detected some language other than the six target languages. Many of these were relatively uncommon languages such as Gaelic that appeared to be miscoded instances of the target languages. Although they could possibly be detected and corrected manually, the purpose of this study is to test the effectiveness of an automated method. Even after the filtering process reduced the number of available changesets that could be analyzed, the result was still a large and geographically representative sample of changesets.

Figure 2-2 shows the number and percentage of changeset comments identified for each language. Portuguese is dominant here, followed by Spanish and then English. The French, German, and Dutch changesets make up about 4% of the total. Note that any of the languages excluded from the analysis (even if they were miscoded instances of the target languages), would constitute less than 2% of the pie if they had been included here.

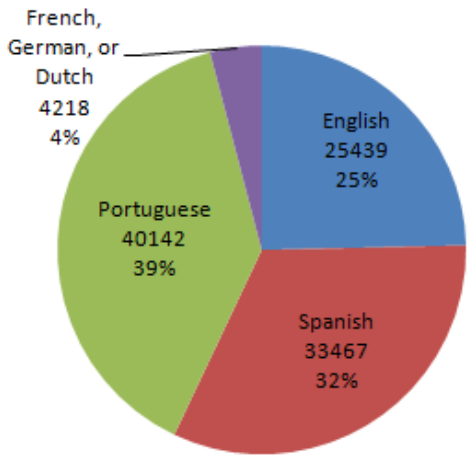


Figure 2-2. Number and percentage of changeset comments in the study dataset, by language.

Figure 2-3 shows the number of contributors employing each language at least once. (Note that the total is greater than the number of unique contributors in the study dataset because some contributors had multiple languages identified among their comments.) Because Portuguese ranks lower here but ranked highest in raw number of changesets, it is clear that some Portuguese-speaking contributors are especially active in the project. Their influence is confirmed in the summary of median changesets per contributor reported in Figure 2-6 later in this chapter.

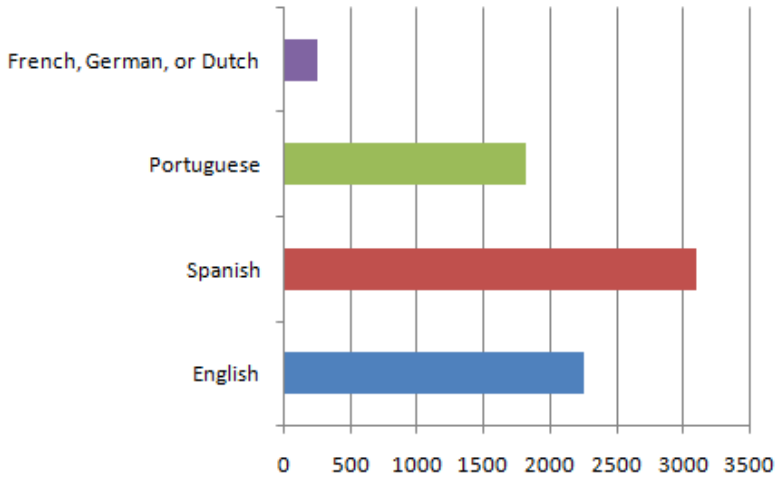


Figure 2-3. Number of contributors using a language at least once in a changeset comment.

Validating the language identification

To evaluate the performance of the language identification software and understand what percentage of these changeset comments were actually in the six target languages, some manual verification was performed. Two researchers familiar with English, Spanish, and Portuguese selected a random sample of 1033 changesets (1%) and checked each comment to verify whether the language had been correctly identified by the langid.py software. For the small number of comments outside these languages, online dictionaries and other language identification software such as Google Translate were used to assist in the verification. The researchers' coded lists were then checked against each other for discrepancies and reconciled through discussion.

Through this random check, it was determined that langid.py had detected the language correctly about 97% of the time. The languages for approximately 2% of the sampled comments were incorrectly identified (often due to the inclusion of place names originating from a different language), and a few records (less than 1%) consisted of multiple languages or the language was indistinguishable. Given these results, it was concluded that the chosen language identification software was suitable for the purpose of processing OSM changeset comments in automated fashion, keeping in mind that a small rate of error might permeate the data. The 97% success rate is higher than that achieved by any of the four language identification platforms tested by Graham et al. (2014a) with Twitter messages, although those tests were run on more complex messages with a greater variety of languages in play.

Summary of spatial patterns among languages

When the changesets are mapped by language used, the spatial patterns typically follow the dominant languages by country (Figure 2-4). Portuguese is widely used in Brazil, while Spanish is used in all other countries except the Guianas (Guyana, Suriname, and French Guiana). In the Guianas, English, Dutch, and French are present in each former colony as expected. German is scattered in pockets throughout the South American map. English, however, appears everywhere.

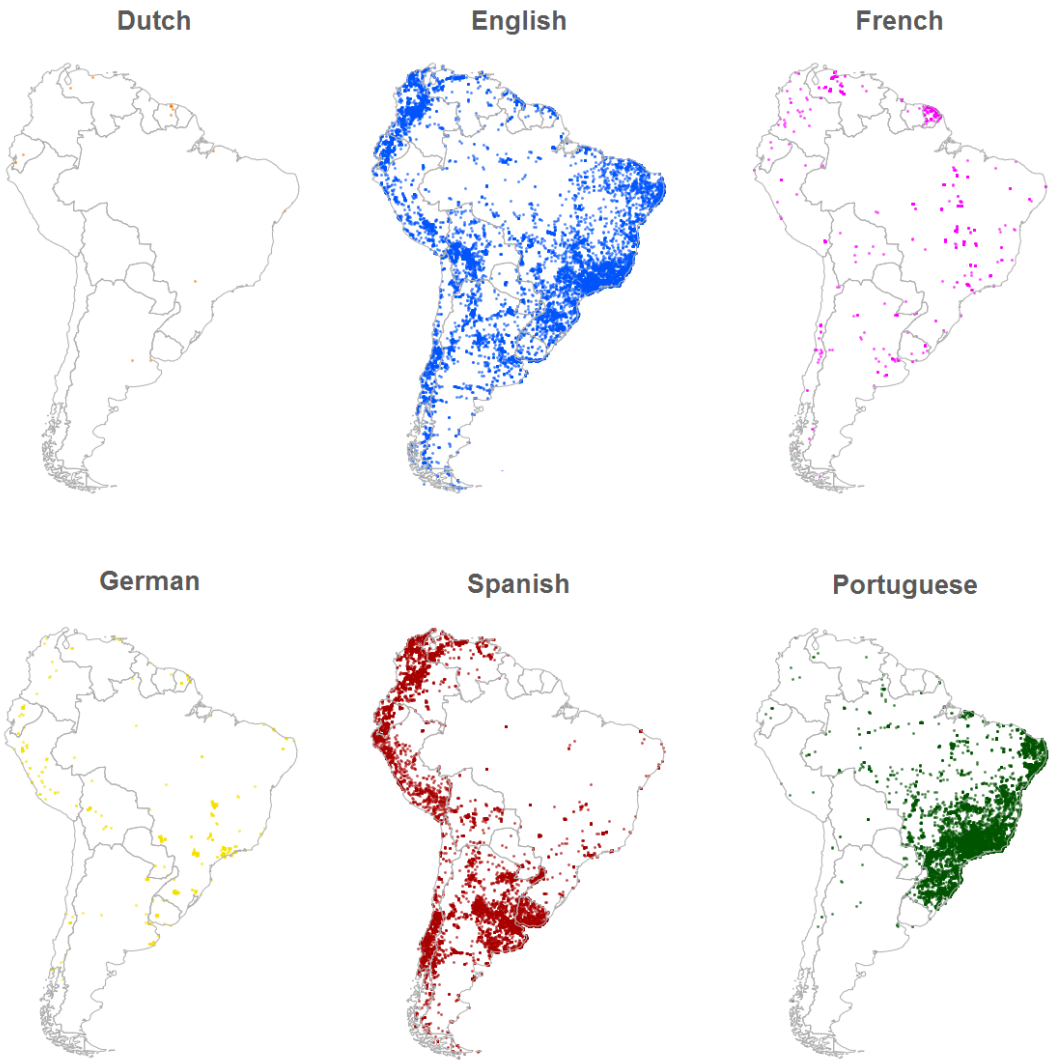


Figure 2-4. Spatial distribution of language use in changeset comments.

It should be noted that some of the contributors (11.4%) used more than one language within their group of changesets. Although it is common for OSM contributors to be multilingual, this figure should be interpreted cautiously. If an active monolingual contributor with many changesets has at least one comment miscoded by the language identification software, he or she

incorrectly appears as multilingual in the analysis. These cases become complex to detect because the most active contributors can create hundreds of changesets, inviting some miscodings even with the 97% success rate observed. Furthermore, heavy contributors mapping a single type of entity sometimes apply the same comment on multiple sequential changesets, thereby amplifying the effect of a miscoding. Further work is needed to develop methods for estimating the percentage of OSM contributors who are truly multilingual.

Evaluating languages against confirmed locations from user wiki and profile pages

If we see a comment in a local language, to what degree of confidence can we infer that the contributor was local in origin? After all, someone might edit South America with a Spanish-language comment, while living in some location outside the continent, such as Spain or Mexico. Furthermore, the use of English does not automatically indicate a nonlocal user, as someone from South America might comment in English with the motive that the text will be understood by more OSM contributors throughout the world. How often do these types of situations actually occur, and what percentage of local users are associated with each language? This question has not been studied in previous literature, and it is necessary to address in order to appropriately interpret the findings of the automated language identification.

As suggested previously, self-reported home locations on OSM user profile and wiki pages can provide a reference dataset of known user origins. The OSM wiki is a framework of web pages editable by anyone in the OSM community. These pages help organize the work around different themes (ie, "Food Security") or geographic regions (ie, "WikiProject Uruguay"). Importantly for our purposes, any OSM user can create a personal page on the wiki to describe his or her interests, technical background, languages, and GPS equipment. Similar, but not equivalent, are profile pages that users create on openstreetmap.org without having to learn the

wiki syntax. Both of these types of pages frequently contain autobiographical text such as "My name is Linda and I'm from Florida". By connecting these profile clues with the language(s) that a contributor is known to employ in his or her comments, I estimated the percentage of the contributors for each language originating from South America using the following methods.

The wiki and profile pages are both available in a known URL format in which the OSM user name is inserted at the end. Using an automated script, all the user names from the South American study dataset of changesets were inserted into the URLs and a web request was made for the HTML of the user's profile and wiki page. For many contributors no web page was returned, meaning that the contributor had not created a profile or wiki page; however, any request that returned a page was manually opened and scrutinized for direct indications of the user's home country. The following types of information were considered acceptable indicators of a person's country of origin: location reported by the contributor directly in the text of the page, location of school or place of employment mentioned on the page, location mentioned in personal web page linked from the page, predominant location mentioned in the page text or diary (blog) where onsite mapping had occurred, or location of OSM user groups the contributor belonged to (eg, "Users in Germany") when such groups were not in logical geographic conflict with each other (some contributors join groups in every country where they have mapped).

This analysis resulted in 567 contributors (8.7% of the total contributors in the study dataset) whose changeset comment languages could be linked with their places of residence at least at the country level. Perhaps not surprisingly, these contributors are more active in OSM than the typical contributor; they accounted for 37,148 changesets (36.0% of the total changesets in the study dataset) with a median of 4 changesets per contributor (as compared with a median of 2 changesets per contributor for the entire study dataset). The percentage of OSM users revealing their location in unstructured profile text and their collective levels of contribution to the project are similar to those found for Wikipedia by Graham et al. (2015).

The percentages of contributors residing inside and outside of South America were then calculated for each language. For example, 225 contributors who wrote comments in Spanish also revealed some location information in their profile or wiki pages. Of these, 86.7% were from South America and 13.3% were from somewhere outside South America (primarily Europe and the United States). Nearly equivalent percentages to this were observed for Portuguese users; however, English was different: out of the 303 contributors who wrote messages in English, 34.7% were from South America[5] and 65.3% were from outside South America. The results for each language are summarized in Figure 2-5. Note that the number of contributors listed in Figure 2-5 exceeds the total of 567 unique contributors because more than one language was detected for some contributors.

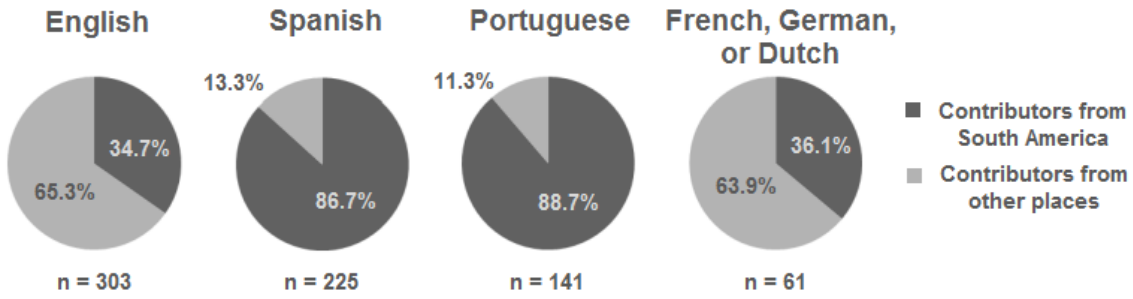


Figure 2-5. Locations of contributors maintaining a profile or wiki page where a location could be detected, grouped by language used in changeset comments

Contributors identified as being from South America tended to add more to the map than contributors from elsewhere. This trend can be seen in the graph of median changesets per user (Figure 2-6) for each language and location group. Portuguese and Spanish-using contributors from South America have the highest median number of changesets, with the especially high activity rates of Portuguese-speaking contributors again evident from this graph. These are followed by all other groups. Although we cannot conclude for certain that these findings

[5] Only one of these contributors was from Guyana, where English is commonly spoken.

extrapolate to the full set of OSM contributors, they do suggest that most of the OSM representation of South America is being built by people who live there and is not dominated by armchair mappers from overseas.The distribution of nonlocal influence varies from place to place, however, as explored in the next section.

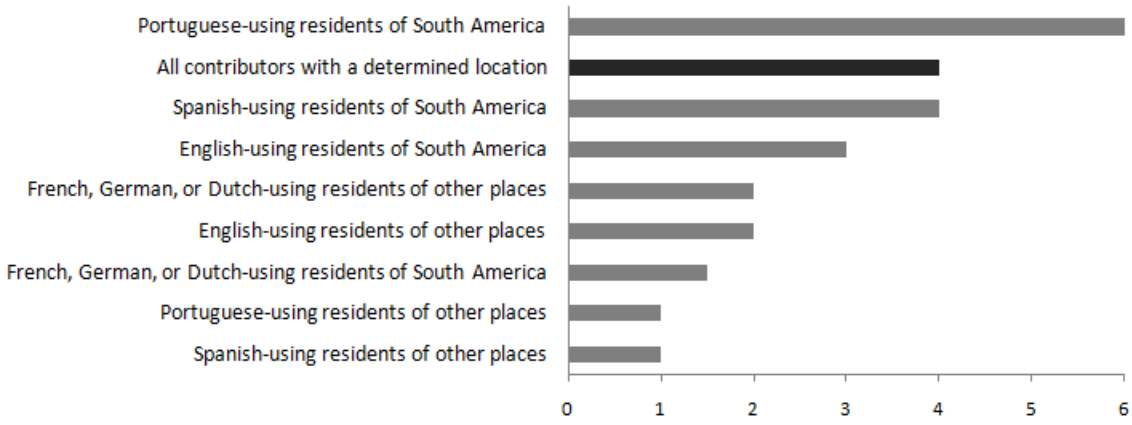


Figure 2-6. Median changesets per contributor when considering language used and determined location

Geographies of English usage

The portion of English-language comments in any particular place is of special interest because of the relatively high percentage of nonlocal contributors associated with English during the profile check. Although we cannot confidently infer a place of origin for any one English speaking user, the above methods suggest that the group of English speaking contributors is more likely to have a greater composition of nonlocal people than a group of Spanish or Portuguese speaking contributors. It follows that mapping the distribution of English-commented changesets could reveal the areas that are experiencing the most nonlocal influence in OSM.

English appears in changeset comments in all parts of South America; however, mapping English changesets as simple dots is not sufficient to understand the presence of English compared to other languages. To further investigate the use of English, bins of roughly equal area

were used to map the dominant language in each (Figure 2-7).[6] The bins are assigned a color based on which language appeared in the largest number of changesets. The bin boundaries are not visible here; instead, proportional symbols at the centroid of the bin provide a relative indication of how many total changesets are present in the bin. Thus, the areas along the highly populated central Atlantic coast of the continent have many changesets , whereas areas in the Amazon have very few or none. Throughout the map, the symbols are drawn with a logarithmic scale to avoid disruptively large and small symbols.

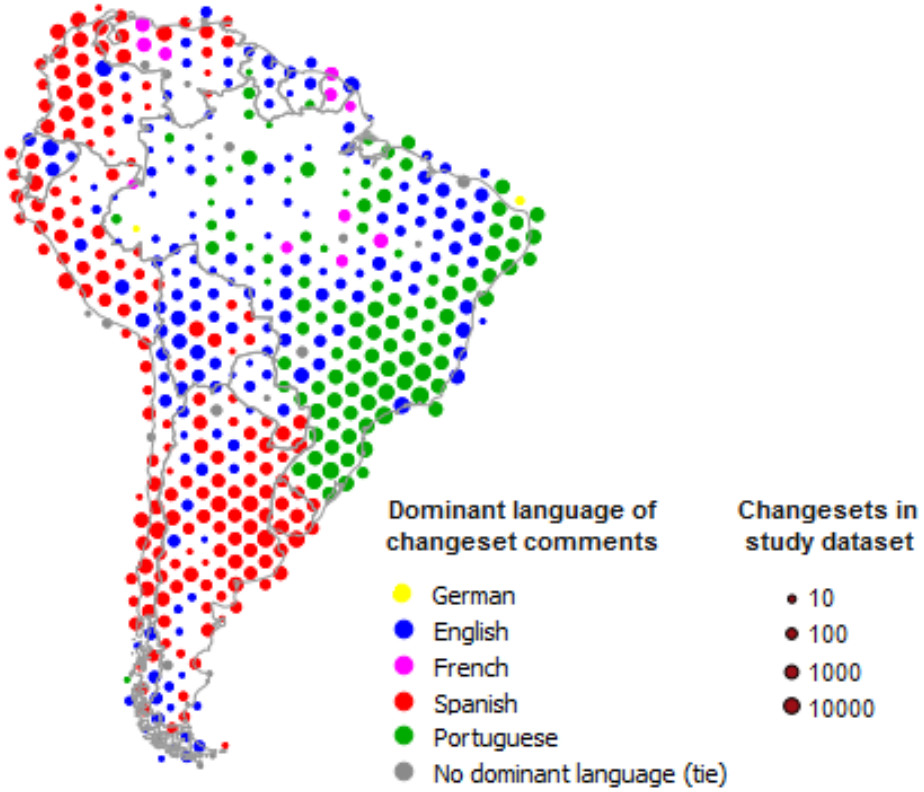


Figure 2-7. Dominant language of OSM contributor comments.

[6] All maps in this study use the changeset centroid to represent the changeset location. Because the studied changesets were limited to 0.5 degrees in width and height, the centroid is expected to be near the actual area of the changes.

In Figure 2-7 a belt of English dominance can be observed running from the northeast coast of Brazil in a southwesterly direction through Paraguay, Bolivia, and Chile. This region further stands out when the percentage of English-commented changesets is mapped with a graduated color scheme (Figure 2-8). Many variables could be tested to determine what, if any, local phenomena affect this overall pattern. In the interest of brevity only two will be analyzed here, economic prosperity and the urban-rural divide.

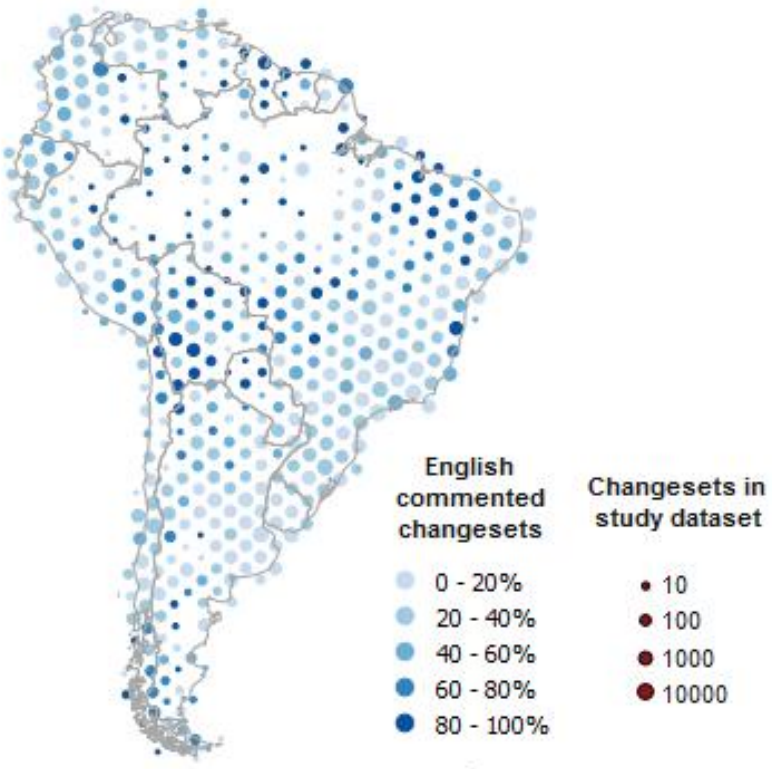


Figure 2-8. Percent of OSM changesets commented in English.

At a glance, the areas high in English language usage appear to cover some of the less wealthy parts of the continent, as well as interior areas far from urban centers. To examine how much this is really the case, I started at the country level scale and examined 2014 estimated gross

domestic product (GDP) at purchasing power parity (PPP) per capita (International Monetary Fund, 2014) for all Spanish and Portuguese-speaking countries in South America. This was compared with the percentage of OSM changesets in the study dataset commented in English within each respective country (Figure 2-9).[7] [8] A significant negative correlation appears between percentage of changesets commented in English and GDP PPP per capita ($r(8) = -0.787$, $p = 0.007$).

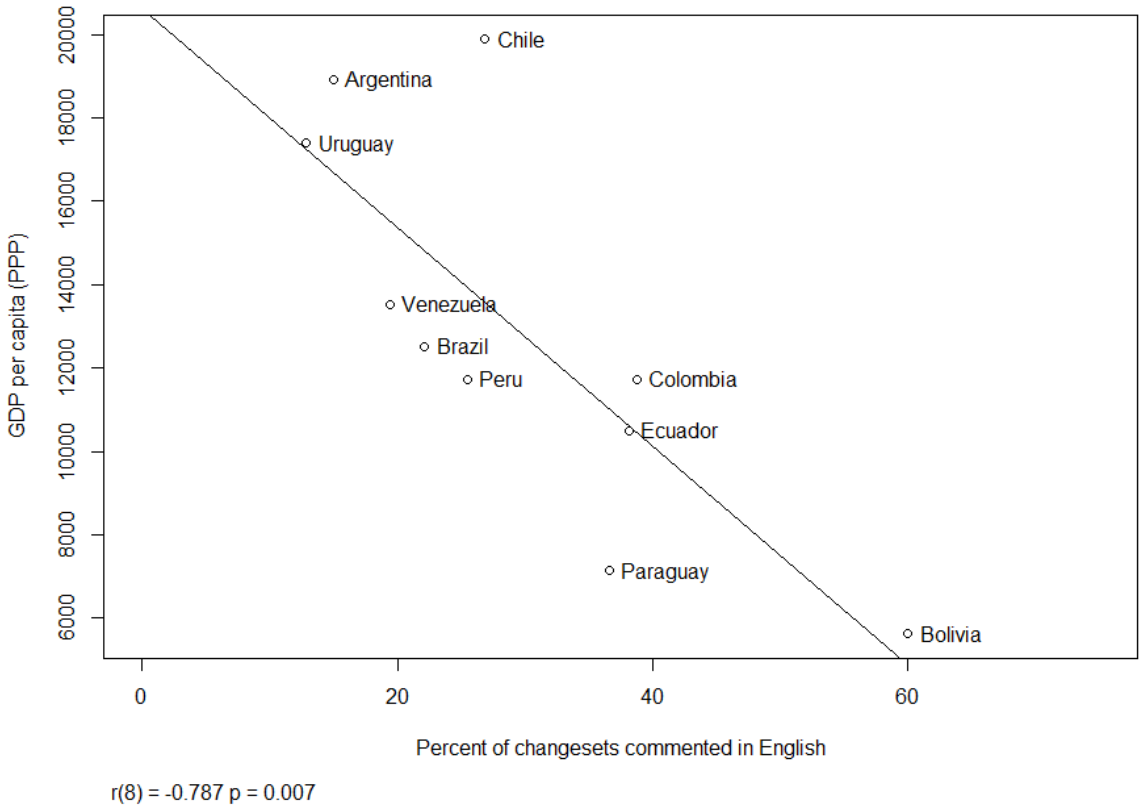


Figure 2-9. GDP (PPP) per capita plotted against percent of OSM changesets commented in English.

[7] The Guianas were excluded from this analysis due the higher percentage of English-speaking population in some parts of this region and the unavailability of consistent statistics for French Guiana.

[8] Although the OSM study dataset covers a period of five years and the economic data is from 2014 only, South American countries see almost no shift when ranked against each other by GDP PPP per capita between 2009 and 2014 (only Columbia and Peru switch places); therefore, for simplicity it was decided to use only the 2014 figures.

Determining whether a relationship between lack of economic prosperity and high English usage holds true at finer-grained jurisdictions is more challenging, as the numbers of OSM contributors at the state and provincial levels are smaller and more varied. At these scales a single very active user can more substantially affect the data, and there is not enough data to reasonably discern year by year trends. Nevertheless, I attempted some analysis in Brazil and Argentina, two countries where OSM has seen a relatively high number of users.

First, 2014 per capita monthly income was taken for all the states of Brazil and compared with the percentage of English-commented changesets in each state (income data from IBGE 2015). As expected given the continent-level test, a negative correlation (albeit weaker and not significant at the 0.05 level) between income and English usage was observed ($r(25) = -0.281$ $p = 0.155$).

Income-related measures are relatively easy to derive and understand across scales, but as a measure of poverty they have some shortfalls, emphasizing economic development over the availability of basic human needs (Yapa 1992, 2015). Therefore, in an effort to consider the influence of poverty in a manner more directly associated with day-to-day necessities of life, I examined the percent of the population in Argentina lacking at least one basic human need, as measured by the Argentine National Institute of Statistics and Census (INDEC 2014). This organization periodically measures the number of people and households whose living conditions fall short of a detailed criteria of needs that includes dwelling quality, sanitary conditions, persons per room, school attendance of children, and capacity of the head of household to provide for the occupants. The most recent available figures are from 2010. The percentage of Argentine residents lacking at least one of these necessities at the provincial level was correlated with the percentage of OSM changesets commented in English, but the correlation is even weaker than that observed with the Brazil analysis ($r(22) = 0.155$ $p = 0.471$). The number of changesets in

Argentine provinces is generally lower than in Brazilian states, and the variation involved makes it unlikely that a strong correlation could be observed at this point, although the picture may become clearer in the years to come as OSM attracts more contributors. Exact figures from all the above tests are listed in tables in Appendix A.

Considering all of these analyses with the previous findings that most English-using contributors in South America are coming from outside the continent, there is some indication that poor places in South America have a lesser degree of local ownership over their maps than affluent places. Further research into this issue is encouraged in order to understand other variables that might be in play and whether the correlations hold at a more local level as OSM increases its volume of data in future years.

Another trend that might be inferred by looking at the maps of languages usage is that English-commented changesets seem to be more common in rural areas. To determine how much this is really the case, the urban and rural dimensions of English usage in OSM were examined by obtaining vector polygons of built-up land use areas from Natural Earth (Natural Earth 2015). The methodology for creating these polygons is described by Schneider et al. (2003). In the built-up areas of South America, 19.4% of changesets were commented in English, whereas in non-built-up areas, 29% of changesets were commented in English. Given the large number of changesets involved, a two-sample t-test found these differences in percentage significant with $p < 0.001$. These findings seem to indicate that nonlocal influence in the South American OSM is higher in rural places than in cities, but that even in rural places the majority of mappers are probably still local in origin.

What do the users of the different languages prioritize?

Does it even matter when large percentages of people are mapping a place from overseas, and if it does, how are the distant contributors mapping differently than the local contributors? To address this question, some further analysis was performed examining which OSM tags are prioritized by English, Spanish, and Portuguese speaking contributors. To maintain focus on countries with mostly Spanish or Portuguese speaking populations, all changesets in the Guianas (Guyana, Suriname, and French Guiana) were excluded from the following part of the analysis.

An automated script was used to tally and rank the frequency of all OSM tags used in connection with the three above-mentioned languages. A tag consists of metadata in key-value form describing some physical entity in the world. For example, in OSM, "amenity" is a key used for denoting a broad assortment of community facilities. Thus, the tag amenity=school is used to denote a school; *amenity* is the key and *school* is the value. In this analysis I focused particularly on four of the most commonly used OSM keys: amenity, highway, shop, and leisure, wanting to see if the users of different languages prioritized differently the values given to these keys.

To understand how the users of the different languages prioritized the different tags, the following procedure was repeated for each key: First, a list was created of all values that ranked within the top 10 of any of the three languages. The rank of each value in English was then compared with the average ranking of the value in Spanish and the value in Portuguese, and then the difference was recorded. This conveyed which values English speaking contributors tended to favor when compared with speakers of the local languages.

The results show that English, Spanish, and Portuguese speakers favor many of the same tags. However, some of the tags favored by speakers of the local languages (Spanish and Portuguese) tend to have a closer connection to providing basic day-to-day needs such as health, transportation, fuel, recreation, and neighborhood "corner store" purchases. In contrast, tags

markedly favored by English speakers are more connected to sites of tourism and consumerism, as well as objects that can be traced from aerial photography.

For example, after ranking all values given to the amenity key, amenity=hospital ranked an average of four places higher for Spanish and Portuguese speaking contributors than for English speaking contributors. amenity=school ranked two places higher and amenity=fuel ranked 1.5 places higher. In comparison, English speakers favored amenity=bench (a difference of 11.5 places in rank), amenity=telephone (10.5 places), amenity=restaurant (2 places), and amenity = fast_food (1.5 places).

When highways were considered, the tag highway=road (5 places) was the highest ranked for English speakers when compared to the average tag ranks for the Spanish and Portuguese speaking contributors. The value "road" is often used when the highway is traced from aerial imagery and the tracer does not have enough in situ evidence to assign a more specific classification. In contrast, highway=bus_stop (7 places) was most favored by Spanish and Portuguese speaking contributors when compared with English speaking contributors. Bus stops are not easily visible from the air, and in some parts of Latin America bus stops are not even marked on the ground, although their locations are widely known by local residents and are critical for reaching places of employment and other necessary errands. They are examples of features that satisfy the needs of everyday residential life that are more likely to be supplied by local mappers.

The full results of this analysis in Figure 2-10 show similar patterns. English speaking contributors placed higher priority on mapping bicycle shops, shopping malls, and auto dealerships, while Spanish and Portuguese speaking contributors tended to map smaller neighborhood businesses such as butcher shops, general stores (shop=yes), and kiosks. When analyzing places of leisure, English speakers marked gardens, stadiums, and tracks while Spanish and Portuguese speakers placed higher priority on marking sports complexes, nature reserves, and

playgrounds. The difference here may be related to which places of leisure are more easily discerned from interpreting aerial photographs, assuming there is a greater percentage of armchair mappers among the English speaking contributors. For example, most OSM contributors can identify a track or a stadium in a photograph, while it is harder to identify playground equipment or know which areas constitute protected natural space without actually visiting a site. Also, tourists seem more likely to visit (or remember) landmark destinations such as stadiums and shopping malls, rather than routine destinations such as playgrounds and butcher shops.

Amenities (amenity=<value>)

amenity value	English rank	Spanish rank	Portuguese rank	English difference
bench	10	31	12	**11.5**
telephone	6	25	8	**10.5**
restaurant	2	3	5	**2**
fast_food	8	10	9	**1.5**
parking	1	2	2	**1**
place_of_worship	4	5	4	**0.5**
bank	7	7	6	**-0.5**
university	11	8	13	**-0.5**
pharmacy	9	6	10	**-1**
fuel	5	4	3	**-1.5**
school	3	1	1	**-2**
hospital	12	9	7	**-4**

Roads (highway=<value>)

highway value	English rank	Spanish rank	Portuguese rank	English difference
road	6	14	8	**5**
living_street	10	10	17	**3.5**
tertiary	2	5	2	**1.5**
unclassified	5	7	5	**1**
residential	1	1	1	**0**
service	7	8	6	**0**
footway	9	9	9	**0**
secondary	3	2	3	**-0.5**
primary	4	3	4	**-0.5**
trunk	11	11	10	**-0.5**
track	8	6	7	**-1.5**
bus_stop	15	4	12	**-7**

Shops (shop=<value>)

shop value	English rank	Spanish rank	Portuguese rank	English difference
bicycle	9	21	14	**8.5**
mall	3	6	5	**2.5**
car	6	13	4	**2.5**
hairdresser	7	10	8	**2**
clothes	4	4	6	**1**
supermarket	1	1	1	**0**
convenience	2	2	2	**0**
bakery	5	5	3	**-1**
car_repair	8	7	7	**-1**
hardware	12	12	9	**-1.5**
kiosk	10	3	13	**-2**
yes	20	9	10	**-10.5**
butcher	26	8	20	**-12**

Recreational sites (leisure=<value>)

leisure value	English rank	Spanish rank	Portuguese rank	English difference
garden	3	7	7	**4**
recreation_ground	9	8	11	**0.5**
stadium	8	9	8	**0.5**
track	10	11	10	**0.5**
park	1	1	1	**0**
pitch	2	2	2	**0**
common	4	3	5	**0**
swimming_pool	5	5	4	**-0.5**
playground	7	6	6	**-1**
nature_reserve	11	10	9	**-1.5**
sports_centre	6	4	3	**-2.5**

Figure 2-10. Rank of OSM tag values by language. Note that in OSM tagging, the value "yes" is used when more specific information is not available.

Conclusions and directions for future research

OSM is created by speakers of many languages from different parts of the world. Identifying and mapping the languages of changeset comments can provide a picture of regional

trends among contributors. Here I have evaluated the degree to which this can be accomplished using a freely available language identification software package. The cross-checking of the detected languages with locations from user profiles helps understand the amount of nonlocal participation inferred by each language.

The geographic mix of contributors affects the composition and richness of the map in any given place. In South America, most features appear to be contributed by editors from South America, rather than long-distance tourists or armchair mappers, although it is possible that this was not always the case in the early days of OSM. Mappers local to the continent have a heavy influence, but their levels of influence vary from place to place.

In this chapter I have shown that one metric of nonlocal influence (in non-Anglophone areas) is the percentage of changesets commented in the English language. Similar methods could be applied elsewhere in the Global South, such as in Africa or southeast Asia, to ascertain how nonlocal influence in the map fluctuates from place to place. Language use among OSM contributors exhibits a marked spatial variation that corresponds to a variety of phenomena. In South America, English-commented changesets (and by extension nonlocal influence) are more prevalent in rural areas than in urban locations. Also, I have explored cases where lower income and deprivation of basic needs are correlated (at varying degrees of strength) with higher levels of mapping by English-speaking contributors. Further analysis is needed to determine how these trends vary from country to country, although the relatively low number of OSM contributions and contributors makes it difficult to arrive at solid conclusions at vey local scales. Other variables could also be examined for possible associations with OSM contributor activity; for example, broadband access was found by Graham et al. (2015) to have a strong correlation with Wikipedia edits at the country level scale.

When interpreting the results of language identification and the cross-checks with user profiles, several limitations deserve mention. Within the set of users who edited OSM for South

America and revealed their locations, I showed that most English speaking contributors are not from South America, and most speakers of Spanish and Portuguese are from South America; however, conclusions about place of origin cannot be made at the individual level based solely on language use. Also, the vast majority of users have said nothing about themselves through a profile or wiki page, and we cannot be sure whether their geographic distribution would match the patterns of more active users who tended to reveal a location. A general survey asking the locations of OSM contributors might be one way to confirm this, although the set of respondents might just closely match the same active group that created profile pages. Other methods seeking to derive location clues from unstructured text (Lee et al. 2013) may hold promise in future analysis when applied to OSM profile pages, contributor comments, and OSM-hashtatgged social media posts.

From my experience reading hundreds of these biographical pages, the user profiles from within South America tend to be briefer in nature than those from the European contributors mapping the continent (who are often OSM "power users"). It is possible that the methods presented in this chapter underestimate the number of users in South America due to a lower propensity by these users to create detailed profiles or any profile at all. The degree of severity of this underestimation is unknown, but would in no way nullify the importance of beginning with the analysis presented here. Another important follow-up study would be to identify the contributors leaving no comments (or very short ones) and record their locations (where indicated in profile/wiki pages) and the types of things they tend to map.

When compared with Spanish and Portuguese speaking contributors, English speaking contributors emphasize features that can be easily traced from aerial photographs or observed in passing, such as roads and stadiums. Also, their favoring of shopping malls, restaurants, fast food outlets, and auto dealerships seems to reflect an interest in sites of tourism and consumerism. On the other hand Spanish and Portuguese speaking contributors emphasize features related to daily

routines such as taking children to school, visiting the corner store, riding the bus, or visiting the doctor. Many of these features can only be observed or verified by someone on site. These local influences make the map more valuable for the residents it serves, while reducing the empty spaces on the map where thousands of people may dwell unnoticed or uncared for in "cartographies of silence" (Harley 1988, Brunn and Wilson 2013).

The comparison of tags favored by local and nonlocal languages could be extended into many other categories of features, especially if combinations of tags are considered. One example would be a study of whether Spanish and Portuguese speakers are more likely than English speaking contributors to add a street name when marking a road. Further inquiries into the tags added by speakers of other languages would also help support or refute the assertions made here about features favored by local and nonlocal contributors. I found that there were not enough tags added by French, German, or Dutch speaking contributors in South America to warrant independent analyses of these languages; however, the tags from all languages other than Spanish or Portuguese might be combined with the English ones to see if the above results are substantively affected.

An analysis of "localness" in OSM could involve a variety of scales. Here I have used a coarse-grained binary approach at the continental level to determine if a contributor should be considered local or not. To some degree the study was forced into this scale by the small number of user profile pages available for validating the findings. Studies at the provincial or municipal level may reveal interesting differences in local vs. nonlocal contributions if more advanced language processing can be used on the contributor comments to ascertain editor locations (eg, geocoding comments such as "this is my street" vs. "I traced this from Bing imagery"). Large-scale, systematic surveys of contributors might also provide insight, although garnering enough participation seems daunting when considering Budhathoki's (2010, p.66-67) blanket survey of

OSM contributors that saw no respondents from South America (the next fewest respondents from an inhabited continent was 16).

Potential users of crowdsourced VGI such as OSM should look beyond the map image and consider the set of contributors that created the data and how the end product might have been affected by them. This is particularly important in low-income regions, places lacking robust Internet infrastructure, and among peoples where online participation is otherwise low or suppressed. Further work is needed to compare the motives of local and nonlocal mappers, and investigate the ways that these types of contributors might be fostered (and retained) in areas of the world where their particular contributions are needed.

Chapter 2 references

Brunn, Stanley D., and Matthew W. Wilson. 2013. “Cape Town’s Million plus Black Township of Khayelitsha: Terrae Incognitae and the Geographies and Cartographies of Silence.” *Habitat International* 39: 284–94.

Crampton, Jeremy W. 2009. “Cartography: Maps 2.0.” *Progress in Human Geography* 33 (1): 91–100.

Elwood, Sarah. 2008. “Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS.” *GeoJournal* 72 (3-4): 173–83.

Girres, Jean-Fran\ccois, and Guillaume Touya. 2010. “Quality Assessment of the French OpenStreetMap Dataset.” *Transactions in GIS* 14 (4): 435–59.

Goodchild, Michael F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211–21.

Graham, Mark, Scott A. Hale, and Devin Gaffney. 2014a. “Where in the World Are You? Geolocation and Language Identification in Twitter.” *The Professional Geographer* 66 (4): 568–78. doi:10.1080/00330124.2014.907699.

Graham, Mark, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. 2014b. “Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty.” *Annals of the Association of American Geographers* 104 (4): 746–64.

Graham, Mark, Ralph K. Straumann, and Bernie Hogan. 2015. “Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia.” *Annals of the Association of American Geographers* 0 (0): 1–21. doi:10.1080/00045608.2015.1072791.

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

Graham, Mark, Matthew Zook, and Andrew Boulton. 2013. “Augmented Reality in Urban Places: Contested Content and the Duplicity of Code.” *Transactions of the Institute of British Geographers* 38 (3): 464–79.

Hagen, Erica. 2010. “Putting Nairobi’s Slums on the Map.” *Development Outreach | World Bank Institute*, July, 41–43.

Haklay, Mordechai. 2010. “How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets.” *Environment and Planning. B, Planning & Design* 37 (4): 682.

———. 2013. “Neogeography and the Delusion of Democratisation.” *Environment and Planning A* 45 (1): 55–69. doi:10.1068/a45184.

Harley, J. Brian. 1988. “Silences and Secrecy: The Hidden Agenda of Cartography in Early Modern Europe.” *Imago Mundi* 40 (1): 57–76.

———. 1990. “Cartography, Ethics and Social Theory.” *Cartographica: The International Journal for Geographic Information and Geovisualization* 27 (2): 1–23.

IBGE (Instituto Brasileiro de Geografia e Estatistica). 2015. “IBGE divulga rendimento domiciliar per capita segundo a PNAD Contínua para o FPE.” http://saladeimprensa.ibge.gov.br/noticias?view=noticia&id=1&busca=1&idnoticia=2833.

INDEC (Instituto Nacional de Estadística y Censos) - Argentina. 2014. “Indicadores sociodemográficos - Condiciones de vida.” http://www.indec.mecon.ar/indicadores-sociodemograficos.asp.

International Monetary Fund. 2014. “World Economic Outlook Database.” http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/index.aspx.

Latif, Sufian, KM Rakibul Islam, Md Monjurul Islam Khan, and Syed Ishtiaque Ahmed. 2011. “OpenStreetMap for the Disaster Management in Bangladesh.” In *Open Systems (ICOS), 2011 IEEE Conference on*, 429–33. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6079240.

Lee, Kisung, Raman Ganti, Mudhakar Srivatsa, and Prasant Mohapatra. 2013. “Spatio-Temporal Provenance: Identifying Location Information from Unstructured Text.” In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, 499–504. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6529548.

Leszczynski, Agnieszka, and Matthew W. Wilson. 2013. “Guest Editorial: Theorizing the Geoweb.” *GeoJournal* 78 (6): 915–19.

Lui, Marco, and Timothy Baldwin. 2011. “Cross-Domain Feature Selection for Language Identification.” In *In Proceedings of 5th International Joint Conference on Natural Language Processing*, 553–61.

———. 2012. “Langid. Py: An off-the-Shelf Language Identification Tool.” In *Proceedings of the ACL 2012 System Demonstrations*, 25–30. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2390475.

Maron, Mikel, Grant Slater, and Steve Coast. 2012. “Google IP Vandalizing OpenStreetMap | OpenStreetMap Blog.” January 17. https://blog.openstreetmap.org/2012/01/17/google-ip-vandalizing-openstreetmap/.

Natural Earth. 2015. *Urban Areas*. http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-urban-area/.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf. 2011. “The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011.” *Future Internet* 4 (1): 1–21.

———. 2013. “Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions.” *Future Internet* 5 (2): 282–300.

Neis, Pascal, and Alexander Zipf. 2012. “Analyzing the Contributor Activity of a Volunteered Geographic Information project—The Case of OpenStreetMap.” *ISPRS International Journal of Geo-Information* 1 (2): 146–65.

Quest, Christian. 2014. “OSM Quality Assurance Thru Cross Checking Statistics and External Datasets.” In *State of the Map 2014*. Buenos Aires, Argentina. http://vimeo.com/album/3134207/video/112233941.

Schneider, A., M. A. Fried, McIver, D. K., and C. Woodcock. 2003. “Mapping Urban Areas by Fusing Multiple Sources of Coarse Resolution Remotely Sensed Data.” *Photogrammetric Engineering and Remote Sensing* 69: 1377–86.

Stephens, Monica. 2013. “Gender and the Geoweb: Divisions in the Production of User-Generated Cartographic Information.” *GeoJournal* 78 (6): 1–16.

Wood, Harry. 2012. “Licence Redaction Ready to Begin | OpenStreetMap Blog.” July 9. https://blog.openstreetmap.org/2012/07/09/licence-redaction-ready/.

Wright, Alyssa. 2013. “Changing the Ratio of OpenStreetMap Communities.” presented at the State of the Map 2013, Birmingham, England. http://lanyrd.com/2013/sotm/scphhf/.

Yapa, Lakshman. 2015. “Why We Cannot All Be Middle Class in America.” In *Routledge Handbook on Poverty and the United States*, edited by Stephen Haymes, Maria Vidal de Haymes, and Reuben Miller, 576–83. Routledge.

———. 1992. “Why Do They Map GNP per Capita.” *Majumdar, SK, Forbes, GS, Miller, EW, Schmalz, RF (Eds.), Natural and Technological Disasters: Causes, Effects, and Preventive Measures. The Pennsylvania Academy of Science, Easton, MA*, 495–510.

Yasseri, Taha, Robert Sumi, and János Kertész. 2012. “Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis.” *PloS One* 7 (1): e30091.

Zook, Matthew A., and Mark Graham. 2007. “The Creative Reconstruction of the Internet: Google and the Privatization of Cyberspace and DigiPlace.” *Geoforum* 38 (6): 1322–43.

Zook, Matthew, Mark Graham, Taylor Shelton, and Sean Gorman. 2010. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake.” *World Medical & Health Policy* 2 (2): 7–33.

Chapter 3

Using small cities to understand the crowd behind OpenStreetMap[9]

Abstract

As businesses and governments integrate OpenStreetMap (OSM) into their services in ways that require comprehensive coverage, there is a need to expand research outside of major urban areas and consider the strength of the map in smaller cities. A place-specific inquiry into the OSM contributor sets in small cities allows an intimate look at user motives, locations, and editing habits that are readily described in the OSM metadata and user profile pages, but often missed by aggregate studies of OSM data. Using quantitative and qualitative evidence from the OSM history of five small cities across North and South America, I show that OSM is not accumulating large local corpora of editors outside of major urban areas. In these more remote places OSM remains largely at the mercy of an unpredictable mix of casual contributions, business interests, feature-specific "hobbyists", bots, and importers, all passing through the map at different scales for different reasons. I present a typology of roles played by contributors as they expand and fix OSM in casual, systematic, and automated fashion. I argue that these roles are too complex to be conceptualized with the traditional "citizen as sensor" model of understanding volunteered geographic information. While some contributors are driven by pride of place, others are more interested in improving map quality or ensuring certain feature types are

[9] This chapter contains a slightly edited version of a paper originally published as:
Quinn, Sterling. 2015. "Using Small Cities to Understand the Crowd behind OpenStreetMap." GeoJournal. doi:10.1007/s10708-015-9695-6.
The original publication is available at Springer via http://dx.doi.org/10.1007/s10708-015-9695-6.

represented. Institutions considering the use of OSM data in their projects should be aware of these varied influences and their potential effects on the data.

Introduction

In the past decade, the growth of a more interactive and mobile-friendly World Wide Web has allowed people with little formal technical training to produce enormous amounts of online information. The practice of "crowdsourcing" harnesses the work of these diverse and distributed Internet users to collaboratively produce information products, such as the free online encyclopedia Wikipedia. The OpenStreetMap (OSM) project, built on a similar model of open contribution, has proven that cartographic information can also be collected through a crowdsourcing approach.

Goodchild (2007) described this type of mapping crowd as "citizens as sensors", collecting and contributing "volunteered geographic information" (VGI) about the world around them. Human sensors are not as easily managed or calibrated as traditional mechanical sensors are, and it has become clear that OSM and other broad-scale VGI projects face challenges with achieving consistent levels of coverage across space (Latif et al. 2011, Neis et al. 2013) and scale (Touya and Brando-Escobar 2013). Contributors bring individual interests, biases, motivations, business objectives, and life experiences to crowdsourced maps that affect the nature and focus of the data (Budhathoki 2010, Stephens 2013, Glasze and Perkins 2015). The fact that VGI is often created in loosely organized and informal environments can cause hesitance in adoption by governments and other institutions who might otherwise find the data valuable (Feick and Roche 2013, Johnson and Sieber 2013), and institutions that do adopt the data may have a limited awareness of its potential shortfalls.

A multitude of studies have been undertaken to assess the positional or semantic accuracy of OSM, but somewhat less common are forays into the rich social influences driving the construction of the data. Coleman et al. (2009) notes that these inquiries are needed with VGI in general. Although Haklay (2014) observes that the terms VGI and OSM should not be conflated, he similarly advocates studying societal aspects and social influences in OSM as pillars of a proposed "OpenStreetMap studies" genre of research. I address this need in the present study by dissecting the size, composition, and priorities of "the crowd" behind OSM, focusing on small cities that might not attract worldwide interest. In many cases, these cities lack locally organized user groups and have fewer institutions where people with existing digital mapping skills might be found, such as colleges or technical firms. Although OSM has amassed an impressive number of streets, local landmarks, and businesses in major urban areas such as Seattle, New York, and Buenos Aires, studying the data and its contributors in smaller cities may provide a more practical picture of the current usefulness of OSM for applications that require comprehensive geographic coverage.

OSM is a social construction, and when we look at OSM in a place, we see the product of a set of choices and priorities by a finite number of individuals regarding how to represent the world in a digital database. What, then, do the intrinsic characteristics of the data tell us about the size and interests of the "crowd" behind OSM in any given place? In response to this question I will explore the following aspects of the crowd in small cities:

- **How big is the crowd in these places?** – How many contributors are active in these cities, and how has the number waxed and waned over time compared to larger cities? This can help form a picture of what level of variety is represented and the amount of scrutiny that the data has received.

- **How is the work distributed among the crowd?** – Given that most OSM contributors make very few edits to the project (Neis and Zipf 2012, Wood 2014) how can we go beyond the raw number of contributors to determine the size of the active crowd in these cities?
- **Where is the crowd coming from?** – Do most of the contributors have local access to the area for onsite data collection, or are they external "armchair mappers" tracing aerial imagery, importing data, and tidying up attributes? The in situ data gathered by local mappers takes more effort to produce and requires a physical presence in or near the mapped area; however, it provides a richness to the map connected to the routines of day-to-day resident life in way that is difficult to achieve through armchair mapping (Quinn in press).
- **What brings the crowd to this place?** – Why have people edited the map in these cities, especially if they appear to have never visited the area? How do these purposes affect the types of entities that they add to (or leave out of) the map?
- **How does the nature of the crowd vary across different spaces and scales?** – How does composition and focus of the crowd differ when examining small cities from different world regions, or a larger city in the same geographic region?

Many of the above questions can intersect, leading to new dimensions of understanding about how the project is constructed by the crowd. For example, in a given location, are the heaviest contributors the ones with local access? And for a given place, how can we describe the growth (or stagnation) of the set of local OSM contributors working on the project on a regular basis? The answers can help indicate whether OSM will always be an unbalanced map across urban and rural divides.

This study focuses on three small cities in the 40,000-60,000 population range from the Pampas region of South America. The cities lie in three different provinces/departments in

Argentina and Uruguay. This region was selected because the few academic studies on OSM outside of the Global North have tended to focus on very large cities. To explore questions about how the results compare with other regions and scales, I also compared data from the three above small cities with those of (a) a large city in the Pampas region with population over 1 million, and (b) two small cities with comparable populations (ie. 40,000-60,000 people) in the Interior Plains of North America.

The publicly available history of OSM activity for the focus cities was downloaded and scrutinized for both quantitative and qualitative clues about the nature of the crowd producing the map in these places. These include the raw number of contributions, the free-form comments they leave when saving their work, the text of the public profiles they create on the OSM website, and the languages they use in these materials. In particular, indications were sought about their motives, geographic locations, amount of content contributed, and special interests or "hobbies" (such as focusing on railroads or power lines) that brought them to edit these cities. This information was used to develop a typology of editor roles describing different contribution behaviors in the map.

Theoretical background and related work

Despite the ability of crowds to quickly produce many gigabytes of data, crowdsourced information products do not appear by magic; they are built piece by piece by human beings. The bottom-up "citizen sensor" means of gathering data presents exciting prospects for the collection of rich place-specific information that perhaps could not be gathered any other way (Goodchild 2007, Elwood et al. 2013); however, a reliance on human sensors introduces vulnerabilities for the rapidly growing set of companies such as Mapbox, CartoDB, and Telenav whose business models lean heavily on open map data, especially OSM. Services provided by these companies

such as base maps and routing often require comprehensive coverage of broad regions such as countries; yet academic studies of OSM coverage and quality typically have been restricted to large urban areas, neglecting the rural landscapes and small cities that many map users will inevitably traverse.

At the same time, researchers studying a variety of OSM-related topics have indicated that the crowd of volunteer mappers remains scarce in many places. After attempting an examination of collaborative editing patterns in OSM, Mooney and Corcoran (2013) concluded: "The idea that there is a large crowd of contributors working together to gather geographic data and build the OSM database is inaccurate." Lin (2011) found OSM participants remarking that often their hometowns were mapped by just a few individuals. And in an analysis of OSM use for disaster response, Zook et al. (2010) suggested that the small and homogenous group participating in crowdsourced mapping calls into question claims that Web 2.0 has sparked real democratic revitalization of digital content production.

The imbalance of attention to certain places and features in OSM was not unanticipated. Near the beginning of the OSM project, founder Steve Coast was asked if any places would get missed or neglected in the map. His response that "no one wants to do council estates" but that nowhere else would get missed (GIS Pro 2007) was investigated quantitatively by Haklay (2010). After showing that OSM coverage was indeed negatively correlated with socioeconomic deprivation in his UK study areas, Haklay rendered the verdict that "OSM is not an inclusive project, shunning socially marginal places (and thus people)." Stephens (2013) added an examination of the ways that the imbalanced gender makeup of OSM editors has influenced map content and community discussions about which entities on the landscape should be incorporated into the OSM ontology of features.

These examples are manifestations of the "virtual black holes" and uneven geographies of information production that Graham (2010) warned could appear in VGI due to technological,

economic, or other cultural barriers faced by people who might otherwise be able to share deep knowledge of a place. Projects such as OSM are consequently constructed by a smaller set of individuals who have the time, inclination, and education necessary to edit the map (Graham et al. 2013), thereby wielding a disproportionate influence in the creation of spatial information. Patterns of digital map construction thus reflect Harley's (1989) observations of cartographic power throughout history, that "to those who have strength in the world shall be added strength in the map".

Why study small cities?

One purpose of the present study is to learn more about how small cities are imparted "strength in the map" in OSM. When we look at a small city in OSM, how likely is it that someone from that city contributed to the map, and is it reasonable to expect that OSM is gaining enough uptake in small cities to be maintained by local contributions over the long term? For all the other people contributing data in this city, what brings them to map places they have never visited where there is otherwise relatively little outside interest?

Small cities were chosen for study for three main reasons. First, most available literature on OSM has focused on either large cities or broader geographic regions such as countries. National-level studies may blunt or hide OSM construction processes occurring (or not occurring!) in smaller domains. Second, maps and apps that require full geographic coverage of an area for navigation, natural hazard planning, etc. need to provide that coverage in small cities just as well as large ones. Finally, with small cities it was conjectured that the entire corpus of edits and editors in a place would be manageable enough that it could be comprehended by a human analyst and studied qualitatively at a deeper level than other studies that have focused on tallying raw numbers of features in the database.

Small cities play an indispensable role in regional economic networks and are home to a notable portion of urban dwellers, even though they are often neglected in academic urban studies (Jayne and Bell 2009). For example, at the time of the 2010 census, 8.8% of US residents lived in "micropolitan statistical areas" centered on urban cores of 10,000 to 50,000 inhabitants and including towns in the immediately surrounding area (United States Census Bureau 2010, 2014). The 2001 Argentine national census found 7.4% of residents living in urban areas containing between 50,000 to 100,000 inhabitants (INDEC 2001a, 2001b). These cities often compete with each other and their larger neighbors to attract businesses, residents, and other amenities, using web presence as a tool (Urban 2002, Grodach 2009).

In a world where our day-to-day decisions are increasingly determined by the results of digital search algorithms and electronic placemarks (Graham et al. 2013, Graham et al. 2014), the presence of a comprehensive online map could spell an advantage for one city over another. If relatively few mappers are involved in the construction of small cities in OSM, it seems reasonable to speculate that mapping activity in these cities could be highly variable. Another potential consequence is that smaller cities remain undermapped when compared to large neighboring cities. Small city residents who use OSM-based services for search, navigation, and other functions may find that their needs are not met as fully as those of their metropolitan neighbors. Perhaps a greater challenge is that amenities and services in their cities languish invisible in the map and its search algorithms, thereby losing patronage to more visible services in competing cities, a digital realization of Harley's (1988) "cartographies of silence" (see also Brunn and Wilson 2013).

Drawing from theories of critical cartography, Perkins (2011) suggests that OSM, like many other mapping products, derives its authority from some level of denial of subjectivity, but that these subjectivities could be easily uncovered in the data. Are these subjectivities starker in cities with a smaller contributor base? "Linus's Law", popularized in the open source technology

world by Raymond (1999), and confirmed to some extent for OSM by Haklay et al. (2010) proclaims that "given enough eyeballs, all bugs are shallow", inferring that a large number of contributors will filter out low quality or unhelpful content. The potential smallness of the OSM contributor crowd in particular places is therefore a subject of special concern. Although the OSM blog reported that the project recently surpassed 2 million registered users (https://blog.openstreetmap.org/2015/03/12/two-million-contributors/), most of these are not actively working on the map. Geographically, OSM contributors are centered in Europe and major urban areas, although there is great variation in contributor activity even among major world cities (Neis and Zipf 2012, Neis et al. 2013).

Small cities as a lab for studying contributor motivation and behavior

VGI in general involves an intense focus on the local and the individual (Goodchild, quoted in Wilson and Graham 2013), yet relatively few studies of OSM have looked to intrinsic aspects of the data to discern the locations and nuances of the contributors beyond the positional and semantic accuracy of their collective edits (Glasze and Perkins 2015). The general motivations and practices of OSM contributors have been studied a bit more thoroughly, using the varying approaches of mass surveys (Budhathoki 2010), personal interviews (Lin 2011), and computational analysis of edit types (Steinmann et al. 2013).

For example, by comparing survey results with OSM contribution statistics, Budhathoki (2010, p. 84) identified a desire to share local knowledge as the motivational factor most closely linked to consistent OSM contribution. This motivation was significantly associated with the number of nodes added, frequency of contribution, and longevity of participation. In this same study, a desire to achieve the goals of the OSM project was significantly related to nodes and frequency, but not longevity. Thus, pride of place seems to be a more salient motivator than pride

of the map, although both are significantly related to at least several aspects of OSM participation.

Budhathoki suggested that these results could lead to a higher level of map detail in places where there are more human sensors, such as big cities and tourist sites, and could cause the potential for underrepresentation of towns or rural areas. Indeed, Zielstra and Zipf (2010) indicated that in Germany, the completeness of the OSM road network decreased outside of large cities to the point of becoming much less usable in rural areas, while in a study of Greater London, Mashhadi et al. (2015) confirmed that the completeness of OSM points of interest was positively correlated with population density.

Taking a more qualitative approach to understanding OSM contributor motivations, Lin (2011) interviewed contributors and identified four social worlds involved in the production of OSM data: (1) business, (2) government, (3) NGO/Third Sector, and (4) individual contributions. This approach acknowledges the influence of a diverse set of stakeholders in the project, whose organizational goals sometimes transcend large groups of employed contributors. Particularly when there is a business interest at stake, pride of place may take a back seat to interests of making sure the map data is comprehensive. In this context, the crowd behind OSM is more complicated than a set of individual sensors contributing their own local knowledges.

The present study makes at least three contributions to the above literature on OSM contributor motivations and actions: First, it draws on the vast body of OSM contributor comments embedded in the project metadata. These comments link qualitative observations by contributors with a specific set of edited features. Second, it uses the text of publicly available profile pages to understand contributor locations and backgrounds, offering a glimpse into the contributors' "social worlds" mentioned by Lin. Third, the study is place-based, allowing the comparison of OSM activity in cities with similar populations or geographies.

The question of scale of edits also plays an important role in this study. If people are editing small cities in OSM as part of some thematic effort across a broader region (such as a state or country), it could signify a general increase in map content for small cities. In this situation, "a rising tide lifts all boats". At the same time, if most edits come from these broad regional level projects, we might also expect a lack of local richness in the data that could miss the amenities and services that make a particular city unique.

The abundant literature on Wikipedia growth may help inform hypotheses on OSM development among small and large cities. For example, Iba et al. (2010) discovered two main article types in Wikipedia: (1) articles of narrow focus created by a small number of subject matter experts, and (2) articles about broad topics, created by thousands of editors. In the case of crowdsourced mapping, perhaps there is an analogy to be drawn between these article types and respectively (1) small cities of a local or regional interest, and (2) large cities of a global interest. To date, literature analyzing OSM activity in small cities has been too scarce to confirm this, leading to the present research.

Methods

For the core of this analysis, as introduced above, three small cities (between 40,000 – 60,000 inhabitants) were selected from the Pampas region of South America. This is an area primarily used for agriculture and grazing, with very low population density outside of its scattered cities and towns. The selected cities were General Pico, La Pampa province, Argentina; Tres Arroyos, Buenos Aires province, Argentina; and Tacuarembó, Tacuarembó department, Uruguay (Figure 3-1). The straight-line distance between any two of these cities is over 400 kilometers. The intent of selecting three cities in this area was to get a feel for the proportion of

OSM contributors active at a regional level versus the proportion that only contribute for one place.

To complement this analysis, the much larger city of Rosario, Argentina (population over 1 million) in the Pampas region was also studied. The motivation behind this was to learn how the contribution patterns in the small cities related to patterns in a larger city in the same region, and what proportion of the contributor sets overlapped between these cities. Argentina's primate city of Buenos Aires was not selected because its massive size causes it to be of disproportionate political and economic importance beyond the bounds of the region. Furthermore, this city has already been examined in Neis et al. (2013).[10]

Figure 3-1. South American cities whose OSM data was studied in this analysis.

[10] Neis et al. (2013) found Buenos Aires to have a relatively low density of OSM contributors when normalized against population and compared with 11 other large world cities. Their methods found slightly over 30% of contributors to the Buenos Aires map to be "External Mappers", with the remainder being "Local Mappers" or "Vicinity Mappers". The "senior mappers" in Buenos Aires had higher average numbers of days active in the project and number of features created than most of the other cities studied.

Finally, to compare how editing patterns in South America relate to locations in the Global North, two small cities of population 40,000-60,000 were selected from the Interior Plains of North America, a region similar to the Pampas with regards to its agricultural and grazing land use and low population density. These cities were Salina, Kansas state, USA; and Brandon, Manitoba province, Canada (Figure 3-2). Although more cities in North America and elsewhere could have been selected for comparison, it was desirable to keep this particular analysis and discussion focused primarily on South America, as there has been less study of OSM in this region. Results by Quinn (in press) indicate that there is substantial influence on the South American OSM coming from residents of other continents, especially in rural areas, further making this region desirable for additional investigation.

Figure 3-2. North American cities whose OSM data was studied in this analysis.

To understand the amount of data available in each city for each contributor, the "full history dump" file was downloaded from the OSM website, covering a time period from 2007 until the final week of the year 2014. The data was clipped to rectangular bounding envelopes

around each city. Only OSM nodes (points) and ways (lines and polygons) were examined in this study. OSM relations are conceptually and technically challenging to clip to a study area and have no visual component, therefore relations were removed or ignored during processing. Using a Python script, I constructed point and line geometries of each version of each feature in the history file. This allowed the creation of small multiple maps (discussed later) to visualize each contributor's body of work.

A separate file containing the OSM changeset history was also downloaded. A changeset contains metadata about a group of edits uploaded to the OSM database at a single time, typically when the contributor invokes the Save option in his or her editing program. A changeset can be loosely viewed as a unit of work. A useful piece of the changeset metadata is the rectangular bounding box of the edits, showing the geographic range over which the edits were applied. Another valuable part of the changeset metadata is the free-form comment that contributors can attach to the changeset justifying or describing their edits. This comment can provide indications of user language, geographic origin, and motivation.

Each contributor's comments were reviewed by both human and machine. I first analyzed each contributor's set of comments and noted any favored edit types or habits, at the same time beginning to develop categories for the most commonly occurring editing practices. I then submitted the comments to the langid.py Python language identification module (Lui and Baldwin 2011, 2012). For each contributor, the most common language detected for that contributor across all work in the OSM project was noted. An analysis of the relationship between language of OSM comments and the actual place of contributor origin can be found in Quinn (in press).

Many OSM contributors edit the map in more than one place (Neis and Zipf 2012). In order to understand how much of a contributor's work was dedicated to a particular city, each contributor's total number of changesets in that city was tallied and compared with his or her total

number of changesets in OSM. This analysis distinguishes the passer-bys and happenstance contributors from those who dedicate a considerable portion of their efforts to the city.

Another metric calculated was the number of days each contributor was active in mapping the city. From these results it was possible to derive the total number of person-days of mapping activity for each city, summing for all contributors the number of days each contributor made an edit to the project. These metrics reveal the most committed contributors and give a picture of the overall level of mapping activity in each city. Counting the number of active days in the project has some advantages over examining raw numbers of edits or changesets. Some people import or digitize line or polygon features with a high density of nodes, complicating a metric of influence based on the number of edited nodes alone. Counting the number of changesets is also a problematic measure of influence in the project, because some editors save their work much more frequently than others, thereby creating many changesets.

Further information about the motivations and geographic origins of the contributors is sometimes available an OSM profile or wiki page if the contributor has chosen to create one.[11] These are publicly available web pages in the openstreetmap.org domain where contributors can optionally offer biographical information about themselves and their work. For each contributor who had made a page, I read the page and noted the language of the page and any editing preferences or hobbies. The Google Translate online translation utility was used for assistance with the relatively few cases where the page was in some language other than English, Spanish, or Portuguese.

The contributors' profile and wiki pages also often provided information about their home cities or countries, offering some understanding of which contributors might be considered to have local knowledge of the mapped area. These home locations were noted whenever a

[11] To give an example of the proportion of users creating profile or wiki pages, in the large city studied here (Rosario, Argentina) there were 50 out of 191 contributors who had either a profile or wiki page detected (26.2%). Twenty of these contributors had created both types of pages.

contributor mentioned one. In the case where the home location was not directly mentioned on the page (or in any personal website linked directly from the page), the contributor location could sometimes be derived from the reported list of special interest user groups that he or she had joined, (for example "Users in Seattle"). Some people join groups from multiple locations, therefore if two geographies were in conflict, the parent geography was recorded (ie, if the user belonged to "Users in Berlin" and "Users in Dusseldorf", Germany was recorded). If the groups fell within different countries and no other clues were available, no geography was recorded.

Results

The above methods revealed new insights on the nature of the crowd, while also reinforcing some findings of previous research (my own and that by others). The differences in contributor patterns between the small cities and the large city do not always fit proportional differences in population size, and reveal that different map construction processes are being enacted at different scales. The results below first report what was learned about the size of the crowd, followed by the geographic origins of the crowd, retention of the crowd, scale of edits, and the individual motivations and editing habits of crowd members.

Size of the crowd

In the small cities studied in this chapter, OSM has not reached the level of being built by a "crowd", but rather is being constructed by a dedicated "handful" of contributors who vary in motivations and activity levels in the project. Table 3-1 shows that each small city was built by just a few dozen individuals.

Table 3-1. Contributor metrics showing the size and activity of the crowd in the cities studied.

City name	**City type and region**	**Number of contributors in OSM history file**	**Total person-days contributed**	**Contributors active only one day here**	**Contributors active more than five unique days here**
General Pico, ARG	Small city – Pampas	35	101	18 (51.4%)	3 (8.6%)
Tacuarembó, URY	Small city – Pampas	25	104	12 (48.0%)	4 (16.0%)
Tres Arroyos, ARG	Small city –Pampas	29	85	17 (58.6%)	3 (10.3%)
Rosario, ARG	Large city – Pampas	191	1028	120 (62.8%)	20 (10.5%)
Brandon, Manitoba, CAN	Small city –North American Interior Plains	46	113	27 (58.7%)	5 (10.9%)
Salina, Kansas, USA	Small city – North American Interior Plains	58	167	33 (56.9%)	5 (8.6%)

The number of person-days reported in each row offer some insight into the activity levels of the contributors in the different cities. For example, Tacuarembó has the smallest crowd (with only 25 contributors), but it has a higher number of person days (104) than General Pico and Tres Arroyos. This suggests that at least some contributors working on the map in Tacuarembó are more actively mapping than their counterparts working on the other two cities.

The large city, Rosario, saw 191 contributors, many times the number seen in the small South American cities studied. This lends credence to the analogy made earlier between OSM and Wikipedia articles invoking the study by Iba et al. (2010): more narrowly focused topics, or

cities, tend to attract smaller contributor crowds. Furthermore, for practical purposes, the actively working crowd in any given city is much smaller than the total number of contributors in that city. The set of contributors never returning to map anything else in the city after their first day of contribution ranges from 48 – 63%. (Steinmann et al. (2013) reported this figure at 53% for the OSM project as a whole.) Furthermore, none of the small cities studied had any more than 5 contributors who were active more than five unique days in mapping. Consequently, a small number of actively mapping contributors can wield a high proportion of influence on the map.[12]

These varying quantities of edits between users are evident in Figure 3-3, which visualizes the nodes and ways modified by each contributor in the city of Tres Arroyos. These maps are arranged by the number of unique days the contributor was active in the project, starting at the top with the most active contributors and moving from left to right, then down to the next row and so on until reaching the bottom part of the graphic where the least active contributors are located. Most contributors have only modified one or a few nodes or ways, while a small group of contributors have modified features throughout the town. The two dense gray splotches near the upper-right of the image were from users who modified address ranges on most of the town's streets.

[12] Parr (2015, p. 132-133) described two US metropolitan areas where the majority of the mapping was performed by a single individual. In one of the cases, 99% of the features in the city were added by one contributor.

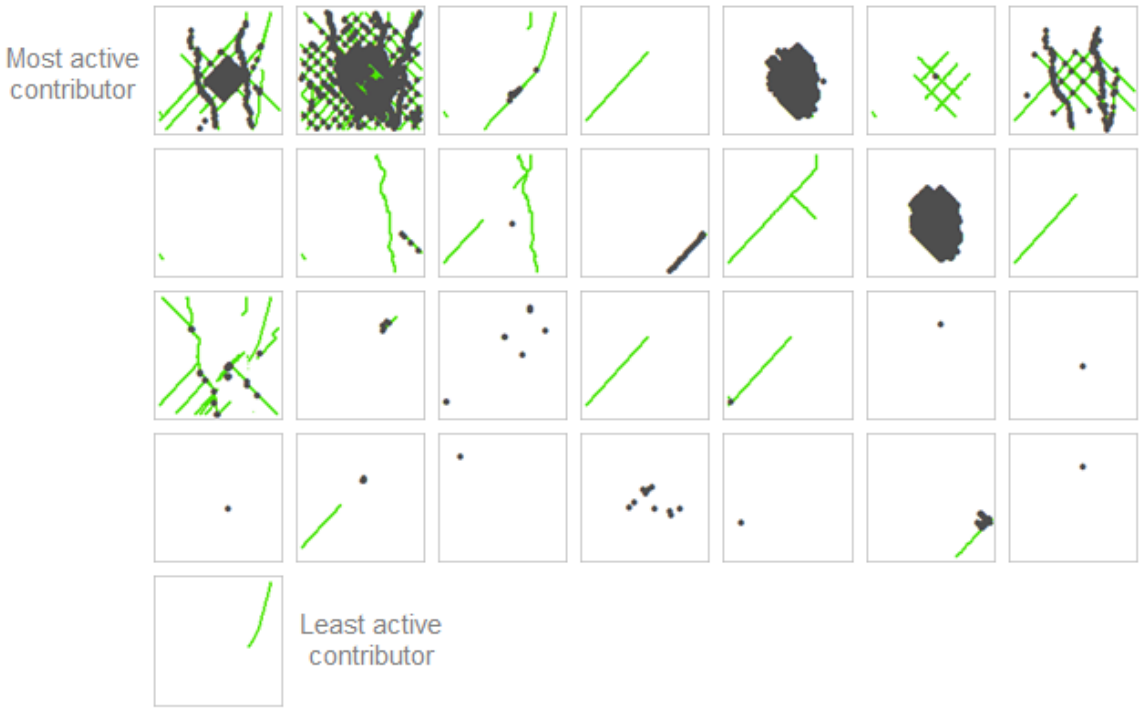


Figure 3-3. Small multiple maps of each contributor's work within Tres Arroyos over the history of the OpenStreetMap project. Edited ways are green lines and edited nodes are gray dots.

The purpose of Figure 3-3 is not to comment on the value of each contributor's work, but rather to show that a small group (and small percentage) of people do most of the work in a place. Anthony et al. (2005) demonstrated that contributions of Wikipedia editors who only make one edit can sometimes provide more value to the project (in terms of longevity and "staying power" of the edit) than those offered by more active contributors. This could be the case in OSM as well, as Parr (2015, p.123) found that the users with the lowest mean spatial error in feature placement tended to have low mapping activity, but high attention to the context of the project and the geographic detail of what they were adding. Further research is needed to determine if the activity levels of a user and the actual longevity of his or her OSM feature contributions are connected.

In viewing Figure 3-3, it's also important to remember that contributors who make only one edit in a particular town are not necessarily "newbies" to the project. Comparing the number

of changesets each contributor made in the town versus his or her total number of changesets in OSM revealed many cases where very active OSM contributors were just passing through the town, either physically or virtually, to map a single feature for one of a variety of reasons discussed later in this chapter.

Geographic origins of the crowd

Although precise locations of each crowd member are not known, a picture of the crowd's familiarity with any particular city can be formulated through multiple indicators. The most direct way to derive an OSM contributor's location is from self-reported information in the contributor's public profile or wiki page on the OSM site; however, many contributors choose not to create these pages and even fewer indicate their home place. For example, profile and wiki pages indicated a location for 36 (18.8%) of the 191 contributors to the map in Rosario (see Table 3-2). Another method to infer the locality of the contributor in non-English speaking countries is to detect the language of contributor comments in the changeset metadata and the language used on any profile or wiki pages created (if these exist). Quinn (in press) used the South American OSM to show that places dominated by Spanish and Portuguese comments have a heavier local influence than places commented in English. Finally, a way of inferring whether a contributor is local to a particular city is to calculate the percentage of the contributor's total OSM changesets that fall within the city. Contributors with a higher percentage of edits in the city would be expected to have more familiarity with the city, although this conclusion is not certain for any particular contributor and becomes less reliable when the user has very few total edits.

Figure 3-4 shows the most commonly detected language in OSM changeset comments for each of the top 10 contributors in the Pampas region cities, where the native language is Spanish. These rankings and all others in this chapter were determined by ordering the contributors by

number of unique days active in OSM in the city.[13] These languages were detected by the langid.py Python module, as described in the methods section. Languages listed are derived from the contributor's work across the entire OSM project, and not just within the target city. If a country of origin could be determined from the user's profile or wiki page, it is listed in parentheses. Contributors who appear in multiple cities are connected by a line.

Contributor rank (by days active in the city)	Tacuarembó, URY (small city)	General Pico, ARG (small city)	Tres Arroyos, ARG (small city)	Rosario, ARG (large city)
1	Spanish (Uruguay)	Spanish	Spanish (Argentina)	Spanish (Argentina)
2	Spanish (Uruguay)	English (Germany)	Spanish	English (Germany)
3	Spanish	Spanish (Argentina)	English (Germany)	None
4	German	Spanish	English	Spanish
5	Spanish	Spanish	Spanish	Spanish (Argentina)
6	English (Germany)	Spanish	Spanish (Argentina)	Spanish
7	English	Spanish	Spanish	Spanish
8	Portuguese	Spanish	French	Spanish
9	French	None	Spanish	Spanish
10	Spanish	English	Spanish	Spanish

Figure 3-4. Languages favored in OSM changeset comments by top contributors in each city. Places of origin, if known, are listed in parentheses.

Most of the top contributors speak Spanish and are likely from the country they are mapping; however, in the smaller cities there is a greater share of contributors employing languages other than Spanish, to the point where Tacuarembó sees five different languages among the top 10 contributors. The smaller cities, therefore, may be receiving more influence from armchair mappers and might have a need for more place-specific knowledge to be added by contributors who can collect data on site.

[13] Ties were broken by favoring the contributor with the larger number of changesets in the city. A further tiebreaker (if needed) favored the contributor with the higher percentage of changesets in the city compared to his or her total number of changesets in the entire OSM project.

When considering all contributor locations reported in profiles (not just the top 10), the most prevalent countries of origin were, in order, Argentina, Germany, Brazil, and Uruguay (Table 3-2). Although this information is likely reliable, it is unclear how well it extrapolates to the entire body of contributors in these cities, just because so many users did not report a location. The presence of Germany is not a surprise as Neis and Zipf (2012) estimated that over a quarter of all OSM users are from Germany. Even if this figure has diminished in the past several years, the table below suggests that German influence in OSM across the world is still substantial. When looking at the number of contributions in the study areas (not reported in this table), German contributors in the Pampas cities tended to be lighter contributors than Argentines and Uruguayans.

Table 3-2. Countries of origin determined from public OSM user profile and wiki pages, in the format <number of contributors: country>.

Tacuarembó, URY (small city)	**General Pico, ARG (small city)**	**Tres Arroyos, ARG (small city)**	**Rosario, ARG (large city)**
18: Unknown	24: Unknown	19: Unknown	155: Unknown
2: Germany, Uruguay	6: Argentina	5: Argentina	12: Argentina
1: Brazil, Canada, Italy	3: Germany	4: Germany	8: Germany
	1: Canada, Uruguay	1: Canada	4: Brazil
			3: UK
			1: Belgium, Bolivia, Canada, Hungary, Sweden, Switzerland, Turkey, Uruguay, USA

Retention of the crowd

The activity level of the crowd in the small cities has fluctuated over time, as shown in Figure 3-5. This graph measures the attention given to each of the five small cities by tallying the number of person-days of activity seen each year. In this calculation, the person-days do not represent units of 24 hours of work (we do not know the number of hours spent on making the edits), rather they are the number of days that a contributor made any contribution, summed for all contributors. Although a slight general increase is visible from the beginning of the project to the present, the numbers often move up and down from one year to another. In contrast, the larger city of Rosario has increased its amount of mapping activity in each of the past five years. This suggests that larger cities may be more likely to amass a consistently growing contributor base over time (most likely including local residents), while smaller cities are at the mercy of more transient mapping activity by passers-by.

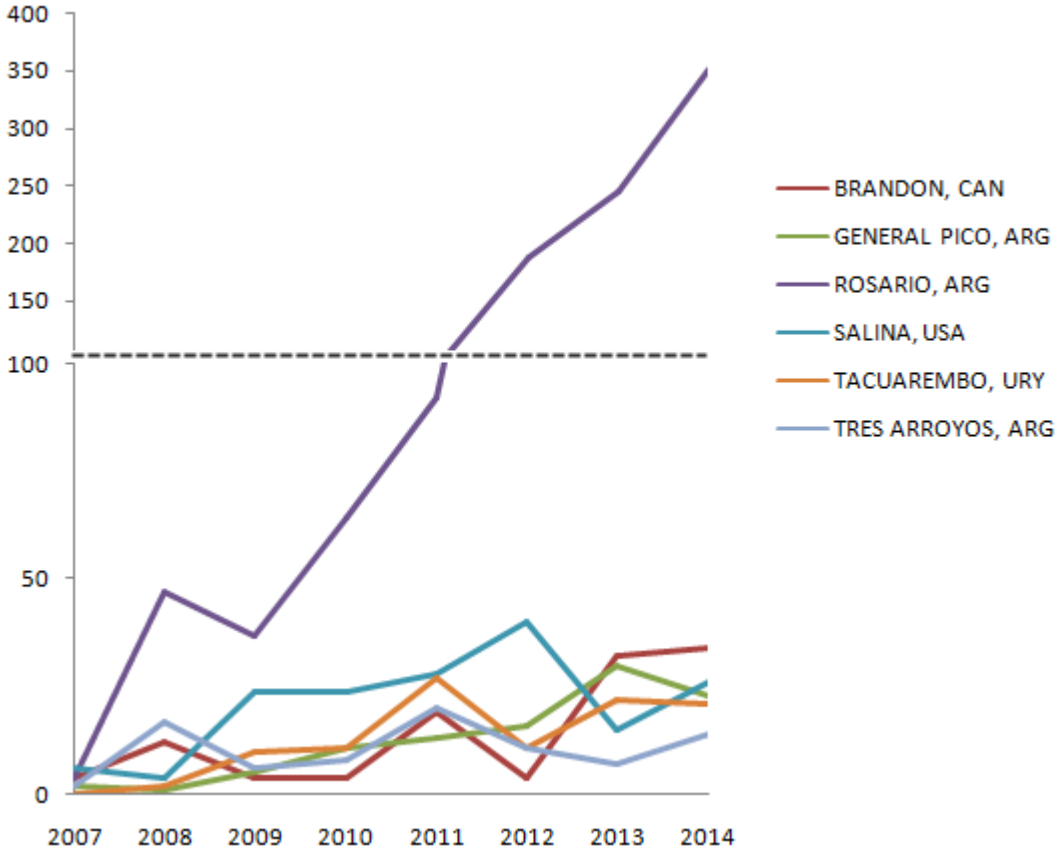


Figure 3-5. Yearly person-days of mapping

Although the larger city of Rosario is increasing its mapping activity at a greater rate, it performs about the same as the other cities when considering the percentage of contributors active during the most recent year (Figure 3-6). Most cities hover around the 25 – 35% range; however, not all cities exhibit the same proportions of active contributors. General Pico, in particular, saw over half of its contributors active in 2014. For many of these contributors, 2014 was the only year they edited the city. Although a group event such as a mapping party or local school project could skew the numbers in this way, there is no evidence that this occurred in General Pico; however, several paid mappers and a bot appeared in the area making fixes during 2014 (discussed further below). This might indicate that more attention is being paid to smaller cities in recent years as part of a broader effort to improve regional OSM quality.

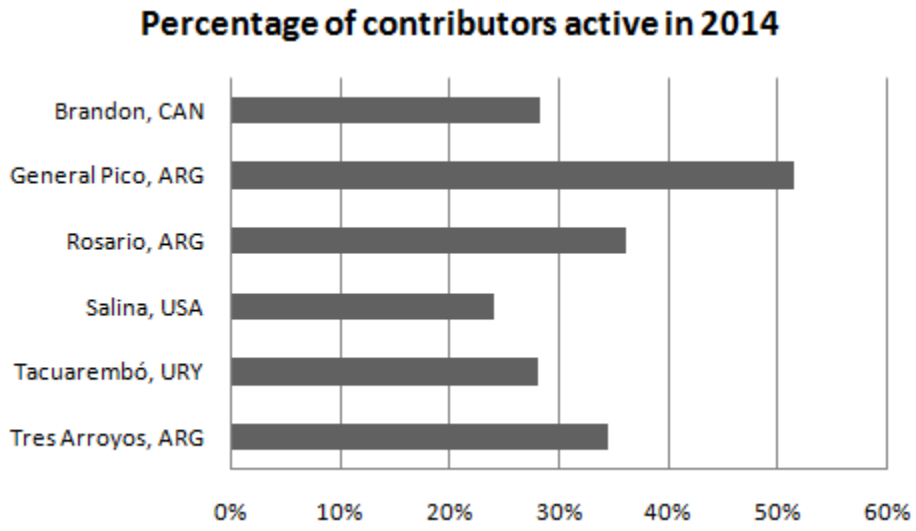


Figure 3-6. Percentage of contributors active in 2014.

Additional evidence of large cities retaining contributors appears when examining the range of active years for the top contributors for each city. I demonstrate this by examining the large city and then comparing it to one of the small cities. Figure 3-7 summarizes the top 10 contributors in Rosario, ranked by total days active in the project. All of the contributors in this list have been active during at least two different years, and over half have been active during at least four consecutive years, with most of those years occurring in the more recent history of the project.

Top 10 contributors in Rosario, Argentina

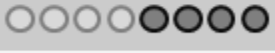
User rank	Active days	Changesets here	Percent of user's changesets here	Active years here '07 '08 '09 '10 '11 '12 '13 '14
1	348	769	42.5%	
2	83	213	1.1%	
3	55	434	33.1%	
4	32	49	62.8%	
5	29	38	1.7%	
6	28	51	10.4%	
7	23	111	4.5%	
8	21	59	51.3%	
9	20	26	12.7%	
10	19	33	47.1%	

Figure 3-7. Activity levels of top 10 contributors in Rosario, Argentina (large city).

In contrast, the top contributors in the smaller cities have been active for fewer years, a natural result of them making fewer contributions overall. A key difference is that most contributors in a given small city have made only a very small percentage of their OSM changesets there, suggesting they do not have a primary interest in the place being mapped and may be less likely to have collected data on the ground there. For example, in Tacuarembó, only two top contributors have made over 10% of their OSM changesets in Tacuarembó (Figure 3-8), whereas in Rosario seven out of the top 10 contributors have made over 10% of their edits in Rosario. Similar patterns exist for the other small cities studied here. Thus, pride of place as a motive for editing OSM may have a better chance of being observed in larger cities, whereas

smaller cities are primarily built by other motives such as pride of the map, in other words, the desire to ensure that high quality OSM coverage extends everywhere.

Top 10 contributors in Tacuarembó, Uruguay

User rank	Active days	Changesets here	Percent of user's changesets here	Active years here '07 '08 '09 '10 '11 '12 '13 '14
1	27	39	0.6%	○○○○○●●●
2	18	31	2.5%	○○●●●○○○
3	9	32	2.1%	○○○○○○●●
4	6	7	0.2%	○○○○●○○○
5	5	15	62.5%	○○○●●○○○
6	5	8	< 0.1%	○○○○●●●○
7	4	7	2.3%	○○○●●○○○
8	4	5	0.2%	○○○●○○○○
9	4	5	< 0.1%	○○○○●○○○
10	3	33	18.8%	○○○○○○○●

Figure 3-8. Activity levels of top 10 contributors in Tacuarembó, Uruguay (small city).

Scale of edits

Each OSM changeset is accompanied by metadata containing the rectangular bounding box of the encompassed edits. This box is defined by the minimum latitude, minimum longitude, maximum latitude, and maximum longitude of the edits. Mapping the changeset bounding boxes can show the scales at which editors typically approach their OSM edit sessions and, by extension, their regions of focus.

The analysis of bounding boxes in all six cities shows that the map is made by a mix of editors working at different scales focused around 1) the city itself, 2) the local region such as a state or province, or 3) the containing country. The map in Figure 3-9 of changeset bounding boxes for General Pico shows these patterns in action.[14] The city itself is surrounded by a dense number of bounding boxes as expected, which are just a small "pinpoint" in these maps. There are also many bounding boxes covering the Pampas region and the La Pampa province in central Argentina, to which General Pico belongs. Beyond that, there are relatively few bounding boxes until reaching the national boundary of Argentina itself. The numerous bounding boxes surrounding the country indicate that some contributors were brought to General Pico because of edits they were applying at a national scale. This systematic approach to OSM at the geopolitical unit level discounts affinity toward a particular town as a motive for the edits, and contradicts an imagination of VGI as a product of citizens sensing only their local domains.

The maps in Figure 3-9 are colored by language used in the changeset comments as determined by me. Envelopes commented with nothing or text where the language was indistinguishable (such as toponyms or tag names) are colored gray, whereas envelopes commented in English are blue and envelopes commented in Spanish are red (No other languages were observed in this city). The maps demonstrate that a notable contingent of the English speakers who made an edit in General Pico were passing through as part of a more systematic editing project at the country level scale or making a very localized update within the town. In contrast, the edits by Spanish speakers tend to be more varied, with many changesets covering the region around General Pico and its province of La Pampa. The changesets whose comments were blank or indistinguishable exhibit both patterns.

[14] A very small number of intercontinental bounding boxes (such as those produced by global data imports) were ignored in this map, although I acknowledge the role of these types of edits in the map creation process.

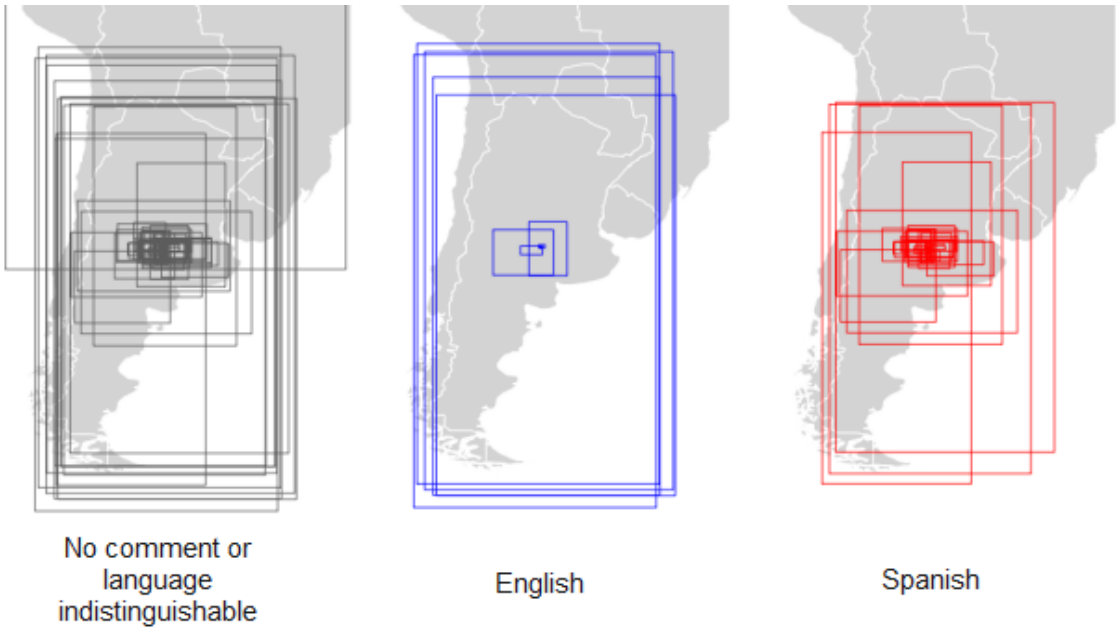


Figure 3-9. Changeset bounding boxes for General Pico, colored by language of the changeset comment.

Contributor motives and activities

Small cities provide a bounded and comprehensible subset of contributors that can be further studied to understand how OSM takes shape in a place. The fact that few of the these contributors seem unlikely to have performed local surveys in the areas they are mapping makes it easier to sort out the wide range of other contributor motives that feed into OSM construction beyond the basic desire to share local knowledge. By also analyzing the large city of Rosario, I found that these motives are not exclusive to small cities; however, they were easier to identify when working with the small and highly variable sets of contributors in small cities.

By analyzing the editor comments in these cities, I have developed a typology of roles played by contributors as they approach the OSM project. Although Steinmann et al. (2013) created useful "contribution profiles" to categorize OSM contributors based on the types of

geometry and attribute data modified, there has not been an attempt to describe user activities based on actions and thought processes mentioned in the changeset comments. The closest has been Parr (2015) who factored changeset comment length into his activity-context-geography model of OSM contribution; however, Parr's model was applied to a massive dataset of all OSM contributions in the US and no qualitative analysis of the comment text was attempted. The analysis below is also unique in that it acknowledges that contributors could be assuming more than one role at a time when they make certain contributions.

The roles always involve mapping new features or fixing/enhancing existing ones. These mapping and fixing roles can be enacted in either a casual, a systematic, or an automated fashion. In combination with these mapping and fixing roles, I have identified some "special interest" motivations. These in particular have the potential to bring armchair mappers to reach beyond the major urban areas into small towns and cities. Table 3-3 shows the identified types of contributor roles, accompanied by changeset comments that are representative of each role. The comments were taken directly from the history of the five small cities in this study. A more detailed discussion of each role follows.

Table 3-3 . Roles describing OpenStreetMap contributor activities, with changeset comments representative of each. * indicates a translation from Spanish by the author.

Role type	**Role**	**Example changeset comments**
Mapper roles	Casual mapper	• "Service station ACA El Bolson"* • "Added SnoMan trail"
	Systematic mapper	• "Stoplights in the downtown area of General Pico"* • "Adding/correcting maxspeed tags on interstates in Kansas. Some from personal knowledge, some from KDOT maps. Removed some that I know were wrong but insufficient resolution on map (like I70 in Topeka - will survey on next trip)"
	Automated mapper (Importer)	• "canvec import in western manitoba " • "adding airports from ourairports.com" • "Active Mines and Mineral Plants in US http://www.data.gov/details/13"
Fixer roles	Casual fixer	• " Fixed connectivity issue spotted thanks to www.maproulette.org • "The name is TaCuarembo, not TaGuarembo :)"
	Systematic fixer	• "Splitting ways"* • "Reclassifying highways with new criteria, UY."* • "hgv=* on high priority corridors"
	Automated fixer (Bot)	• " Removing surrounding whitespace and empty tags"
Special interest motivations (always enacted through mapper or fixer roles above)	Topophilic mapper	• "Fast food. Serves Hamburgers and Ice Cream. Kind of like Dairy Queen but can be more reasonable" • "My parents' house"*
	Paid mapper	• " fix unconnected roads (http://osmlab.github.io/to-fix/?error=unconnected_major1) "
	Feature-specific hobbyist	• "railway work in MB" • "power lines"
	Event or "crisis" mapper	• (Not observed in these towns, but documented elsewhere)

Below I describe these roles in more detail, accompanied by some commentary about how they may appear in editor comments in small cities.

Casual mapper

Neis and Zipf (2012) showed that most OSM contributors only make one or a few edits during the duration of their involvement with the project. Sometimes these casual contributors are residents of small cities, mapping items from their town.

In other cases, a heavy OSM user has visited or passed through on vacation, business travel, and so forth. Such a contributor's influence on the town may be limited to a "casual" one or two changesets, which are still potentially of great value because they originate from a local survey.

Systematic mapper

In the study areas it is apparent that a few contributors have taken upon themselves the work of adding certain feature types in a systematic and comprehensive fashion across the entire city. These include street names, address ranges, stop lights, and other entities that make the map more useful for routing and other functions. In some cases, a single contributor was instrumental in initiating more than one of these projects in a given city.

Some degree of local knowledge or ground survey experience is helpful when addressing the project in this role, although some activities such as manual tracing of building footprints from remotely sensed imagery may be performed by "armchair mappers".

Contributors in the systematic mapper role tend to have high numbers of edits in the OSM project in general. They are what Steinmann et al. (2013) called All-Rounders. Systematic mappers are also often heavily involved in the "fixer" roles described below.

Automated mapper ("Importer")

Provided there is buy-in from the local mapping community, OSM allows users to import geographic data whose license terms are compliant with the Open Database License. These imports sometimes cover scales as large as a country, and therefore wind up encompassing small cities. The US Census TIGER street data import is one such example that affected the city of Salina, Kansas in this study. Another import brought in a dataset of airports from around the world.

Imports can affect vast amounts of data in one transaction, therefore a contributor with a small number of changesets and days active can still have an enormous effect on the map for a long period of time. Because of the skill and community communication required, contributors who can act in the importer role are rare; however, their actions are highly visible.

Casual fixer

OSM editors assume the casual fixer role when they notice existing features that could use correction or enhancement some way, such as through adding or updating attributes or adjusting geometric topology. These actions could be spurred by noticing something that contradicts their own local knowledge when browsing the map in a place of interest. At other times, contributors familiar with the OSM project may notice aspects of data that contradict or fall short of the community-established attribute schemas.

Contributors can also receive suggestions of places to fix from third-party applications that randomly display features deemed as likely errors in the map. MapRoulette and MapDust are examples of such applications that could bring contributors to small cities and towns that they would not otherwise attend to. In some cases these applications are built by companies who have based their business models on OSM data and have an interest in high-quality coverage across the landscape. This business activity may therefore be driving an improvement of OSM quality in small towns.

Systematic fixer

Some OSM contributors focus on tidying up and improving existing metadata and geometry information instead of, or in addition to, mapping new features. McConchie (2013) calls these contributors "map gardeners", after the term "wiki gardener" used in Wikipedia contexts. They adjust nonstandard tags, re-align geometries, split up ways into routable segments, clean up imported data, etc. in a systematic fashion across geographies, influencing small cities as well as large ones.

The work of systematic fixers is critical for OSM's quality. OSM power users often play both the systematic mapper and systematic fixer roles.

Automated fixer (Bot)

The OSM community allows for automated programs (also known as "bots") that do data cleanup and maintenance activities in a manner more global and consistent than could be achieved by manual work. For example, a bot may remove trailing whitespace from tags, or

adjust punctuation or spelling to meet certain conventions. Bots are often deployed on a global, continental, or country scale in OSM, and therefore their work is evident in small cities.

Topophilic mapper

The topophilic mapper is driven by expansive local knowledge and pride of place, resulting in participation across both mapping and fixing roles in either a casual or systematic fashion. In the small cities studied here, contributors fitting the topophilic mapper role are sometimes not native to the small city itself, but rather from the surrounding region. In these cases, they can still access the city for ground surveys and other rich data, and they may have more interest in the city than others simply because it has a greater chance of being part of their lives. Using a ranking mechanism where contributors are sorted by days active in the city, changesets in the city, and percent of all OSM changesets in the city (see Figures 3-7 and 3-8 for examples), three of the same contributors appear in the top 10 list for Rosario, General Pico, and Tres Arroyos, Argentina. Similarly, in Salina, the top mapper is based in urban Kansas, but not Salina itself, and in Tacuarembó, one of the most prolific contributors has made thousands of edits throughout Uruguay.

Paid mapper

In the past several years, some tech companies specializing in web mapping have begun relying heavily on OSM to provide contextual background data, routing, and other services. As these companies base their services and strategies on OSM, their incentive to see a high quality map grows, and they are willing to invest in paying employees to fix errors and build out geographies needed by clients (Barth 2015). The same has occurred with governments wanting to

use OSM for basemaps while ensuring that the data meets a particular standard of accuracy and overall quality (see McHugh (2014) as an example). This paid mapping work can affect small cities where errors or holes may have persisted in the data. Paid mapping is likely to be systematic in nature, whether it involves mapping new features or fixing existing ones.

In the cities studied here, several edits made by a mapping corporation's data quality team were found. Their edits focused on fixing unconnected roads, a feature needed for cross country navigation and logistics applications, whether in the routing engine itself or in the base map. Ultimately we cannot be certain of the purpose of these particular efforts, but they were easily detected because mappers hired by this particular corporation clearly identify themselves and their employers via OSM profile pages. Other paid mapping efforts may not be so readily noticed, depending on the transparency policy of the employer and the nature of the work (whether the OSM editing is full time or just a small part of a broader job).

Feature-specific hobbyist

One of the motives for nonlocal users to edit OSM is a desire to map certain types of entities that are closely related to personal hobbies or interests. These can include railroads, power lines, bicycle trails, places of worship, and so forth. Many contributors identify in their user profiles that they tend to specialize in one or more of these types of features.

The desire to add hobby-related features across a particular geography, such as a country, can bring OSM users to new places and cause them to contribute to the map construction within small cities and towns. For most hobbyists, it is not so much interest in the place nor interest in the map that drives the work, but rather interest in the mapped entity itself. A substantial proportion of OSM entities in small cities appear to be contributed because of hobbyists virtually "passing through" the map as they expand coverage of their hobby interest through imagery

tracing or imports. At other times, the hobby work is more casual and occurs because a hobbyist physically visited the area for some reason.

Event mapper

Although not detected in any of the small towns studied here, it is well documented that the desire to help with humanitarian or "crisis" mapping has brought many OSM editors to work on the map in places they have never visited (Zook et al. 2010, Peylen et al. 2015). Such efforts were instrumental in creating maps for aid personnel following the 2010 Haiti earthquake, Typhoon Yolanda in the Philippines in 2013, and the 2014 Ebola outbreak in West Africa. The resulting mapping work affects small cities as well as large ones, and in some cases heavy mapping occurs in places that saw little or no attention previously. This mapping is systematic in nature, but can be limited in depth if nobody is available to supplement remotely traced data with local knowledge.

Conclusions

Viewing the history of OSM in any given place reveals a chance conglomeration of individuals with all kinds of motives and interests, including feature-specific hobby work, local surveying, and map quality assurance. Studying small towns and cities provides bounded study sets of contributors, whose sparse numbers of systematic local mappers actually make it easier to identify the other types of contributor personas involved in the project. Some contributors are driven by pride of place, some by pride of the map, and some by pride of a certain type of feature related to personal hobbies. Just as small cities often spring up around junctions of long-distance highways and thoroughfares, the digital maps of small cities in OSM represent a crossroads of

virtual mappers passing through for their own unique reasons. These myriad influences behind OSM reveal that VGI cannot always be conceptualized as the product of altruistic citizen sensors who are sharing personally-collected tidbits of spatial knowledge.

In each of the small towns studied, the "crowdsourced" data of OSM was actually created by several dozen individuals, only a few of whom were active in the project on a consistent basis. The amount of work performed by each contributor varies greatly, and can be better conceptualized by mapping each contributor's work side-by-side and studying the number of days and changesets contributed to the project. Because of the relatively small contributor set, the few active contributors and their personal interests and priorities wield a powerful influence on what types of entities are emphasized in the map. Interestingly, some of the top mappers in any given city may actually do most of their work in other places in OSM. Regional contributors who have some local knowledge and access to the place on the ground can play a great role in building the map in small cities.

Different visualization approaches can help with distinguishing casual vs. systematic vs. automated contributions. For example the small multiple maps in Figure 3-3 exhibit great "blobs" of data for users who have employed systematic or automated contribution strategies. Likewise, the changeset extent boundary maps in Figure 3-9 indicate where systematic and automated contributions have been employed at different geographic scales. Introducing interactivity in the visualization strategies may offer greater potential for understanding OSM contributor characteristics and editing habits. This will be explored in the next chapter.

In comparison with larger cities, OSM in small cities appears to be struggling with attracting and retaining new users. Increasing the number of topophilic mappers seems to be in the best interest of data quality and comprehensiveness in OSM's small cities. Recruitment through secondary or vocational schools may be a promising channel for finding new active mappers in places where there are no large universities or communities of technical professionals.

Also, urban areas with active local OSM groups can reach out to hold mapping parties in nearby small towns and cities, thus diffusing the knowledge of OSM throughout a region. The wide availability of video conferencing and screen sharing technology in many areas of the world might further expand OSM outreach, training, and events to places that could not otherwise be visited by groups of experienced mappers.

In order to examine a few places in depth, the present research is limited to a strategically selected handful of cities. In further research, it would be worthwhile to make a comparison with cities in countries like Germany or the UK where OSM has reached a greater level of maturity than in the Americas. Also, studies of suburban or peri-urban areas in OSM might reveal the geographic limits of OSM influence emanating from large cities with active mapping communities.

Organizations considering using OSM data should be aware that the data may greatly decrease in quantity and quality when moving outside major cities due to the general lack of attention and scrutiny paid to the map. Improving OSM in small cities is in the best interest of companies who are leaning their business models on OSM-based routing and logistics services; governments who are using OSM as a basemap for their services; applications that must provide global, national, or state-level map coverage; and humanitarian agencies that supply maps for crisis response in rural areas.

Chapter 3 references

Anthony, Denise, Sean W. Smith, and Tim Williamson. 2005. "Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia." *Hanover: Dartmouth College*. http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf.

Barth, Alex. 2015. "The Paid Mappers Are Coming." In *State of the Map US 2015*. http://stateofthemap.us/the-paid-mappers-are-coming/.

Bell, David, and Mark Jayne. 2009. "Small Cities? Towards a Research Agenda." *International Journal of Urban and Regional Research* 33 (3): 683–99.

Brunn, Stanley D., and Matthew W. Wilson. 2013. "Cape Town's Million plus Black Township of Khayelitsha: Terrae Incognitae and the Geographies and Cartographies of Silence." *Habitat International* 39: 284–94.

Budhathoki, Nama Raj. 2010. "Participants' Motivations to Contribute Geographic Information in an Online Community." Ph.D., United States -- Illinois: University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/bitstream/handle/2142/16956/1_Budhathoki_Nama.pdf?sequence=2.

Coleman, David J., Yola Georgiadou, Jeff Labonte, and others. 2009. "Volunteered Geographic Information: The Nature and Motivation of Produsers." *International Journal of Spatial Data Infrastructures Research* 4 (1): 332–58.

Elwood, Sarah, Michael F. Goodchild, and Daniel Sui. 2013. “Prospects for VGI Research and the Emerging Fourth Paradigm.” In *Crowdsourcing Geographic Knowledge*, 361–75. Springer. http://link.springer.com/chapter/10.1007/978-94-007-4587-2_20.

Feick, Rob, and Stéphane Roche. 2013. “Understanding the Value of VGI.” In *Crowdsourcing Geographic Knowledge*, 15–29. Springer. http://link.springer.com/chapter/10.1007/978-94-007-4587-2_2.

GisPro. 2007. “The GiSPro Interview with Steve Coast,” October.

Glasze, Georg, and Chris Perkins. 2015. “Social and Political Dimensions of the OpenStreetMap Project: Towards a Critical Geographical Research Agenda.” In *OpenStreetMap in GIScience: Experiences, Research, and Applications*, edited by Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, 143–66. Lecture Notes in Geoinformation and Cartography. Switzerland: Springer.

Goodchild, Michael F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211–21.

Graham, Mark. 2010. “Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth.” *Tijdschrift Voor Economische En Sociale Geografie* 101 (4): 422–36.

Graham, Mark, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. 2014. “Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty.” *Annals of the Association of American Geographers* 104 (4): 746–64.

Graham, Mark, Matthew Zook, and Andrew Boulton. 2013. “Augmented Reality in Urban Places: Contested Content and the Duplicity of Code.” *Transactions of the Institute of British Geographers* 38 (3): 464–79.

Grodach, Carl. 2009. “Urban Branding: An Analysis of City Homepage Imagery.” *Journal of Architectural and Planning Research*, 181–97.

Haklay, Mordechai. 2010. “How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets.” *Environment and Planning. B, Planning & Design* 37 (4): 682.

———. 2014. “OpenStreetMap Studies (and Why VGI Not Equal OSM).” *Po Ve Sham - Muki Haklay’s Personal Blog*. August 14. https://povesham.wordpress.com/2014/08/14/openstreetmap-studies-and-why-vgi-not-equal-osm/.

Haklay, Mordechai, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. 2010. “How Many Volunteers Does It Take to Map an Area Well? The Validity of Linus’ Law to Volunteered Geographic Information.” *The Cartographic Journal* 47 (4): 315–22.

Harley, J. Brian. 1988. “Silences and Secrecy: The Hidden Agenda of Cartography in Early Modern Europe.” *Imago Mundi* 40 (1): 57–76.

———. 1989. “Deconstructing the Map.” *Cartographica: The International Journal for Geographic Information and Geovisualization* 26 (2): 1–20.

Iba, Takashi, Keiichi Nemoto, Bernd Peters, and Peter A. Gloor. 2010. “Analyzing the Creative Editing Behavior of Wikipedia Editors: Through Dynamic Social Network Analysis.” *Procedia-Social and Behavioral Sciences* 2 (4): 6441–56.

INDEC (Instituto Nacional de Estadística y Censos). 2001a. “Ciudades Que Superan Los 50.000 Habitantes En 2001.” http://www.indec.mecon.ar/nuevaweb/cuadros/74/habitat2.xls.

———. 2001b. “Cuadro 2.1 Total País Según Provincia. Población Censada En 1991 Y 2001 Y Variación Intercensal Absoluta Y Relativa 1991-2001.” http://www.indec.gov.ar/micro_sitios/webcenso/censo2001s2/Datos/01000C21.xls.

Johnson, Peter, and Renee Sieber. 2013. “Situating the Adoption of VGI by Government.” In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, edited by Daniel Sui, Sarah Elwood, and Michael Goodchild, 65–81.

Latif, Sufian, KM Rakibul Islam, Md Monjurul Islam Khan, and Syed Ishtiaque Ahmed. 2011. “OpenStreetMap for the Disaster Management in Bangladesh.” In *Open Systems (ICOS), 2011 IEEE Conference on*, 429–33. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6079240.

Lin, Yu-Wei. 2011. “A Qualitative Enquiry into OpenStreetMap Making.” *New Review of Hypermedia and Multimedia* 17 (1): 53–71.

Lui, Marco, and Timothy Baldwin. 2011. “Cross-Domain Feature Selection for Language Identification.” In *In Proceedings of 5th International Joint Conference on Natural Language Processing*, 553–61.

———. 2012. “Langid. Py: An off-the-Shelf Language Identification Tool.” In *Proceedings of the ACL 2012 System Demonstrations*, 25–30. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2390475.

Mashhadi, Afra, Giovanni Quattrone, and Licia Capra. 2015. "The Impact of Society on Volunteered Geographic Information: The Case of OpenStreetMap." In *OpenStreetMap in GIScience: Experiences, Research, and Applications*, edited by Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, 125–41. Lecture Notes in Geoinformation and Cartography. Switzerland: Springer.

McConchie, Alan. 2013. "From Wiki Gardening to Map Gardening: Analyzing Contribution Patterns in OpenStreetMap." In *State of the Map US 2013*. San Francisco, California. http://vimeopro.com/openstreetmapus/state-of-the-map-us-2013/video/68097490.

McHugh, Bibiana. 2014. "Government as a Contributing Member of the OpenStreetMap (OSM) Community." In *FOSS4G 2014*. Portland, Oregon. https://vimeo.com/album/3606079/video/106226528.

Mooney, Peter, and Padraig Corcoran. 2013. "Analysis of Interaction and Co-Editing Patterns amongst OpenStreetMap Contributors." *Transactions in GIS*.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf. 2013. "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions." *Future Internet* 5 (2): 282–300.

Neis, Pascal, and Alexander Zipf. 2012. "Analyzing the Contributor Activity of a Volunteered Geographic Information project—The Case of OpenStreetMap." *ISPRS International Journal of Geo-Information* 1 (2): 146–65.

Parr, David. 2015. "The Production of Volunteered Geographic Information: A Study of OpenStreetMap in the United States." Doctoral dissertation, San Marcos, Texas: Texas State University. https://digital.library.txstate.edu/handle/10877/5776.

Perkins, C. 2011. “Researching Mapping: Methods, Modes and Moments in the (im)mutability of OpenStreetMap.” *Global Media Journal: Australian Edition. 2011;5(2):1-12.* https://www.escholar.manchester.ac.uk/uk-ac-man-scw:197593.

Peylen, Leysia, Robert Soden, T. Jennings Anderson, and Mario Barrenechea. 2015. “Success & Scale in a Data-Producing Organization: The Socio-Technical Evolution of OpenStreetMap in Response to Humanitarian Events.” In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4113–22. ACM.

Quinn, Sterling. in press. “A Geolinguistic Approach for Identifying Locally Contributed Content in OpenStreetMap.” *Cartographica*. doi:10.3138/CART.MS3301.

Raymond, Eric. 1999. “The Cathedral and the Bazaar.” *Knowledge, Technology & Policy* 12 (3): 23–49.

Steinmann, Renate, Simon Gröchenig, Karl Rehrl, and Richard Brunauer. 2013. “Contribution Profiles of Voluntary Mappers in OpenStreetMap.” In *Online Proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information, 16th AGILE Conference*. http://flrec.ifas.ufl.edu/geomatics/agile2013/papers/steinmann_ACTIVITY_AGILE_2013.docx.

Stephens, Monica. 2013. “Gender and the Geoweb: Divisions in the Production of User-Generated Cartographic Information.” *GeoJournal* 78 (6): 1–16.

Touya, Guillaume, and Carmen Brando-Escobar. 2013. “Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets.” *Cartographica: The International Journal for Geographic Information and Geovisualization* 48 (2): 134–43.

United States Census Bureau. 2010. “Interactive Population Map.” United States Census Bureau. http://www.census.gov/2010census/popmap/.

———. 2014. “Metropolitan and Micropolitan Statistical Area Totals Dataset: Population and Estimated Components of Change: April 1, 2010 to July 1, 2014.” https://www.census.gov/popest/data/metro/totals/2014/CBSA-EST2014-alldata.html.

Urban, Florian. 2002. “Small Town, Big Website?: Cities and Their Representation on the Internet.” *Cities* 19 (1): 49–59.

Wilson, Matthew W., and Mark Graham. 2013. “Neogeography and Volunteered Geographic Information: A Conversation with Michael Goodchild and Andrew Turner.” *Environment and Planning A* 45 (1): 10–18. doi:10.1068/a44483.

Wood, Harry. 2014. “The Long Tail of OpenStreetMap.” In *State of the Map 2014*. Buenos Aires, Argentina. http://vimeo.com/album/3134207/video/112438218.

Zielstra, Dennis, and Alexander Zipf. 2010. “A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.” In *13th AGILE International Conference on Geographic Information Science*. Vol. 2010. http://agile2010.dsi.uminho.pt/pen/shortpapers_pdf/142_doc.pdf.

Zook, Matthew, Mark Graham, Taylor Shelton, and Sean Gorman. 2010. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake.” *World Medical & Health Policy* 2 (2): 7–33.

Chapter 4

Exploring social influence in a crowdsourced map using geovisual analytics[15]

Abstract

Crowdsourced data products such as Wikipedia and OpenStreetMap (OSM) have grown so massive and ubiquitous that it is sometimes easy to forget they are merely the sum of individual human efforts, often tinged with error and bias. Most studies surrounding OSM focus on the positional or semantic accuracy of the data while passing over the individual nuances, motives, and geographies that persist in the map. We use a geovisual analytics tool called *Crowd Lens for OpenStreetMap* to understand characteristics of both the "crowd" and the unique individuals that constructed OSM in specific places. The tool reveals the size of the crowd behind OSM and how this fluctuates across different map extents. It uses small multiple maps to visualize each contributor's piece of the crowdsourced whole, and links OSM features with the free-form commit messages supplied by their contributors. Users of Crowd Lens can filter the contributor list by preferred language (in other words, the most common language across all changesets by a contributor), OSM attribute tags applied, and other characteristics. We describe the development and evaluation of Crowd Lens, showing how a user-centered design process with multiple stages and testing audiences (including geospatial technology professionals) helped shape the tool design. We use Crowd Lens to examine small cities of regional import in six continents, revealing large discrepancies in the attention of the crowd between the Global North and the Global South. Our findings are potentially useful to institutions deliberating OSM's

[15] This chapter was developed as an academic paper under the following citation:
Quinn, Sterling and Alan MacEachren. in preparation for submission. "Exploring Social Influence in the Global Crowdsourced OpenStreetMap Using Geovisual Analytics."

fitness for use in different contexts, as well as researchers studying how crowdsourced products can be better comprehended with visual analytics methods.

Introduction

In the past two decades the increasing pervasiveness and interactivity of the Internet has facilitated the creation of enormous stores of crowdsourced information. These projects such as Wikipedia, Yahoo! Answers, and OpenStreetMap (OSM) rely on the collective knowledge and experiences of contributors to constantly augment and improve the information product. Although it is easy to think of the crowd as amorphous and "out there" somewhere, each piece of information can be linked to an individual contributor in a physical location, and by the same logic, any large body of information can be traced to a finite contributor set.

As the potential value of these crowdsourced projects grows for both corporations and individuals, it becomes important to understand more about the crowd that contributed the information. For example, with any piece of the above-mentioned projects such as a Wikipedia article, Yahoo! Answer thread, or OSM town map, answers to the following might help potential users of the project understand how much credibility can be placed in its information:

- How many contributors have ever influenced this information? Just one, or many?
- How many contributors recently modified this information, such that we might guess how many are likely to enhance the information in the near future?
- If there is a lot of information, was it primarily contributed by one person, or was the work spread out more evenly among multiple individuals?
- Has systematic "vandalism" ever occurred with this information, and if so, what has been the severity and frequency of the defacement?

- What biographical or demographic information about the contributors can be deduced from the information products (or associated profile pages) that would help us understand the positionality and credibility of the contributors?

Understanding each person's contribution to a crowdsourced project can be cognitively and computationally challenging due to the amount of data and the number of individuals involved. Since many entities in these projects have multiple versions resulting from change over time, the complexity can even apply to a single item such as an article in Wikipedia or a highway in OSM. Furthermore, obtaining the raw historical data archives can require technical skill and computing resources. The volume of data to analyze is often very large and supplied in semi-structured text-based formats that do not immediately lend themselves to visualization. For these reasons, approaches from the research domain of visual analytics are particularly suited for addressing the above questions.

Visual analytics methods and tools attempt to facilitate analytical reasoning using interactive visual interfaces (Thomas and Cook 2005). Although static graphs, charts, and even maps could be made of any one phenomenon within a dataset, the ability to interact with the data and dynamically create new views of its varying dimensions can help drive the formulation of new questions and answers. Often visual analytics interfaces consist of multiple linked views that help the analyst get a picture of how different parts of a phenomenon are related (Roberts 2007). Visual analytics is particularly suited for addressing numerous dimensions of a crowdsourced project, including the demographics and scope of the contributor set, the timeline of contribution, and the nature of the contributed content itself.

This chapter describes *Crowd Lens for OpenStreetMap,* a visual analytics tool we have designed, implemented, and evaluated for learning about the construction of the OSM project in a place. OSM is a crowdsourced geographic database that is built online by volunteers in a format

similar to Wikipedia's, wherein anyone is invited to edit any geographic feature at any time. Contributors can modify feature geometries, feature attributes, or both. As a matter of practice, attributes are specified using a set of community-defined "tags". At the time of this writing, OSM boasts over 2.2 million registered users (although only a much smaller portion is active)[16], and OSM coverage has come to rival commercial and government-produced alternatives for reference mapping in some areas of Europe and the United States (Kounadi 2009, Haklay 2010, Neis et al. 2011, Zielstra and Hochmair 2011, Graser et al. 2013). Coverage in other parts of the world is more varied (Neis et al. 2013).

The additional dimensions of geographic and attribute space differentiate OSM from some of the other purely text-based crowdsourced projects such as Wikipedia and make OSM a particularly interesting case for study using visual analytics. Because the desire to share local knowledge plays heavily into OSM contribution patterns (Budhathoki 2010), the physical location of the contributors is perhaps of greater relevance than with other crowdsourced projects. Additionally, the spatial nature of the data invites the introduction of a mapping component into the tool, placing Crowd Lens in the category of "geovisual analytics" tools proposed by Andrienko et al. (2007).

Crowd Lens is place-based, and (in the demonstration of its capabilities reported upon here) incorporates a view of six pre-processed towns of roughly the same size from six different continents. The tool provides a manipulable overview of the contributor set behind each place, while also enabling detailed qualitative inquiry into each individual contributor. All the information displayed directly in Crowd Lens is derived from the publicly available OSM history files, although links are provided to contributor profile pages where these exist. A demo of Crowd Lens can be viewed at http://www.sterlingquinn.net/apps/crowdlens/index.html.

[16] Recent numbers of registered and active users in the OSM database are regularly reported at http://wiki.openstreetmap.org/wiki/Stats.

The main goal of Crowd Lens is to help geospatial technology professionals to make more informed decisions about whether to adopt (or continue using) OSM data. Because accurate and comprehensive spatial data often takes an extraordinary amount of time, money, and human resources to collect (learned the hard way by Apple Maps during its rocky initial launch), the maturing free and open source OSM database has begun to appear more attractive to businesses as well as governments who want to supplement or replace their existing spatial data in a relatively low-cost manner. Examples of these adoptions of OSM include a government-backed transit authority offering OSM base maps for its route planning application (McHugh 2014), a cloud-based web mapping company offering a variety of custom-styled OSM base maps (Barth 2015), and a location-based services company offering a mobile route-finding application using OSM street data (Van Exel 2014). In each of these cases, OSM plays a critical role as a free alternative to other maps that may be expensive or less flexible with their ability to make changes to the data. Furthermore, OSM may be the only mass spatial data option in places where governments do not offer maps and commercial mapping companies such as Google have not found it economically compelling to maintain detailed spatial data.

Each of the above three real world applications would experience some added vulnerability from associating itself with the unorganized, or at best loosely organized, contributor set associated with OSM. Unlike Wikipedia, there is no hierarchy of privileged editors in OSM (other than a small “Data Working Group” within the OpenStreetMap Foundation) and no infrastructure for "locking" controversial features. Incoming data is not pre-checked by any kind of gatekeeper for geometric or semantic accuracy, and vandalism slipping into the project can be tricky to control (Ballatore 2014) as even Google has learned with its similar Map Maker feature (BBC News 2015). Although there are automated scans of OSM data by bots and other programs looking for logistical flaws and other anomalies, one of the most effective bug-prevention mechanisms might be a large and active contributor set that will quickly

spot and fix problems in the data (Raymond 1999, Haklay et al. 2010). The intent of Crowd Lens is to reveal the size, activity, and interests of the OSM contributor set in a given place, while also offering an exploration-friendly overview of any anomalies in data contribution including imports, bot activity, and vandalism. Another goal of the tool is to get beyond a simple "node counting" approach to quantifying OSM data to expose more qualitative information about the contributions, such as the free-form comments, or "commit messages", left by contributors when saving their work in the project. Thus Crowd Lens accommodates the visual analytics mantra of Keim et al. (2008a), facilitating an initial overview analysis of the data, followed up with more detailed filtering and retrieval of details about more specific items of interest (in this case, the activities of individual OSM contributors).

The remainder of this chapter develops as follows. We first provide some more thorough background about the intentions of visual analytics research and its potential application to (a) geographic questions and (b) any projects of a crowdsourced nature. We then describe the design and development of Crowd Lens, while discussing how the tool addresses particular needs of analysts studying OSM and, to an extent, crowdsourced information projects in general. We then explain three evaluation methods that were undertaken to guide the development of Crowd Lens and assess the tool's effectiveness at augmenting OSM users' understandings of the data. We discuss the results of those evaluations, then offer some of our own insights about OSM data that we learned through the use of Crowd Lens. We conclude with remarks about potential future directions of development for Crowd Lens and similar tools.

Relevant literature: What visual analytics offers OpenStreetMap studies, and vice versa

OSM falls into a category of spatial data often called volunteered geographic information (VGI), which allows individuals who may have little traditional training in cartography or

geographic information systems (GIS) methods to still participate in the creation of spatial data (Goodchild 2007). In many cases, VGI contributors supply hyperlocal information that they alone can provide or know. When broadly considered, the term VGI can encompass phenomena as diverse as crowdsourced traffic speed databases, geotagged social media posts, and citizen engagement "Report a problem" apps (for example, the Federal Emergency Management Agency Mobile App "Disaster Reporter" feature that enables citizens to upload disaster photos). [17] Considering the fundamental differences between these many flavors of VGI, Haklay (2014) has proposed that OSM has become a large and unique enough phenomenon that it should be considered as its own strain of inquiry within VGI research, which he names "OpenStreetMap studies". Areas of investigation within this proposed OpenStreetMap studies includes the trustworthiness of OSM, OSM's use as a big dataset in other strains of computing research, the completeness of the OSM data, societal impacts of the data, and social practices in OSM contribution. Considering the abundant archives of OSM data history that are available (including all geometry and attribute changes with their associated timestamps), visual analytics has much to offer to these various pillars of OpenStreetMap studies. This includes inquiries into the nature of the crowd behind the project.

Visual analytics emerged in the same time frame as the popularization of the term "big data", and was fueled by the need to make sense of constantly incoming streams of text, video, images, and other data sources, often in real time. Early research agenda materials on visual analytics aimed to support the US homeland security sector following the 9/11 attacks, by focusing on how further such events could be anticipated, or at least mitigated, using abundant sources of digital information (Thomas and Cook 2005). Since that time, visual analytics has been applied in many other domains, such as understanding large corpora of news stories and social media posts (Dou et al. 2012), cycles of geopolitical events (Peuquet et al. 2015), histories of

[17] http://www.fema.gov/mobile-app

judicial decisions (Collins et al. 2009), and so forth. Andrienko et al. (2007) promoted the incorporation of mapping and location awareness components into visual analytics to help understand data with a spatial component. "Geovisual analytics", therefore, recognizes that much big data is associated with location coordinates or geometries, and that geographers may therefore be well suited to addressing big data analysis tasks (Burns and Thatcher 2014).

From the start, the process of human analytical reasoning formed a key part of visual analytics research, and remains a focal topic. Various models have been proposed to describe how the human brain retrieves information from a visual interface, stores that information for further use, and then connects it with other pieces of information in a process of knowledge construction (for example, MacEachren et al. (1999) provide one early articulation of such a model prior to visual analytics being coined as a label for work combining visualization and computational methods). Many of these models are iterative in nature, incorporating repeated "loops" of foraging for information, evaluating the usefulness of that information, and synthesizing the information to achieve new knowledge or "sense making", a process whose success is often influenced by the tacit domain knowledge of the analyst (Card et al. 1999, 10; Pirolli and Card 2005; Sacha et al. 2014). Thus nearly all visual analytics tools support some level of user interaction and iteration (Thomas and Cook 2005). Burns and Skupin (2013) opine that this focus on knowledge construction and analytical reasoning may be the main component separating geovisual analytics from traditional geovisualization; however, other important focuses of study that distinguish visual analytics and its subdomains from other research include computational methods, analytic methods, human-computer interaction (HCI) with visual interfaces, and dealing with heterogeneous information sources (Keim et al. 2008b, Ribarsky et al. 2009).

Visual analytics tools for crowdsourced data

On the topic of crowdsourced data, visual analytics (as well as "traditional" information visualization) have been applied toward both understanding the data generation process and making sense of the resulting data. Here the focus is on the former, representative examples of the latter include Chan et al. (2008), Endert et al. (2013), Fuchs et al. (2013), Cho et al. (2016), and Dou et al. (2015). Wikipedia in particular has received much attention from researchers using visual methods to understand crowdsourcing processes. A single Wikipedia article may be constructed by dozens of contributors, each with their own purposes for modifying or augmenting the article text. Viegas et al. (2004) created a history flow visualization showing how these individual efforts cause the Wikipedia article to take shape over time. The edits of each contributor appear in a different color and are stacked in a column representing the position of the text within the article. As these columns connect across a horizontal timeline, they expose periods of deliberate vandalism, the introduction of new ideas, and conflict between contributors.

The idea of controversy and "edit wars" in Wikipedia has also received attention from visual analytics researchers. For example, Borra et al. (2015) developed a method to identify and visualize the most contentious passages of an article. Suh et al. (2007) focus more on the relationships between editors in controversial Wikipedia topics, attempting to identify the most active and polarizing editors. They employ a force-directed graph layout algorithm to identify groups of editors most prone to delete each other's work. The tool also reveals clusters of mediators and anti-vandalism bots attempting to increase the quality of the articles.

Brandes and Lerner (2008) built a similar tool that examines edits to controversial Wikipedia articles and places all contributors on an ellipse based on their position in the debate. Contributors who tend to undo each other's edits are pulled to opposite sides of the ellipse. Contributors themselves are symbolized by ellipses whose sizes, orientations, colors, and

connecting lines represent aspects of their involvement in the controversy. The analyst can define a slice of time to filter the graph. This is helpful for understanding the evolution of the discussion, and identifying steady contributors versus those who are active for only a short period.

These tools are designed for analysts with some domain knowledge of Wikipedia contribution patterns, bots, and edit wars. However, Boukhelifa et al. (2010) describe a "skin" for Wikipedia that allows the casual user to gain a quick idea of the amount of attention and controversy an article has generated and make inferences about the quality of the content. A left hand panel on the skin depicts metrics of the number of contributions, the timeline of the edits, the number of words on the discussion page, and so forth.

A smaller, but still useful, body of research employs visual analytics to study OSM data contributions. Some inquiries have focused on developing cartographic renderings (whether static or interactive) for OSM metadata. These include "version contour lines" (Van Exel 2011a), a "temperature" map of OSM community attention (Van Exel 2011b), and maps of the time since the most recent edit of any feature (Barron et al. 2013). In the realm of interactive cartography, Roick et al. (2011, 2012) used hexagonal cells to aggregate OSM metadata such as the date of most recent feature edit, the average version number of a feature, the number of points of interest, and so forth. These maps are bundled in an interactive tool called OSMatrix, available to browse online over the European continent.[18]

Users of OSMatrix can discover which areas of the map have been most actively edited (presumably indicating superior quality) and which areas have laid dormant. The tool reveals stark variations in user activity across some international boundaries, as well as more localized anomalies such as a region of numerous edits east of Paris. Countries such as France and the Netherlands show high volumes of edits and data imports. Other countries such as Spain and Portugal show fewer edits, although activity is strong in urban centers, coastal regions, and

[18] http://koenigstuhl.geog.uni-heidelberg.de/osmatrix

transportation corridors. OSMatrix indicates that map editing patterns may follow general online map viewing patterns noted by Fisher (2007).

In another visual analytics approach to exploring OSM contributions, Trame and Keßler (2011) submit an interactive tool to display which geographic entities within the current map view have been edited the most. The color of a feature varies depending on the number of edits it has experienced, and a heatmap approach is taken to create a partially transparent surface of edit intensity. “Hot” areas of the map could identify controversial regions, or they could simply indicate high interest in the area if editors are simply adding more attributes to existing features.

Exploratory visualizations of OSM data and metadata can also help with identifying data anomalies and errors. The OSM Inspector[19] tool by Geofabrik presents a navigable map with an overlay of detected problems such as "self-intersecting ways" and "duplicate node in a way". These can be clicked to retrieve more information about the feature in question and lead the analyst toward a decision about whether and how to fix the issue.

The visual analytics tools for OSM described above are mostly concerned with the characteristics of the geographic data and rate of data production; whereas many of the Wikipedia tools include analysis of the people who contributed the data in an attempt to understand more about article credibility and the motives of the authors. Exploratory visualizations of individual OSM contributor work and habits have been limited; some promising avenues for identifying the most active contributors with local knowledge were proposed by Napolitano and Mooney (2012) but not developed into an interactive tool. The Crowd Lens tool described in this chapter takes OSM visual analytics into the human realm, allowing an exploration of the composition and activity patterns of the OSM contributor crowd. Users of geospatial technology, as well as those interested in crowdsourced data assembly more generally are invited to try Crowd Lens and

[19] http://tools.geofabrik.de/osmi/

consider how the activities of the crowd might affect the credibility and suitability of the OSM data.

Design and development of Crowd Lens

Crowd Lens is an interactive tool that runs in a web browser. In a single display containing multiple linked views, it provides a filterable overview of the OSM contributor crowd behind any given place, as well as a range of drill-down options to explore detailed aspects of the data and the actions of any one selected contributor. The Crowd Lens user interface is shown in Figure 4-1.

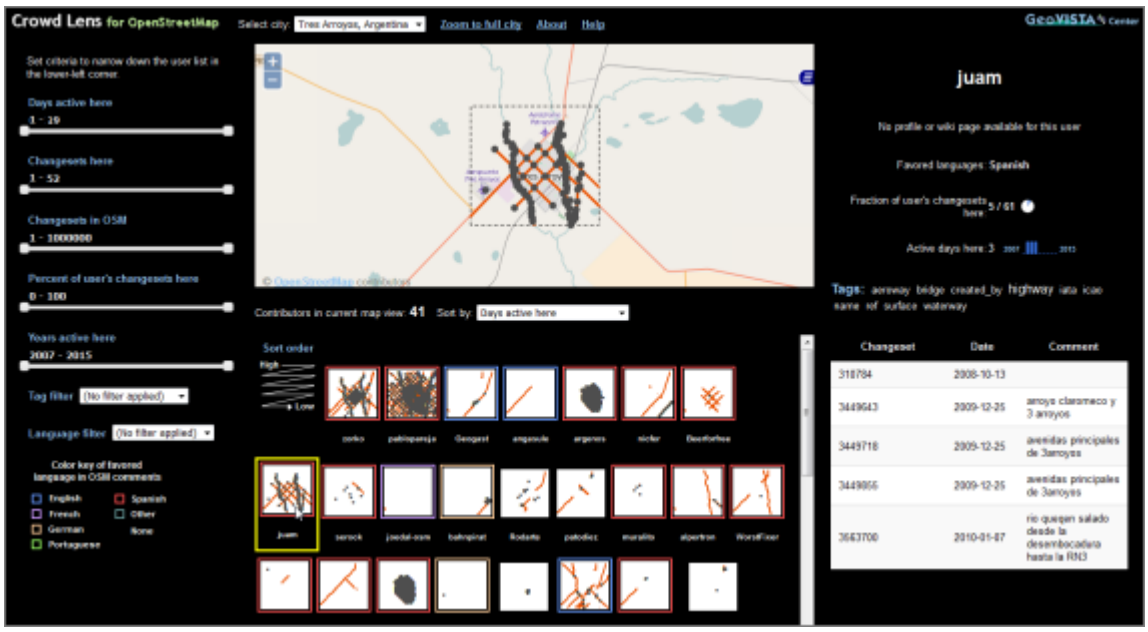


Figure 4-1. Crowd Lens user interface.

Authors of Chapter 2 in the first visual analytics research agenda identify three main types of tasks associated with visual analytics tools: Assess, Forecast, and Develop Options (Thomas and Cook 2005). Crowd Lens primarily falls into the first category as an assessment

tool, designed to help the analyst conceptualize what has happened with OSM up to the current point in time in any particular place; however, the findings could facilitate educated guesses about what OSM contributor trends may continue in the near future (thus supporting at least rudimentary forecasting). Findings from Crowd Lens might also help with developing strategies for filling in sparsely represented places in OSM with a greater variety of contributors and mapped entities, or for dealing with various kinds of bias identified in the data.

Crowd Lens is a data exploration tool that, in addition to finding answers to questions about OSM development, also helps facilitate new questions, a goal articulated for visual analytics more generally by Bivand (2010). For example, a Crowd Lens user might notice that some of the same OSM contributors show up in different cities on opposite corners of the globe, leading to further investigation about the origins and influence of "power users" on the worldwide map. A foray into these user profiles might reveal mappers who are being paid to improve the data on behalf of a company that relies on OSM for its spatial services (see Barth 2015 for an example), leading to additional inquiries and insights about the political economies that take shape around free and open data and how those affect all data users.

Crowd Lens evolved from a set of static maps and other graphics derived by Quinn (2016) examining OSM contributor patterns in small cities. In that study, a series of abstracted small multiple maps was created to compare the relative amount of activity between OSM contributors in a bounded place. It was observed that abnormal contribution patterns such as automated imports, bot activity, and exceptionally active users jump out of these graphics and invite continued investigation. To enable a deeper analysis of the data in multiple cities, it was desirable to link these graphics with an interactive map of the contributions and the free-form comments linked with each OSM contribution. Additional contributor metadata could also be summarized and displayed together in an effort to create the best-rounded profile of a contributor possible using publicly-available data.

Crowd Lens is built with the OpenLayers toolkit, an open source library for JavaScript-based development of geospatial web applications. The data for the visualization comes from two large archive files: (1) the OSM "full history dump" (48 GB compressed), and (2) the OSM changeset history (1.3 GB compressed), both of them processed from the December 28, 2015 release of the data, the final archives in that year. The OSM history dump contains the geometries and attributes (ie, tags) of current and past geographic features contributed to the OSM database. Spanning back beyond 2007 in some cases, it captures most (but not all) of the features ever recorded in OSM.[20] Similarly, the OSM changeset history is a record of how these geographic features connect to different contributors' efforts in the project. A changeset represents one session of work produced by a contributor. A changeset is created each time a contributor clicks the option to "Save" his or her edits in an OSM editor, thereby pushing all modified features into the OSM database.

When adding a changeset to OSM, the contributor can type a comment describing and justifying the edits made. The corpus of changeset comments provides qualitative information about contributor motives and habits not previously linked to any interactive map visualization that we could find. It also gives a clue as to the preferred languages of each contributor, which were derived for visualization in Crowd Lens using the langid.py Python language identification module (Lui and Baldwin 2012, Quinn in press).

The top menu bar of Crowd Lens allows the user to select a city for analysis. To support the user evaluations reported here, six small cities were selected from different continents. They are Hereford, United Kingdom; Hervey Bay, Australia; Johnstown (Pennsylvania), United States; Kadiri, India; Suhum, Ghana; and Tres Arroyos, Argentina. These are non-suburban cities of regional importance. The rectangular study areas surrounding each city were positioned to

[20] Some features missing from the history files include very early items contributed to the project, as well as items redacted from the database due to some contributors' refusals to accept a license change in 2012.

contain roughly 50 – 100 thousand inhabitants. This choice of cities builds on Quinn's (2016) inquiry into small cities as a more telling barometer of overall status of OSM in a region than might be observed in large metropolises; in this way it complements the work of Neis et al. (2013) who made a comparison of OSM activity in different world regions using major urban centers as a point of focus.

The main panel of the Crowd Lens interface displays the contributor activity in the city and is divided into four user interface components that are intended to be used in a left-to-right fashion: the crowd filters, the main map, the contributor list, and the individual contributor panel. These four components are described below beginning with the contributor list, which is the crux of application activity and the center point of focus of the tool.

Contributor list

The contributor list (Figure 4-2) conveys the size of the contributor crowd active in any one place, while giving an overview of each contributor's relative amount of influence on the OSM map contents. It also allows the selection of any one contributor to learn further details.

The contributor list is comprised of small multiple map images giving a geographic overview of each contributor's work in OSM in the selected city. The maps can be sorted from top to bottom alphabetically or based on activity criteria such as the contributor's number of unique days of activity in this city, number of changesets the contributor made in this city, number of changesets the contributor made in the OSM project, and percent of the contributor's changesets in the OSM project that occur in this place.

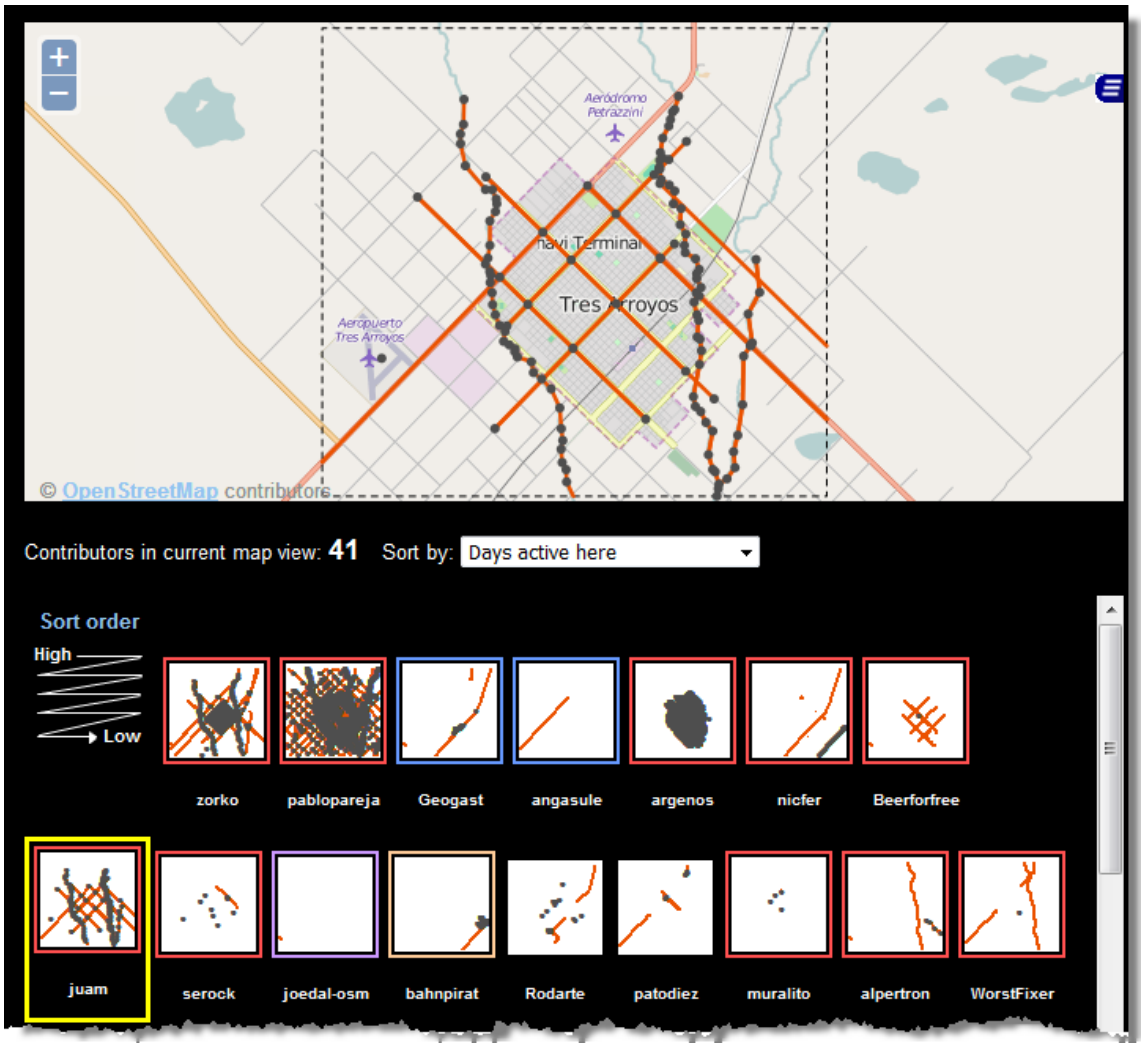


Figure 4-2. Contributor list and interactive map in Crowd Lens

The small map images are surrounded by a border of color representing the contributor's preferred language in OSM changeset comments. Cities with a higher percentage of contributors speaking the native language are expected to contain more map features that would be of value to the daily routines of the city's residents (Quinn in press). The languages reported in Crowd Lens were detected automatically by the langid.py Python module. In some cases, the preferred language determined does not match that of the contributor's changeset comments displayed in Crowd Lens. Although no language identification software is perfect and miscodings are possible with langid.py (Quinn in press), we have found that a much more common cause of these

mismatches is when a multilingual contributor tailors his or her language to the place being mapped. For example, a German contributor might use German when editing Germany, but may switch to English when editing other countries to better communicate with a more international contributor set. In this situation, Crowd Lens might report the contributor's preferred language as German, while the displayed comments in the study area are all in English. The inclusion of all the contributor's OSM comments in the calculation (rather than just the ones in the study area) has the effect of identifying contributors who are less likely to be local to the area.

Crowd filters

The Crowd Lens contributor list is displayed unfiltered at first, in other words, in contains all contributors found in the OSM history files in this city. To better understand and isolate subsets of the crowd for further analysis, Crowd Lens offers multiple direct manipulation filtering tools. For example, the contributor list can be narrowed down by panning and zooming the map, as only the contributors who made any edits in the current map extent are shown in the user list. The list can also be narrowed by adjusting the crowd filters on the left hand side of the interface.

The crowd filters (Figure 4-3) allow a customized view of the contributor list to display only the OSM contributors that meet certain thresholds or criteria of contribution activity. The filters allow for dynamic queries of the contributor list (Shneiderman 1994), in other words, the ability to adjust the query inputs in a point-and-click fashion and see immediate feedback. For example, to get an idea of the number of active contributors in a city, the crowd filters can be used to narrow down the contributor list to just those who were active more than five days in the city and made at least one edit in the most recent year of the project.

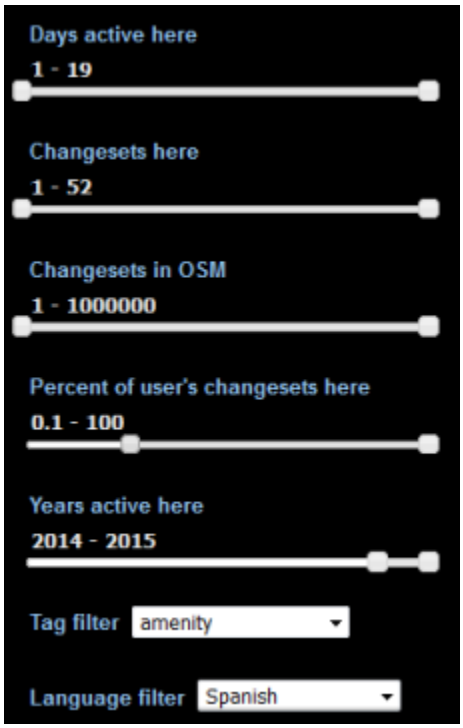


Figure 4-3. Crowd filters in Crowd Lens

The crowd filters include slider bars allowing the Crowd Lens user to set high and low boundaries of these statistical thresholds. In some cases the slider ticks are linear in nature, in other cases they are logarithmic to more easily navigate values that could become very small (such as percentage of a contributor's total OSM edits falling within this particular city). The sliders can also set ranges for the contributor's number of unique days of activity in this city, number of changesets the contributor made in this city, number of changesets the contributor made in the OSM project, and percent of the contributor's changesets in the OSM project that occur in this city.

Two other filters are available through dropdown lists. One restricts the contributor list based on preferred language in the OSM changeset comments. For example, this filter could be used to see all users whose most often-detected language in the comments was Spanish. A second dropdown filter narrows down the list to only contributors who used a particular tag. For

example, this dropdown could be used to display only the contributors who added street addresses in this city.

Map

The interactive map in Crowd Lens serves two functions. Most importantly, it shows the work of the selected contributor in a more detailed fashion than can be obtained by looking at the small map images in the contributor list. The OSM geometries produced by the contributor are overlaid on the current OSM map tiles for geographic context. A Crowd Lens user can click one of the geometries and see all associated features in the changeset highlighted. The corresponding changeset comment for that feature is also highlighted in the individual contributor panel (described below).

As mentioned previously, the map also acts as a filter for the contributor list; therefore, panning and zooming the map offers a way to see the crowd size in any particular neighborhood or other sub-geography of the city.

Individual contributor panel

After applying the crowd filters and/or examining the contributor list, it may be desirable to learn further details about particular contributors, due to factors such as an extraordinary number of edits by that contributor, or a dense cluster of edits in a particular neighborhood of interest. Therefore, the Crowd Lens user can click any small map image in the contributor list and see a description of that person's work displayed in the individual contributor panel. This panel contains statistics such as the selected contributor's preferred language, number of changesets (in this city and in all of OSM), and years active in this city (Figure 4-4).

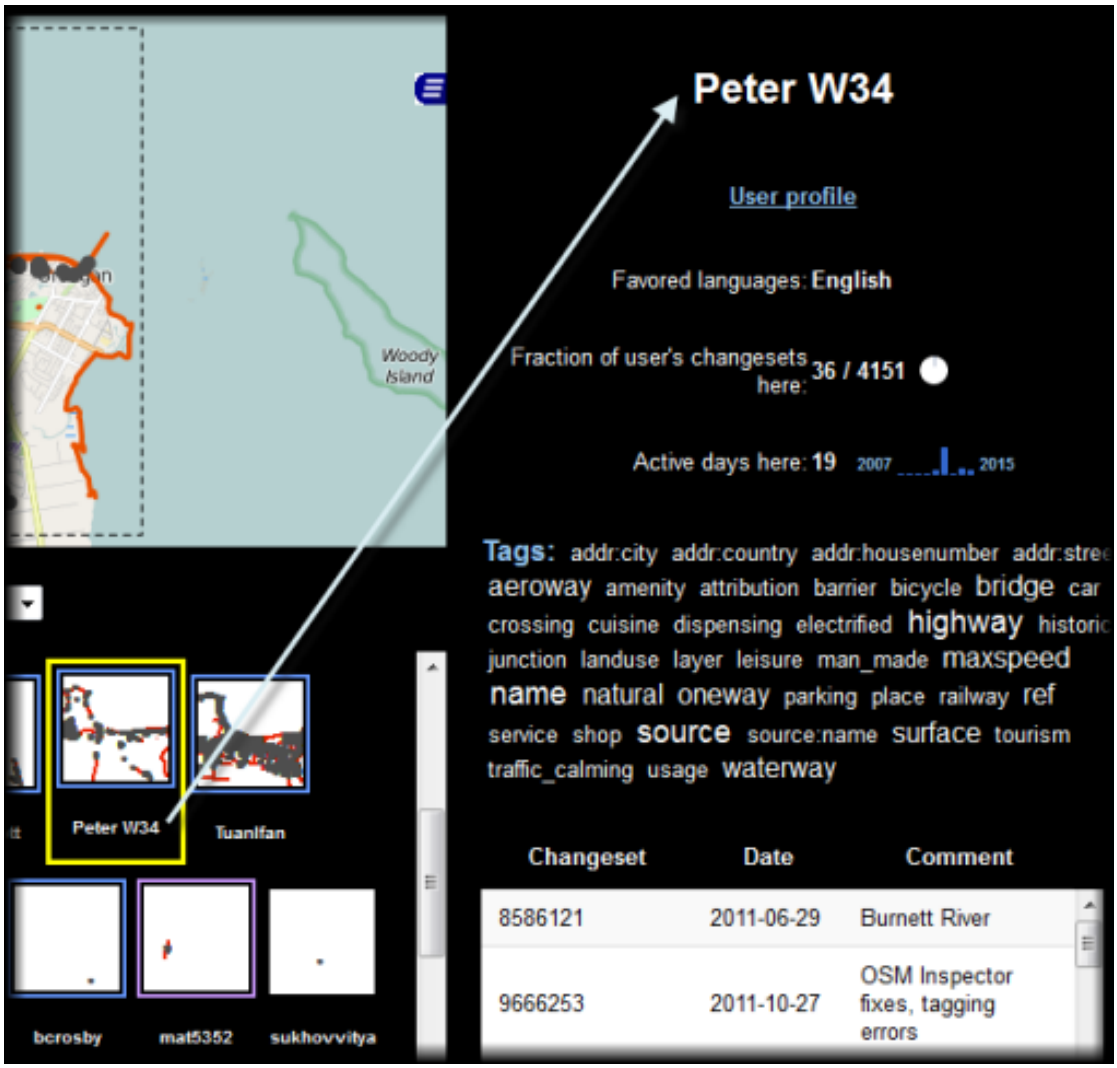


Figure 4-4. Interaction between contributor list and individual contributor panel

The individual contributor panel also offers a "tag cloud" showing all tags applied by the selected contributor, with more heavily used tags appearing in larger font size. This helps identify the types of entities that the contributor tends to add (and does not add) in OSM. The tag cloud also serves as a crowd filtering mechanism, as clicking any tag will narrow down the contributor list to show only the people who applied that tag.

Beyond the raw statistics of how much content the contributor added to OSM, the individual contributor panel also offers qualitative information that can help the analyst get a more well-rounded feel for a contributor's story and role in relation to the OSM project. To this

end, the individual contributor panel offers a data table showing all changesets produced by the selected contributor. The table reports the changeset date, identification number, and the free-form comment that the contributor submitted upon saving the changeset. This table is interactive, such that clicking any record highlights the corresponding changeset items on the map, and clicking an item on the map highlights the corresponding record on the table.

The individual contributor panel only contains information that can be derived from the OSM history files. Sometimes further information about a contributor is available in publicly-available OSM profile or wiki pages that he or she has created. When such pages are detected to exist, the individual contributor panel contains hyperlinks to them. In some cases these pages contain the city of origin, the profession of the contributor, and notes about special projects, hobbies, or editing habits undertaken by the contributor.

Using the components together

Taken together, the user interface items in Crowd Lens provide overviews of crowd characteristics behind a place in OSM, while also enabling a deeper qualitative study of the individual interests and motives each contributor brought to the project, and how these shaped the map over time. Crowd Lens is designed such that end users can follow Keim et al.'s (2008a) visual analytics mantra[21] of "analyse first – show the important – zoom, filter and analyse further - details on demand." The initial city-wide map and contributor list present an overview suitable for initial analysis, which can then be filtered by zooming and setting filters to see the most important set of contributors for the analysis task at hand. Details of this list can be contained by

[21] This mantra is based on Shneiderman's (1996) information seeking mantra of "overview first, zoom and filter, then details-on-demand".

clicking the individual contributor's small map image and seeing the contributor's full activity metadata.

Evaluation of Crowd Lens

The development of Crowd Lens followed a user-centered design process, wherein an iterative series of user studies was conducted to assess the tool's usability and utility (Slocum et al. 2003, Robinson et al. 2005). Roth et al. (2015) describe how a mix of "discount" studies using 5 – 10 participants during the formative stages of the tool can lead to more positive changes than a single large summative study executed after the tool's deployment. Each evaluation cycle generates feedback and lessons learned that can improve the product usability and utility.

The Crowd Lens evaluations included (1) a usability test early in the development cycle to assess the intuitiveness and basic functionality of the user interface, and (2) an evaluation by geospatial technology professionals of a later version of Crowd Lens, designed to assess the utility of Crowd Lens in their work and learn how the tool affected their confidence in the use of OSM data. In between these two rounds of evaluation, a scenario and claims analysis exercise (Rosson and Carroll 2002) was conducted to help with prioritizing remaining development tasks and identifying which potential new items of functionality would be most relevant to end-user workflows. Beyond helping to improve the effectiveness of Crowd Lens, these evaluations helped generate more general insights about designing visual analytics tools for understanding the construction of large crowdsourced information products.

Early usability testing

An early development version of Crowd Lens was tested for usability by seven participants consisting of graduate and undergraduate students in the social sciences at a major US research university. Participants were required to have some previous exposure to GIS and OSM because the primary target audience for Crowd Lens is professionals who are involved with making decisions about if or how to use OSM in their work. Thus, it was important for the initial assessment (even though only focused on interface usability) to include participants with prior knowledge about OSM and its use in GIS applications.

The testers were asked to use Crowd Lens to answer a series of objective questions pertaining to facts about the OSM data. The questions required switching between the different cities offered by Crowd Lens, using the sorting functions on the contributor list, and reading statistics from the individual user panel. (At the time of this test, the crowd filters did not exist.) Examples of questions include, "Which contributor to 'City X' has over 10,000 changesets in OpenStreetMap?" and, "Which of the cities had the most users active over 10 days in the project?" As testers attempted to answer these questions with the tool, their interactions were captured by screen recording software. Think-aloud and looking-over-the-shoulder approaches were deliberately avoided in order to minimize any interference with the users' natural reasoning processes and computer interaction habits (Ribarsky et al. 2009, Dou et al. 2010).

For each tester, the Crowd Lens session was ended after a maximum of 15 minutes. This was followed by a semi-structured interview wherein testers were each asked to identify the most difficult and easy tasks for them and comment on the user interface components that facilitated or impeded their work. Testers were also invited to suggest any additional features they felt might be useful in the tool, given their own experiences with geospatial applications and OSM.

Purposes of this test were to evaluate how easy or difficult it was for users to find functionality in the user interface, as well as to solicit some early feedback about features that might be added to the tool. The test also showed which functions were most commonly used and which were ignored, although in this type of test the functions invoked are influenced by the question set.

Results of early usability testing

In general, testers were able to figure out the answers to the questions without any assistance or previous exposure to the user interface. Five of the seven testers got all seven questions correct within the allotted time period. One tester missed a single question and one tester missed two questions. Testers commented that they liked the ability to sort the small map images, and that the colored border of the map images was helpful for quickly getting a view of the contributors' preferred languages.

The most common items of feedback were that not enough small map images were visible on the screen at a time, making it difficult to navigate through the images or get an overview of which users were predominant within a city. In response to this, the percentage of screen space devoted to the user list was increased and the size of the map images was decreased in future iterations of the tool.

Other users commented that the slow responsiveness of the tool impeded their interpretation of the display. This problem was especially pronounced in the larger cities. It was eventually mitigated by developing an alternative data structure in which the OSM data were stored in and displayed from separate files for each user, rather than filtering a single large file for the city itself. This made the application more scalable. An animated "loading" graphic was also added to give users a visual cue about when they should begin to interpret the map results. This

graphic becomes visible only when the tool must load large amounts of new data, such as when the displayed city is switched.

Other tester suggestions or observations eventually added to Crowd Lens included:

- Making the tag cloud into a clickable filtering mechanism. The testers liked the tag cloud and thought it effective for seeing the contributors' domains of work, but many of them tried to click the original static version with no result.
- Adding an alphabetical sort option to the user list
- Adding an explanatory graphic at the top of the user list explaining that the small map image sorting occurs from left to right, starting at the top as do the words on an English-language page

Scenarios of use and claims analysis

The usability testers submitted more suggestions than could be reasonably implemented in the available development time frame, some of them which seemed only peripheral to the tool's purpose for answering questions about crowdsourcing. To understand which features would be most helpful for accomplishing the tool's objectives and prioritize them for development, we undertook a thought exercise called a "scenario of use". Rosson and Carroll (2002) describe scenarios of use as a way to form a design by telling a story about how an end user is anticipated to interact with a tool's user interface (in their words, "narrative descriptions of envisaged usage episodes"). Critical user interface features and their interaction points in the scenario can then be evaluated through a "claims analysis" identifying the pros and cons of the design. See MacEachren et al. (2008) for an example of a scenario of use and claims analysis applied to an interactive, exploratory geovisualization web application.

Although scenarios of use are more often implemented in the early stages of design, they became important at the evaluation phase of Crowd Lens in order to keep the project focused on the core objective of enabling analysis of the crowdsourced process of OSM creation and the identification of geographic variations in that process. In our situation, enough time remained in the development cycle to make substantial changes, thus it was felt that it was not too late to implement the scenario of use. In particular, developing the scenarios of use and the claims analysis helped to guide subsequent system refinement as well as design of the utility evaluation outlined in "Testing by geospatial technology professionals" below.

Below are two scenarios of use describing hypothetical interactions with Crowd Lens by two geospatial technology professionals with distinct use goals. Although the scenarios are fictional, the contexts of OSM use described therein were influenced by a range of formal and informal input collected at multiple industry conferences such as State of the Map and FOSS4G.

Scenario 1: Municipal web mapping of community services

Dan is a GIS analyst with some beginning level web development skills. His job is to create online web maps for the (fictional) city of Oysterville. These maps allow citizens to find city services such as community centers, voting locations, and schools, with the option of getting driving directions to them. In these interactive maps Dan has been relying on Google Maps as a base map to provide geographic context, but Dan recently became aware of OpenStreetMap and is considering the potential effects of switching the default basemap to OSM. He has browsed the OSM basemap in his town and feels that it might have the potential to include more detailed features recognizable to residents (e.g. local landmarks and businesses), while offering the ability to make fast updates and fixes. Furthermore, he and his team lead like the idea of not having to

worry about staying within the legal or financial bounds of any current or future license agreement with Google or other institutions.

At the same time, Dan is aware that there are fewer than 100,000 residents of his city and wonders how many of these people even know about OSM or would have the time and training to fix and enrich the map on a regular basis. Dan knows that the time and scope of his day job will prevent him from making frequent edits to OSM himself. In this scenario, he is interested in knowing (1) how many people are actively contributing to OSM in his town, (2) what percentage of these contributors would likely have knowledge of local features that could only be obtained by an on-the-ground visitor, such as road or business names, (3) what these likely-local contributors tend to work on mapping in OSM, and (4) what kind of information is being added by the rest of the contributors. These questions could give him a feel for how the local territory is influenced in OSM.

Dan opens Crowd Lens and selects his town from the dropdown list. He notices the familiar OSM basemap of his town, overlaid by nodes and ways from what Dan determines to be one contributor. He guesses this because of the set of square images along the bottom panel, showing each contributor's work in the town. The first image is surrounded by a yellow frame, indicating that that contributor is selected, and the small map within the frame matches the same pattern of nodes and ways Dan sees overlaid on the main map of OSM.

In the right-hand panel Dan sees some information about this contributor, including a link to the contributor's OSM profile page, the number of changesets offered by this contributor in just this town and in the OSM project as a whole, the number of days this contributor worked on the town, and a sparkline graph of the years this contributor worked on the town. The contributor's favored language is also listed, which is English, the language predominantly spoken in Dan's town (although the town has a sizable minority of Spanish speakers). Below all this is a table showing the changesets made by the contributor and comments associated with

each. Dan discovers that if he clicks one of the small map images in the lower panel, he can change the currently featured contributor in the right hand panel.

From this initial view, Dan learns that Crowd Lens gives him detailed information about a single contributor, but he is more interested in some of the general contribution patterns and the overall size and nature of the crowd. He scans the interface for these broader indicators and notices that the number of contributors to the total map view is reported on the screen. This is 76, informing Dan that 76 people have worked on the project here. Dan notices a series of filters allowing him to narrow down this set of contributors even further using different criteria.

Dan wants to take a guess at how many of these contributors out of the 76 are local to his town, so he uses one of these dropdown lists to filter the contributors to those whose preferred language is English, postulating (but not knowing for certain) that most of the contributors in his town will choose to employ that language when editing OSM. The number of contributors is subsequently reduced to 64. Using the remaining language filters he finds that most of the remaining contributors favor the German language, not Spanish as he initially expected.

Dan then sorts the English-favoring contributors by % changesets in the target area and notices that only a few contributors have more than 10% of their changesets in his town. Most of these contributors only made one or two changesets in the project. Dan concludes that they are very likely local (or have visited the local area personally), but are not active contributors. On the other end of the spectrum, some of the contributors with the lowest percentage of changesets in the target area are very active in OSM as a whole, but Oysterville is a place they've only edited once or twice. Scanning the changeset tables for a few of these active contributors, Dan notices that the comments they attach to their edits typically reveal that they were making data fixes in the town, either randomly through applications like MapRoulette or systematically as part of a regional effort to clean up OSM tags to conform with established OSM standards.

In his effort to find active local contributors, Dan then decides he wants to filter out contributors who only made a few changesets, and contributors with a very low percentage of their changesets in Oysterville. A series of slider bars allows him to set this filter. Ultimately Dan's query displays contributors active more than 5 days in this town who made more than 10 changesets and over 50% of their edits in this town. This filters out everybody, meaning there is no one who meets the criteria. Remembering the logarithmic scale on the slider bar for the percentage of contributor's total edits in Oysterville, Dan adjusts the 50% threshold down to 10% and gets four contributors as a result.

Dan wants to explore these four contributors in more depth, so he clicks their small maps in the lower panel to examine detailed metadata about each. While doing this, he notices one of them was not active in the project recently, so he adds another filter to show only contributors active at least once in the past year. This narrows the list to two contributors. Dan notices from the tag clouds in the right hand panel that one of the contributors only fixes roads (and may still very well be an armchair mapper tracing imagery), while the other tends to do a variety of tasks ranging from fixing tag structure to adding new points of interest, to changing street names. Dan decides that this latter contributor has almost certainly spent extensive time in Oysterville and is the only contributor who can be guessed with a high degree of confidence to be an active local mapper. Dan clicks the link to the user's wiki page and finds that this contributor belongs to an OSM community group of users from Dan's state (ie, "Users in California"), but has not reported any further location information.

After using Crowd Lens to explore the nature of the crowd behind the Oysterville map, Dan continues to have some mixed feelings about using OSM as a background in his own projects. On one hand, he knows there is probably at least one local mapper who is liable to continue adding rich geographic information that may be unattainable in mass produced commercial maps; however, Dan is somewhat disappointed to see that there is not an active local

community of contributors focused on improving the Oysterville map. He realizes that if he decides to adopt OSM, his office may need to devote some time and resources to help instantiate such a community.

Scenario 2: Assessing road data quality for a non-profit organization

Rosa works for a non-profit organization that distributes aid to tropical communities affected by severe seasonal flooding. Part of her work involves assessing which communities are most vulnerable to this type of natural disaster, and then determining the quality of the maps and routes that her crews would use to reach these areas in the event of an emergency. After two years in this job, it has become apparent to Rosa that commercially produced maps are lacking sufficient detail in some of the smaller towns and cities in the region where she works.

At a professional conference, Rosa hears that OSM sometimes has better coverage than other online maps and she wonders if it may prove to be a more reliable alternative for her line of work. She investigates some of the cities in her region on OpenStreetMap.org and notices that they generally suffer from a mismatch in coverage levels between large cities and small cities, similar to what she observed in commercial maps; however, the road networks at least seem to be mapped out comprehensively in OSM.

It is important for Rosa's group to understand the quality of motor vehicle routes in and around the town, therefore Rosa is interested in knowing when this road data was introduced into the OSM project and how actively it is being maintained. Furthermore, she would like to get some indication of how much mapping on the road network was performed by local contributors as opposed to being simply traced or imported by remote contributors who may have no direct local knowledge. She figures that local mappers would be more likely to keep the map updated in the event of roads being washed out, closed, or re-routed, and they will be more immediately aware

of changes to roads generated by local construction projects, new housing developments, and so forth. On the other hand, if the data was placed there a long time ago by one person, it may be unlikely that new data will be added in the near future and Rosa may consider her alternate plan of eventually paying some of her field crew to gather and update the data.

Rosa decides that Crowd Lens would provide an opportunity to develop some answers to the above questions. She selects her city and notices that it has a total of 43 contributors. From the small maps in the lower panel, she can tell that some of these contributors only added nodes and did not do extensive work on the road network. At the same time, she notices from some of the contributors' small maps that they have edited many of the road lines.

Rosa clicks one of these contributors whose map shows road lines. She sees the associated edits appear on the map. She clicks one of the road lines and notices that a row in the changeset table in the right-hand side of the map has been highlighted. From the information in the table, she can tell that this data was added in 2011 and that its contributor commented that the changeset was a data import. While she is looking at the table, Rosa also notices from the tag list above that the word "highway" is very large. Rosa now remembers that the highway tag is used to classify all types of roads in OSM.

Rosa clicks the highway tag in the list, and immediately the user list at the bottom of Crowd Lens shrinks to show only the contributors who used this tag, in other words, all the contributors who worked on roads. The tool reports that this involves 27 contributors.

Rosa decides to explore the map some more and zooms in to an area on the edge of town where she knows the roads will be most susceptible to hazards such as landslides and long term closure. As she zooms in, the list of contributors in the bottom panel narrows to include only those who modified items in the current map view. Rosa further filters these by number of changesets and years active in the project to show only the road editors in this area who made

more than two changesets and had some activity in the past two years. This results in seven contributors.

Rosa is then able to focus her time on a more detailed investigation of these seven contributors. She sorts them by number of changesets in this area and then starts examining the contributor information systematically, starting with the one who has created the most changesets in the area. For each contributor, Rosa clicks the profile or wiki page link (when available) for clues about the contributor's place of origin. She also scans the table of changesets comments to understand any motives or source information about the contributor's edits. After studying these for a while, Rosa finds three contributors who either (1) mentioned in their profile or wiki pages that they were from the local town or (2) mentioned in the comments that they had done some local surveying (e.g. with GPS, bicycles, or field work papers). The rest of the mappers seem to be "armchair mappers" focused on tidying up tags, importing data, or tracing roads to make fixes and fill in blank space.

Rosa just used Crowd Lens to identify and investigate the subset of OSM contributors working on certain feature types and geographies most pertinent to her work. Rosa is encouraged to find that there is indeed local highway mapping activity going on in the area, and she leaves her Crowd Lens session with a more confident feeling about using OSM in her projects. At the same time she does not believe the three local mappers are enough to maintain all the road data, and for some of the most critical routes she still plans to send an advance field crew to verify the location and quality of the roads. She has decided to instruct this field crew to enter any adjustments directly into OSM so that they can integrate their improvements with the data already available and share it with others on an open platform.

Claims analysis and repercussions for Crowd Lens development

Following the development of the above scenarios, a claims analysis was performed to evaluate some of the advantages and drawbacks of critical aspects included in or planned for the user interface. The analysis is shown in Table 4-1.

Table 4-1. Claims analysis related to the Crowd Lens scenarios of use.

The ability to narrow down the contributor list through a series of filters...
+ Allows the quick calculation and further exploration of the subset of contributors who participated in a particular way
+ Gives the end user a way to explore which parameters affect the crowd size the most
- May get end users into a state where they forget they are only viewing a subset of contributors
- May tempt the user into getting overly interested in tweaking settings and controls and/or exploring the work by specific individual contributors, rather than thinking about the broader context of and patterns of contributions
Showing the crowd filters and the individual user details on the same screen...
+ Lets the end user immediately know that he or she can explore the general characteristics of the crowd as well as the individual nuances of each user
+ Cuts down on the clicks and "window management" that the end user needs to perform to open and shut interface components
- Makes the view of the screen more visually cluttered
- May be annoying for the analyst who solely wants to focus on crowd characteristics, or who is only interested in individual characteristics
Showing just a single user's contributions as a vector overlay on the map (rather than showing all the data)...
+ Cuts down on visual clutter in the map screen
+ Allows a direct link between clicks on map features and the user changeset table
+ Keeps map loading speed and responsiveness reasonable
- May mislead people into thinking that a single user's contribution actually represents all the available data
Representing each user with a thumbnail map image...
+ Gives the analyst a quick view of each user's relative contribution and how it compares with the other users' contributions
+ Gives a visual clue that the analyst might be able to change the main map by clicking the thumbnail
- Takes up additional space and makes it difficult to see all the users listed in the view at the same time
- May be initially confusing to those who do not understand that the image is showing a (very abstracted) map

At the time that these scenarios were created, the crowd filters (ie, the slider bars for narrowing down the user list) had not been developed yet. From these scenarios of use and claims analysis, it was decided that the filters would introduce many new possibilities for understanding subsets of the crowd, and that it would be wise to invest most of the remaining development time in implementing the filters.

It could be argued that initial development efforts on Crowd Lens might have been better directed had the scenario of use been created prior to any implementation; however, the feedback from the usability testing was valuable to have in mind when producing the scenarios.

Testing by geospatial technology professionals

A final round of testing assessed how insights derived from Crowd Lens might affect professionals' confidence levels in OSM data for use in spatial data services and applications. More generally, we also evaluated the utility of Crowd Lens at addressing the tasks and concerns faced by technical OSM users. This round of testing occurred after the improvements from the other studies above had been implemented and a beta version of the tool had been deployed on a publicly-visible demo server and informally tested by several geography students associated with the authors' university. We invited 10 geospatial technology professionals working with digital maps or GIS on a regular basis to try Crowd Lens. These testers were recruited from the authors' professional networks formed through years of participation in the geospatial industry and academia. They represented a broad range of domains including software development, environmental consulting, utilities management, municipal government, and others. To be considered for participation in this study, testers were required to have used OSM in some way to

support their mapping or GIS work (or to have considered using it), thus having some exposure to the OSM contribution and tagging practices.

The evaluation was conducted entirely online in a web browser. We summarize the instructions and questions below, noting that the full text of the evaluation instructions and survey is available in Appendix B.

First, a welcome screen presented an introduction to the research and information required by the institutional review board (IRB) for the informed consent process. Because the IRB judged the study to be exempt, testers' advancement to the next screen was interpreted as their willingness and eligibility to participate.

Before testers opened Crowd Lens, a short pre-assessment was administered asking testers about their previous experiences with digital mapping and GIS, and how they have used (or considered using) OSM in their work. Testers were also asked to rate their level of confidence in OSM for their work purposes on a five-point scale ranging from "I have major concerns with the quality of OSM data for my work purposes" (a score of 1) to "I am highly confident in the quality of OSM data for my work purposes" (a score of 5).

Testers were then required to read an introductory documentation page on Crowd Lens describing the purpose of the tool and pointing out the main user interface components (see Appendix C for the entire text of this page). Following this, they were asked to use the Crowd Lens tool for as long as they liked (with a minimum of 10 minutes) to explore the available OSM data in any way they wanted. As they did so, they were asked to try to form an understanding of how OSM had taken shape over time in different towns through the efforts of individual contributors. They were also asked to consider how the development of OSM in these towns affected their own perception of the quality of OSM data with respect to their own projects.

Following the use of the tool, testers were also asked to share any insights they had gained about OSM data from using this tool that they had not known or considered previously,

and describe how those insights might affect the way that they worked with OSM data in the future. Testers were then invited to suggest any improvements to the tool that they felt might allow them to develop new insights about OSM that would be valuable in their work. Finally testers were asked how the tool had affected their confidence in the quality of OSM data for their work purposes, using a five-point scale ranging from "I feel less confident in the quality of OSM data for my work purposes after using this tool" (a score of 1) to "I feel more confident in the quality of OSM data for my work purposes after using this tool" (a score of 5).

Questions in the survey were largely open-ended because we anticipated that the professionals recruited as testers used OSM in an expansive variety of ways and we wanted to learn about their approaches for using and interpreting OSM data. The purpose of this test was to assess the utility and usefulness of the tool and we did not want to create closed-ended questions reflecting (even inadvertently) any preconceived notions from us about how people would use the tool. We did not ask testers to complete any specific task during the evaluation because we wanted them to tailor the use of Crowd Lens to their own interests, and we felt the intuitiveness of the interface had already been adequately vetted through the early usability test and subsequent improvements to the interface described previously. We assessed the results of the survey by reading the responses to each question and keeping a tally of recurring themes. We also noted tester insights that were particularly detailed, innovative, or outside the realm of what we had expected.

Results of testing by geospatial technology professionals

Testers' past professional experiences with OSM included tasks from a range of basic to advanced skill levels. For instance, some occasionally used OSM as a basemap, others used it to supplement datasets from governments and commercial sources, and several worked directly with

OSM source data they had extracted for use in GIS analysis projects. It seemed that few, if any, testers used OSM as their primary data source at work, although this was not asked explicitly.

On the five-point scale of confidence in OSM data quality for work purposes (with 5 being the highest), the mean result was 3.8, with seven out of the 10 testers reporting a score of 4 and only one tester reporting a score on the lower end of the scale (Figure 4-5). Most testers reported that they had found OSM suitable for their work purposes and three mentioned that OSM was typically better than other data sources they encountered; however, many of these same people maintained some concerns about errors persisting in the OSM data. These concerns were characterized by a tester who remarked, "Open Street Map [*sic*] data looks fairly clean for my area, but I have found areas with older information which makes me question the data as a whole." Other challenges mentioned with OSM included a difficulty of convincing colleagues of its data quality and the lack of certain types of data in OSM such as elevation values. The tester who reported a confidence level of 2 noted the legal pitfalls of using crowdsourced geographic data in maps used in litigation. No testers directly mentioned biases of human focus (on places or entities) as a factor affecting their perception of OSM data quality, although three of them noted a concern with the variation of OSM coverage and accuracy across space.

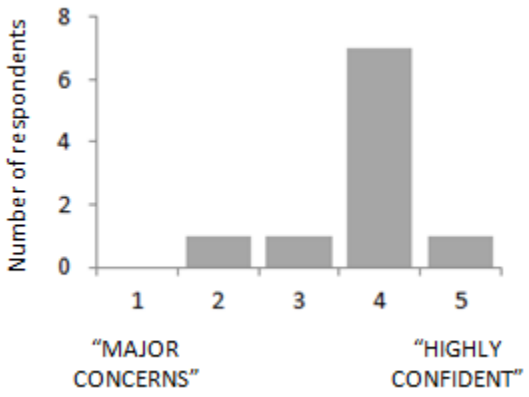


Figure 4-5. Responses to the pre-assessment question "How confident are you in the quality of OpenStreetMap data for your work purposes?"

One objective of the testing was to understand what (if any) insights about OSM data that testers derived from their interaction with Crowd Lens. In this sense of the word, "insight" is more akin to the accumulation of new information rather than a moment of cognitive breakthrough, although the step-by-step accumulation of new information and knowledge from visual analytics tools likely does facilitate such breakthroughs (Chang et al. 2009). The most common insight that testers reported garnering from their experimentation with Crowd Lens was an appreciation of the magnitude of edits supplied by the most active OSM contributors, marked by a gap in activity levels between the most active contributors in a city and the least active. Several testers looking closely at the metadata noticed a division in the habits of these most active contributors: there are those who tend to stay local in their edits and those who edit a wider variety of places across the globe. One tester also observed a middle group of contributors (between heavy and light) that tended to focus on one type of task, such as adding addresses or working on water features.

Four of the 10 testers mentioned the potential for Crowd Lens to be used as a tool for assessing OSM data suitability or quality. One tester noted the pitfalls of relying on a raw count of editors, musing that "the number of editors may be a good indicator of the accuracy of the data, but it also may not..." while another warned, "one has to keep in mind that there may be 'hit and runs' where a one-time or inexperienced user may make a change." Thus, a large set of contributors displayed in Crowd Lens does not entirely dispel the fear of errors permeating the dataset.

Two testers suggested that Crowd Lens could be used for identifying experts in particular topics such as hiking, parks, or pedestrian infrastructure, who could then be recruited to help with targeted mapping projects or other initiatives. Although we had not anticipated the use of Crowd Lens in this way, it is enabled by the inclusion of the tag filter and tag cloud. Words are sized in

graduated fashion in the tag cloud such that larger words indicate more attention by the contributor to that particular tag. The tag filters allow a view of just the contributors who applied a particular tag. Future development work in this vein could focus on grouping tags to create a more readable and manageable list of theme-based filters, or highlighting the map features and changeset table records associated with the use of the selected tags or themes.

Despite all the testers sharing at least one new insight they gained about OSM from their use of Crowd Lens, several reported that the tool would not affect the way they worked with OSM because they already tended to use OSM data as a "last resort" when other sources were not available. Thus, quality seems to be less of a concern when the alternative option is to have no data. Overall, the testers' confidence levels in the quality of OSM for their work purposes went up after using Crowd Lens. On the 1 to 5 scale, with 1 being less confident after using the tool and 5 being more confident after using the tool, the mean of the responses was 3.7 (Figure 4-6). We were somewhat surprised to see that no one submitted a score lower than 3 because we felt that some testers would be alarmed by the small number of contributors in some of the cities and the relatively small amounts of work done by a large portion of the contributors. Asked to explain their answers, testers reported that they were comforted by seeing the number of contributors and the amount of attention rendered by the most active ones. A tester remarked, "seeing the large number of different contributors makes me more confident that errors will be fixed", while another was "very impressed at the scale of edits and dedication by volunteers." Testers whose level of confidence remained unmoved (ie., submitted a score of 3) cited lingering concerns with data accuracy that they had seen in the project, although one remarked, "While I do not feel more or less confident about the quality of OSM having just tried out the application, I feel more empowered to assess the quality of OSM across space".

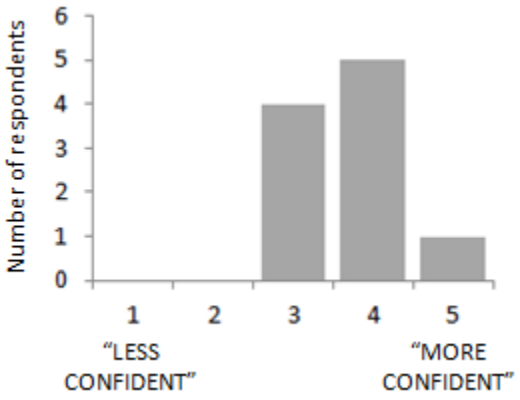


Figure 4-6. Responses to the post-assessment question "To what degree has this tool affected your perception of the quality of OpenStreetMap data for your work purposes?"

Testers offered numerous suggestions for improving Crowd Lens to provide additional insights. A popular request was to offer the end user of Crowd Lens the freedom to select any geographic bounding box for analysis. This goal remains out of scope for the initial implementation of Crowd Lens due to the extensive amount of processing required on the data to precalculate the statistics, geometry files, and small multiple map images. For example, the OSM history dump must be traversed to construct the past record of geometries modified by each contributor, and the OSM changeset history must also be scanned to pull out the contributor comments and other relevant metadata displayed in CrowdLens. These databases are so large that we decided to extract just the data for our study areas and work with smaller files containing only these data subsets. Search and indexing optimization techniques that could scan the entire OSM history files and construct these data structures on the fly would be necessary in order to achieve the goal of allowing any bounding box for analysis. Even still, to limit the amount of data returned, it would be judicious to place a cap on the study area size and scale allowed in a single request.

Two testers mentioned a desire for faster performance of the tool. The most computationally intensive portion of Crowd Lens is the back-end spatial processing that determines on the fly which contributors have influenced the current map bounding box. This might be alleviated by the introduction of some manner of spatial index to more rapidly eliminate the consideration of those contributors whose work is far away from the map extent. Another barrier to performance occurs with drawing the contributor geometries as vector graphics in the browser. This is typically a boon to performance and interactivity, but with the contributors who have added thousands of features, the number of graphics to be loaded and drawn can slow down the web browser. This might be alleviated by using a map image produced server-side through a web map service (WMS) or by serving image tiles from a pre-generated cache. Such an approach would sacrifice the interactive re-styling that is possible with vector graphics, and would introduce a new geospatial server tier into the application architecture. It was avoided for this version of the tool but could be implemented if larger datasets needed to be handled. Approaches using WebGL also hold promise for rendering large numbers of graphics in the browser, and would retain some of the interactive benefits of client-drawn vectors.

Finally, a tester commented that "the tool had too many adjustable elements that made the app confusing." This is presumably referring to the many filters available through slider widgets and dropdown menus on the left-hand panel. Although the claims analysis exercise had anticipated that some users might not like the visual clutter of the filters, we had perhaps underestimated how overwhelming the multitudinous available settings would seem to some users. A potential way to address this confusion was suggested by another tester: define some standard filter settings that would guide the end user toward learning certain results. For example, there may be thresholds of certain measures of participation (such as number of days active, percentage of OSM changesets in the target city, etc.) that would characterize the most active contributors who are most likely to have local knowledge of the place being mapped. Perhaps the

end users of Crowd Lens could be offered some pre-generated settings or default choices that would help them toward finding these interesting subsets of the crowd. These settings could be constructed through the domain knowledge of OSM experts, or they might be generated through data mining approaches trying to detect and suggest the most "interesting" trends in the metadata (Shneiderman 2002, Beale 2007).

Overall, the view of each OSM contributor's work offered by Crowd Lens had the effect of increasing some testers' confidence levels in the overall quality of OSM data. Others felt the tool was interesting and could possibly help them assess data quality in individual situations, but did not alleviate their worries about errors infiltrating the crowdsourced data.

Discussion of OSM contributor characteristics observed using Crowd Lens

We now offer some of our own observations about OSM data that we garnered from exploring data with Crowd Lens. A glance at any of the cities in Crowd Lens confirms that most of the work in OSM is done by a relatively small percentage of individuals, a statistic noted by various observers throughout the project's history including Neis and Zipf (2012) and Wood (2014). Each city has a group of individuals who has contributed across the geographic scope of the city, and each city has a large group who only created one changeset. Note that many users in the latter group have thousands of other changesets in OSM and came to these cities to make a single fix or addition. In other cases they were drawn to these cities for the purpose of making automated fixes and imports, as the text “bot” appears in several of the user names with large numbers of nodes and ways modified. Further examination of the contributor profile and wiki pages reveals the nature of the bots. In most cases their function is to add, remove, or adjust attribute tags to conform to a certain semantic standard.

Several of the cities have a small group of contributors who have made one or just a few edits in the OSM project as a whole, all within the study area boundary. Crowd Lens may be especially helpful in revealing the activities and motivations of these one-time OSM contributors who come to the project for one specific mission, or decide to leave the project after a single edit session and do not return. These contributors are easy to find by sorting the contributor list by "% changesets here". The science of one-time crowdsourcing contributors is a topic not deeply addressed in the literature, although Anthony et al. (2005) showed that in Wikipedia such one-time "Good Samaritan" edits can have more staying power than edits produced en masse by more prolific users. In Crowd Lens, the city of Tres Arroyos shows six out of 41 editors who produced 100% of their changesets in the town. All of them had fewer than 10 changesets. The city of Hervey Bay, Australia also saw six editors (out of 85) who made 100% of their changesets in the city. One added an automobile electric shop, while another added a gem/mineral club and a museum. These types of establishments might be valuable to the livelihoods of contributors from the perspectives of business or pleasure, and could provide similar benefits to local residents viewing the map. We have observed small business owners create an OSM account solely for the purpose of making sure their store is represented on the map. Others may be drawn to OSM for a short period as part of a school assignment or a social event such as a "mapping party" (Perkins and Dodge 2008, Hristova et al. 2013).

The language filters reveal that English is widely used for OSM changeset commenting by contributors to all cities studied. The only city without an English-favoring majority was Tres Arroyos, where most contributors used Spanish in their OSM comments. German-using contributors appear in all cities studied, confirming the popularity of the OSM project among Germans identified by Neis and Zipf (2012). The cities of Suhum, Ghana and Kadiri, India saw no contributors favoring the various regional languages used among the populace of these areas such as Twi, Ewe, Telugu, or Hindi, a finding that could reflect both a desire from the more local

mappers to communicate in English to reach a broader audience, as well as a heavy influence from mappers living elsewhere, particularly the Global North.[22]

Overall, comparisons of contributor activity reveal stark differences between the cities from the Global North (here including Australia), and those in the Global South. The total number of contributors for each city is shown in Figure 4-7. An investigation of the edits and contributors in Suhum and Kadiri reveals relatively little profile information and sparse contributor comments. Interestingly, it appears that bot activity has only affected the top three cities in Figure 4-7. Since bots and import activities are controversial and have stirred debates about their effects on OSM communities (McConchie 2015), it will be interesting to see if OSM develops in different ways in the Global South, relying on smaller contributions that contain more local knowledge. Tres Arroyos may be an example of this phenomenon, having seen systematic OSM editing activity by local-language contributors, who have created a detailed street map without being driven by bots or mass imports.

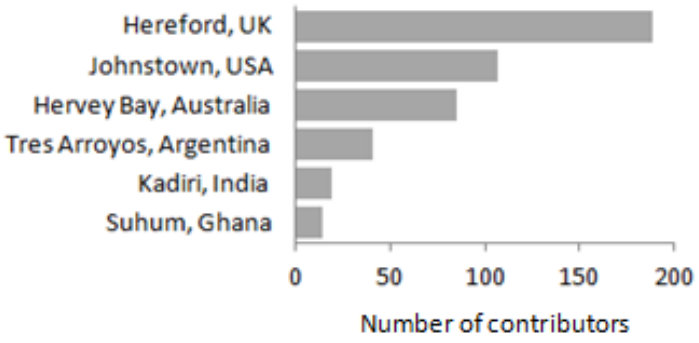


Figure 4-7. Number of unique contributors in OSM history files for years 2007–2015.

[22] I do not speculate which reason is more likely, because neither of these two cities saw any contributor making over 20% of his or her edits there. This makes it difficult to guess any particular contributor that might be local to the area. The influence of OSM "power mappers" is stronger in Kadiri, where 12 out of 19 contributors have logged over 1000 changesets in the OSM project (as opposed to 4 of 14 in Suhum).

Conclusion

It can be easy to think of the "crowd" behind a crowdsourced information product as something amorphous and nebulous, without known or knowable bounds. This leads to the risk that the crowd is overly feared or revered, with end users of the crowdsourced product not really knowing how much faith they can put in the resulting information quality and fitness for use. With the introduction of the Crowd Lens geovisual analytics tool, we have given a new view of the OSM contributor set that helps users of OSM better comprehend the social influences that underlie the data. Crowd Lens starts with a view of the overall aspects of the contributor crowd in a place, while allowing the user to drill down to any microgeographic region or any individual contributor record. Qualitative data such as contributor comments are visually linked to the corresponding bits of contributed information, allowing the user to understand connections between a database edit and the contributor's view of the edit's purpose.

Crowd Lens was shaped by a user-centered design process that saw multiple rounds of user evaluation during development. Early usability testers offered a multitude of suggestions about how the tool could be made easier to use and interpret. We developed scenarios of use and an accompanying claims analysis to understand which of these suggestions to prioritize to facilitate the most salient workflows. Upon showing a refined product to geospatial technology professionals, we found that when their confidence of OSM data quality was affected by the tool, it was only increased (Figure 4-6). The professionals seemed comforted by the view of the many contributors who have influenced OSM (some with thousands of edits); however, the size of the crowd alone was not enough to erase their apprehension about errors creeping into OSM data.

In addition to the evaluations described in this chapter, we suggest several additional approaches for testing and refining the usability and utility of this tool: (1) On the usability side, an explicit task analysis could be performed wherein the items in the claims analysis would be

directly tested. For example, do the crowd filters really cause end users to forget that they are viewing just a subset of the contributors? (2) From the utility side, the tool could be shared with researchers, social scientists, and professionals who work with non-geographic crowdsourced data products such as Wikipedia on a regular basis. They could be asked about how their insights from Crowd Lens might apply to this other crowdsourced data. These testers might also be solicited for suggestions about how a similar visual analytics tool might be designed and tailored for their own crowdsourced data platforms of interest. We are confident that elements of our design and evaluation approaches could be applied toward the analysis of other large repositories of crowdsourced data, such as Wikimapia, the millions of geotagged articles on Wikipedia (see Graham et al. 2014 and http://www.geonames.org/wikipedia/), or even "citizen science" projects such as those on Zooniverse.org. As OSM and these other crowdsourced datasets grow in popularity and increase in attractiveness compared to the more expensive alternatives, we anticipate that interactive visual analysis of contributor crowds will only grow in demand.

Chapter 4 references

Andrienko, Gennady, Natalia Andrienko, Piotr Jankowski, Daniel Keim, M.-J. Kraak, Alan MacEachren, and Stefan Wrobel. 2007. "Geovisual Analytics for Spatial Decision Support: Setting the Research Agenda." *International Journal of Geographical Information Science* 21 (8): 839–57.

Anthony, Denise, Sean W. Smith, and Tim Williamson. 2005. "Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia." *Hanover: Dartmouth College*. http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf.

Ballatore, Andrea. 2014. "Defacing the Map: Cartographic Vandalism in the Digital Commons." *The Cartographic Journal*. http://www.maneyonline.com/doi/abs/10.1179/1743277414Y.0000000085.

Barron, Christopher, Pascal Neis, and Alexander Zipf. 2013. "iOSMAnalyzer – Ein Umfassendes Werkzeug Für Intrinsische OSM Qualitätsuntersuchungen." In *AGIT 2013*. Salzburg, Austria. http://koenigstuhl.geog.uni-heidelberg.de/publications/2013/Barron/Barron_et_al_iOSMAnalyzer@agit_2013.pdf.

Barth, Alex. 2015. "The Paid Mappers Are Coming." In *State of the Map US 2015*. http://stateofthemap.us/the-paid-mappers-are-coming/.

BBC News. 2015. "Google Suspends Map Maker because of Vandalism," May 12. http://www.bbc.com/news/technology-32704566.

Beale, Russell. 2007. “Supporting Serendipity: Using Ambient Intelligence to Augment User Exploration for Data Mining and Web Browsing.” *International Journal of Human-Computer Studies*, Ambient intelligence: From interaction to insight, 65 (5): 421–33. doi:10.1016/j.ijhcs.2006.11.012.

Bivand, Roger S. 2010. “Exploratory Spatial Data Analysis.” In *Handbook of Applied Spatial Analysis*, edited by Manfred M. Fischer and Arthur Getis, 219–54. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-03647-7_13.

Borra, Erik, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. “Societal Controversies in Wikipedia Articles.” In *ACM CHI Conference on Human Factors in Computing Systems*. Seoul, South Korea. http://delivery.acm.org/10.1145/2710000/2702436/p193-borra.pdf?ip=75.102.66.179&id=2702436&acc=ACTIVE%20SERVICE&key=A792924B58C015C1.782FA3A5BE459501.4D4702B0C3E38B35.4D4702B0C3E38B35&CFID=576281334&CFTOKEN=52919993&__acm__=1453138571_53d71f8c8e046b620c611c07a60ac97f.

Boukhelifa, Nadia, Fanny Chevalier, and J. Fekete. 2010. “Real-Time Aggregation of Wikipedia Data for Visual Analytics.” In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, 147–54. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5652896.

Brandes, Ulrik, and Jürgen Lerner. 2008. “Visual Analysis of Controversy in User-Generated Encyclopedias∗.” *Information Visualization* 7 (1): 34–48.

Budhathoki, Nama Raj. 2010. “Participants’ Motivations to Contribute Geographic Information in an Online Community.” Ph.D., United States -- Illinois: University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/bitstream/handle/2142/16956/1_Budhathoki_Nama.pdf?sequence=2.

Burns, Ryan, and André Skupin. 2013. “Towards Qualitative Geovisual Analytics: A Case Study Involving Places, People, and Mediated Experience.” *Cartographica: The International Journal for Geographic Information and Geovisualization* 48 (3): 157–76.

Burns, Ryan, and Jim Thatcher. 2014. “Guest Editorial: What’s so Big about Big Data? Finding the Spaces and Perils of Big Data.” *GeoJournal*, October. doi:10.1007/s10708-014-9600-8.

Card, Stuart, Jock Mackinlay, and Ben Shneiderman. 1999. *Readings in Information Visualization: Using Vision to Think*. San Francisco, California: Morgan Kaufmann Publishers.

Chan, B., L. Wu, J. Talbot, M. Cammarano, and P. Hanrahan. 2008. “Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration.” *IEEE Transactions on Visualization and Computer Graphics* 14 (6): 1213–20. doi:10.1109/TVCG.2008.178.

Chang, R., C. Ziemkiewicz, T.M. Green, and W. Ribarsky. 2009. “Defining Insight for Visual Analytics.” *IEEE Computer Graphics and Applications* 29 (2): 14–17. doi:10.1109/MCG.2009.22.

Cho, I., Wewnen Dou, D. X. Wang, E. Sauda, and W. Ribarsky. 2016. "VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 210–19. doi:10.1109/TVCG.2015.2467971.

Collins, Christopher, Fernanda B. Viegas, and Martin Wattenberg. 2009. "Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora." In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, 91–98. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5333443.

Dou, Wenwen, I. Cho, O. ElTayeby, Jaegul Choo, Xiaoyu Wang, and W. Ribarsky. 2015. "DemographicVis: Analyzing Demographic Information Based on User Generated Content." In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 57–64. doi:10.1109/VAST.2015.7347631.

Dou, Wenwen, William Ribarsky, and Remco Chang. 2010. "Capturing Reasoning Process through User Interaction." *International Symposium on Visual Analytics Science and Technology 2010*. http://webpages.uncc.edu/~wdou1/publications/2010/EuroVAST2010-InterfaceDesign.pdf.

Dou, Wenwen, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou. 2012. "Leadline: Interactive Visual Analysis of Text Data through Event Identification and Exploration." In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, 93–102. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6400485.

Endert, A., R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook. 2013. "Typograph: Multiscale Spatial Exploration of Text Documents." In *2013 IEEE International Conference on Big Data*, 17–24. doi:10.1109/BigData.2013.6691709.

Fisher, Danyel. 2007. "Hotmap: Looking at Geographic Attention." *Visualization and Computer Graphics, IEEE Transactions on* 13 (6): 1184–91.

Fuchs, Georg, Natalia Andrienko, Gennady Andrienko, Sebastian Bothe, and Hendrik Stange. 2013. "Tracing the German Centennial Flood in the Stream of Tweets: First Lessons Learned." In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 31–38. GEOCROWD '13. New York, NY, USA: ACM. doi:10.1145/2534732.2534741.

Goodchild, Michael F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–21.

Graham, Mark, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. 2014. "Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty." *Annals of the Association of American Geographers* 104 (4): 746–64.

Graser, A., M. Straub, and M. Dragaschnig. 2014. "Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph." *Transactions in GIS* 18 (4): 510–26.

Haklay, Mordechai. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning. B, Planning & Design* 37 (4): 682.

———. 2014. “OpenStreetMap Studies (and Why VGI Not Equal OSM).” *Po Ve Sham - Muki Haklay’s Personal Blog*. August 14. https://povesham.wordpress.com/2014/08/14/openstreetmap-studies-and-why-vgi-not-equal-osm/.

Haklay, Mordechai, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. 2010. “How Many Volunteers Does It Take to Map an Area Well? The Validity of Linus’ Law to Volunteered Geographic Information.” *The Cartographic Journal* 47 (4): 315–22.

Hristova, Desislava, Giovanni Quattrone, Afra Mashhadi, and Licia Capra. 2013. “The Life of the Party: Impact of Social Mapping in OpenStreetMap.” In *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM2013)*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6098/6362.

Keim, Daniel A., Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008a. “Visual Analytics: Scope and Challenges.” In *Visual Data Mining*, edited by Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika, 76–90. Lecture Notes in Computer Science 4404. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-71080-6_6.

Keim, Daniel, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008b. "Visual Analytics: Definition, Process, and Challenges." In *Information Visualization*, edited by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, 154–75. Lecture Notes in Computer Science 4950. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-70956-5_7.

Kounadi, Ourania. 2009. "Assessing the Quality of OpenStreetMap Data." *Msc Geographical Information Science, University College of London Department of Civil, Environmental And Geomatic Engineering*. ftp://ftp.cits.nrcan.gc.ca/pub/cartonat/Reference/VGI/Rania_OSM_dissertation.pdf.

Lui, Marco, and Timothy Baldwin. 2012. "Langid. Py: An off-the-Shelf Language Identification Tool." In *Proceedings of the ACL 2012 System Demonstrations*, 25–30. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2390475.

MacEachren, Alan M., Stephen Crawford, Mamata Akella, and Gene Lengerich. 2008. "Design and Implementation of a Model, Web-Based, GIS-Enabled Cancer Atlas." *The Cartographic Journal* 45 (4): 246–60. doi:10.1179/174327708X347755.

MacEachren, Alan M., Monica Wachowicz, Robert Edsall, Daniel Haug, and Raymon Masters. 1999. "Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographical Visualization with Knowledge Discovery in Database Methods." *International Journal of Geographical Information Science* 13 (4): 311–34.

McConchie, Alan. 2015. “Tracing Patterns of Growth and Maintenance in OpenStreetMap.” In *State of the Map US 2015*. New York, NY, USA. http://stateofthemap.us/tracing-patterns-of-growth-and-maintenance-in-openstreetmap.

McHugh, Bibiana. 2014. “Government as a Contributing Member of the OpenStreetMap (OSM) Community.” In *FOSS4G 2014*. Portland, Oregon. https://vimeo.com/album/3606079/video/106226528.

Napolitano, Maurizio, and Peter Mooney. 2012. “MVP OSM: A Tool to Identify Areas of High Quality Contributor Activity in OpenStreetMap.” *The Bulletin of the Society of Cartographers* 45 (1): 10–18.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf. 2011. “The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011.” *Future Internet* 4 (1): 1–21.

———. 2013. “Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions.” *Future Internet* 5 (2): 282–300.

Neis, Pascal, and Alexander Zipf. 2012. “Analyzing the Contributor Activity of a Volunteered Geographic Information project—The Case of OpenStreetMap.” *ISPRS International Journal of Geo-Information* 1 (2): 146–65.

Perkins, Chris, and Martin Dodge. 2008. “The Potential of User-Generated Cartography: A Case Study of the OpenStreetMap Project and Mapchester Mapping Party.” *North West Geography* 8 (1): 19–32.

Peuquet, Donna J., Anthony C. Robinson, Samuel Stehle, Franklin A. Hardisty, and Wei Luo. 2015. “A Method for Discovery and Analysis of Temporal Patterns in Complex Event Data.” *International Journal of Geographical Information Science* 29 (9): 1588–1611. doi:10.1080/13658816.2015.1042380.

Pirolli, Peter, and Stuart Card. 2005. “The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis.” In *Proceedings of International Conference on Intelligence Analysis*, 5:2–4. Mitre McLean, VA. http://vadl.cc.gatech.edu/documents/2__card-sensemaking.pdf.

Quinn, Sterling. in press. “A Geolinguistic Approach for Identifying Locally Contributed Content in OpenStreetMap.” *Cartographica*. doi:10.3138/CART.MS3301.

———. 2016. “Using Small Cities to Understand the Crowd behind OpenStreetMap.” *GeoJournal*. doi:10.1007/s10708-015-9695-6.

Raymond, Eric. 1999. “The Cathedral and the Bazaar.” *Knowledge, Technology & Policy* 12 (3): 23–49.

Ribarsky, William, Brian Fisher, and William M. Pottenger. 2009. “Science of Analytical Reasoning.” *Information Visualization* 8 (4): 254–62. doi:10.1057/ivs.2009.28.

Roberts, J.C. 2007. “State of the Art: Coordinated Multiple Views in Exploratory Visualization.” In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV ’07*, 61–71. doi:10.1109/CMV.2007.20.

Robinson, Anthony C., Jin Chen, Eugene J. Lengerich, Hans G. Meyer, and Alan M. MacEachren. 2005. "Combining Usability Techniques to Design Geovisualization Tools for Epidemiology." Cartography and Geographic Information Science 32 (4): 243–55. doi:10.1559/152304005775194700.

Roick, Oliver, Julian Hagenauer, and Alexander Zipf. 2011. "OSMatrix–grid-Based Analysis and Visualization of OpenStreetMap." *Proceedings of the 1st European State of the Map*. http://koenigstuhl.geog.uni-heidelberg.de/publications/2011/Roick/Roick_2011_SotM.pdf.

Roick, Oliver, Lukas Loos, and Alexander Zipf. 2012. "A Technical Framework for Visualizing Spatio-Temporal Quality Metrics of Volunteered Geographic Information." In . http://koenigstuhl.geog.uni-heidelberg.de/publications/2012/Roick/Roick_OSMatrix_Geoinformatik2012.pdf.

Rosson, Mary Beth, and John Carroll. 2002. "Scenario-Based Design." In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, edited by Julie Jacko and A Sears, 1032–50. Lawrence Erlbaum Associates.

Roth, Robert E., Kevin S. Ross, and Alan M. MacEachren. 2015. "User-Centered Design for Interactive Maps: A Case Study in Crime Analysis." *ISPRS International Journal of Geo-Information* 4 (1): 262–301. doi:10.3390/ijgi4010262.

Sacha, Dominik, Andreas Stoffel, Florian Stoffel, Bum Kwon, Geoffrey Ellis, and Daniel Keim. 2014. “Knowledge Generation Model for Visual Analytics.” *IEEE Transactions on Visualization and Computer Graphics* 99 (PrePrints): 1. doi:10.1109/TVCG.2014.2346481.

Shneiderman, Ben. 1994. “Dynamic Queries for Visual Information Seeking.” *IEEE Software* 11 (6): 70–77.

———. 1996. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, 336–43. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=545307.

———. 2002. “Inventing Discovery Tools: Combining Information Visualization with Data Mining1.” Information Visualization 1 (1): 5–12. doi:10.1057/palgrave.ivs.9500006.

Slocum, Terry A., Daniel C. Cliburn, Johannes J. Feddema, and James R. Miller. 2003. “Evaluating the Usability of a Tool for Visualizing the Uncertainty of the Future Global Water Balance.” *Cartography and Geographic Information Science* 30 (4): 299–317. doi:10.1559/152304003322606210.

Suh, Bongwon, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. 2007. “Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations.” In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, 163–70. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4389010.

Thomas, James J., and Kristin A. Cook. 2005. “Illuminating the Path: The Research and Development Agenda for Visual Analytics.” Pacific Northwest National Laboratory (PNNL), Richland, WA (US). http://www.osti.gov/scitech/biblio/912515.

Trame, Johannes, and Carsten Keßler. 2011. “Exploring the Lineage of Volunteered Geographic Information with Heat Maps.” *GeoViz, Hamburg, Germany*. http://carsten.io/trame-kessler-geoviz2011.pdf.

Van Exel, Martijn. 2011a. “A New OpenStreetMap Visualization: Version Contour Lines.” *Oegeo*. June 20. https://oegeo.wordpress.com/2011/06/20/a-new-openstreetmap-visualization-version-contour-lines/.

———. 2011b. “Taking the Temperature of Local OpenStreetMap Communities.” *Oegeo*. September 19. https://oegeo.wordpress.com/2011/09/19/taking-the-temperature-of-local-openstreetmap-communities/.

———. 2014. “OpenStreetMap and Telenav; Past, Present and Future.” In *State of the Map 2014*. Buenos Aires, Argentina. https://vimeo.com/album/3134207/video/112305387.

Viégas, Fernanda B., Martin Wattenberg, and Kushal Dave. 2004. “Studying Cooperation and Conflict between Authors with History Flow Visualizations.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 575–82. http://dl.acm.org/citation.cfm?id=985765.

Wood, Harry. 2014. “The Long Tail of OpenStreetMap.” In *State of the Map 2014*. Buenos Aires, Argentina. http://vimeo.com/album/3134207/video/112438218.

Zielstra, Dennis, and Hartwig Hochmair. 2011. “Digital Street Data: Free versus Proprietary.” *GIM International* 25 (7). http://www.gim-international.com/issues/articles/id1739-Digital_Street_Data.html.

Chapter 5

Conclusion

The goal of this dissertation was to develop a suite of methods and tools to better reveal geographic variations in the OpenStreetMap (OSM) contributor set in a given place, while also illuminating the unique interests and practices exercised by individual contributors toward the construction of the map. This goal finds its origin in questions of power in cartography (Harley 1989), geographies of digital divides (Graham 2011), and the often-heralded idea that volunteered geographic information (VGI) can facilitate the democratization of mapping (Haklay 2013). I focused the goals and objectives on OSM throughout (rather than a broader set of VGI data types) due to OSM's global nature and its rapid uptake as an alternative to institutionally produced datasets. In doing so, I contribute to the emerging research subdomain of "OpenStreetMap studies" (Haklay 2014).

The body of this dissertation described three academic papers, each offering its own research objectives and approaches toward the broader goals described above. The first study evaluated the degree to which the language mix of the OSM contributor set could be taken as an indicator of nonlocal influence on the map (particularly in places like the study area of South America where English is not commonly spoken). Acknowledging a degree of uncertainty, it was shown that groups of edits from English speaking contributors are more likely to contain more nonlocal influence than groups of edits from Spanish or Portuguese speaking contributors. When languages used in the OSM contributor comments were automatically detected and mapped, spatially uneven patterns of language use emerged. English use occurred everywhere, but was more prominent in the more rural and less wealthy parts of South America. Users of the local languages Spanish and Portuguese tended to put more emphasis on mapping things associated

with the routines of everyday life in South America such as going to school, buying food at the store, riding the bus, and so forth, while English speaking users tended to map sites of tourism, mass consumption, or things that could be traced from an aerial photograph. These findings suggest that some places have less local control over the map than other places, and that a high degree of local control leads to a map with a greater connection to residents' livelihoods. Thus, cartographies of silence may persist in subtle ways in a digital map produced by armchair mappers, omitting the entities that create a sense of place for local residents. This can occur even in areas that appear to be already mapped with a street network and other basic data. Follow up studies to confirm these effects and their magnitude would be a useful contribution to OpenStreetMap studies and the VGI literature in general.

In order to get a deeper look at who wields power in the construction of OSM, the second study in this dissertation examined in depth the OSM contributor "crowds" in three small South American cities, comparing them to the crowds that mapped two small cities in North America and one large city in South America. Small cities were chosen for focus in this study because they have a limited and bounded contributor set that can be examined in great depth. They also have traditionally been under-studied in the literature, even though maps in small cities are needed just as much as maps in large cities for applications that purport to offer global coverage. After examining small multiple maps of each contributor's work, as well as the entire set of metadata comments offered by contributors, it became clear that the "crowd" in the small cities studied is more like a "handful" of contributors, with only a few being very actively involved in the project. A typology of mapper roles was constructed from this analysis, showing that contributors are drawn to the project not only by a pride of local knowledge, but also by a desire to fulfill personal and institutional interests. In small cities, many contributors even come at random to fix the map, brought by applications that automatically detect anomalies in OSM data. Recruitment of local mappers at secondary schools and community colleges may be a way to supplement the map with

local knowledge in these cities, as the contributor crowds there are not seeing the same rate of increase as in larger neighboring cities.

The above observations about OSM in small cities were gathered through static maps and graphs produced after mining the OSM history files and their metadata. In order to facilitate greater interaction with the OSM data and better evaluate the data's fitness for use in professional contexts, a geovisual analytics tool called Crowd Lens was created and tested. This was discussed as the third paper in this dissertation. Crowd Lens is designed to help potential users of OSM understand the size and characteristics of the crowd behind a city, while allowing filtering on the contributor set to better identify subsets of contributors with similar characteristics. It also allows the analyst to compare users, side by side through sortable series of small multiple maps of each person's work on OSM, and to retrieve detailed metadata about any particular contributor. This tool makes it easier to pinpoint automated activities (such as imports or bots) and systematic mapping efforts from power users, while allowing a further investigation into the backgrounds and motives of these contributors. It also facilitates the identification of the most active subset of the crowd in any particular place, going beyond the raw number of contributors, points, and lines in the dataset to find out how much human attention is really being paid to the map. Finally, it reveals discrepancies in attention to the map among similar-sized cities in different parts of the world.

Taking a user-centered design approach, we employed multiple stages and methods of evaluation to gauge the usability and utility of the Crowd Lens environment. Feedback collected from early usability tests and scenarios of use led to the inclusion of additional filters to help summarize and identify subsets of the crowd, such as those contributors who had employed a certain tag, or who had reached a particular number of active days working on the map in the city of interest. The initial usability study result fed into design of two scenarios of use and an associated claims analysis. This guided further development as well as a follow-up evaluation of

Crowd Lens' utility. For that evaluation, we invited feedback from geospatial technology professionals who use OSM in their day-to-day work. Crowd Lens increased some of these testers' levels of confidence in OSM data quality, while others remained wary of the potential for error to enter the crowdsourced data. Surprisingly to us, the tool did not reduce anyone's confidence in the data, in spite of its ability to show the relatively small numbers of active contributors in some places and to make clear that for some places many of the contributors are non-local.

Although testers were largely positive about the Crowd Lens environment, they mentioned a desire for improvement in the speed of the tool and a greater flexibility in choosing the geographic areas for analysis. Addressing these challenges may require substantial architectural changes that would be suitable for further research.

The three studies in this dissertation are unique in their focus on the human side of OSM construction. Although some effort in other literature has been made to learn about Wikipedia contributors and their interactions, most research involving the OSM metadata has been focused on studying the raw quantity and accuracy of the data. Although some surveys and interviews of OSM contributors have been conducted by Budhathoki (2010) and Lin (2011), I find in this dissertation that there is a lot that can be learned directly from OSM user profile pages and changeset metadata. Although these sources such as contributor comments and maps of each person’s work may not provide as much depth of information as personal visits, it is available on a much broader scale and can be summarized through innovative and interactive visual methods.

Using the methods and tools developed in this dissertation, we see that much OSM work is done by very dedicated individuals who are experienced at applying semantic and topological schemas and can interpret aerial photographs of places far from their hometowns. Others have done extensive ground surveys using GPS equipment. Many of them map for love of a particular type of hobby, such as railroads or bicycle trails, rather than for love of a particular place. A

growing number of contributors map on behalf of institutions such as corporations or governments that have decided to use OSM to support their geographic information services. It is sometimes difficult to ascertain these motives by looking at the raw geometries, but analysis of the contributor comments, profiles, languages used, and other metrics can help provide more form, bounds, and texture to our conceptualization of the crowd behind the crowdsourced map.

Challenges

The longer OSM exists and the more it increases in popularity, the larger and more unwieldy the dataset becomes, especially the full history archives. At the time of this writing the size of the OSM full history dump is about 50 GB compressed. Although this is cumbersome to manage on a typical laptop or desktop machine and outside the grasp of viewing in conventional text editors, utilities such as Osmconvert[23] are available to clip and uncompress geographic envelopes of the data without the need to uncompress the entire file. Clipped or not, the OSM history is much smaller than the terabytes of information involved with other sources of geographic data such as remotely sensed imagery collections. A bigger challenge than the size of the OSM history is finding user-friendly tools to work with it. Although QGIS and other providers now offer open source point-and-click tools for downloading the latest data from OSM, options for working with the history dump files are much less common and the ones that exist are not approachable for novices. The handful of experimental tools created by the OSM user community typically require knowledge of Linux commands and package management; there is no graphical user interface (GUI) environment for extracting any piece of the OSM history or converting it to common geographic information system (GIS) file formats. I believe some of this is due to the challenge of constructing historical geometries and working with features that have

[23] http://wiki.openstreetmap.org/wiki/Osmconvert

multiple versions. Since OSM nodes and ways are stored separately and referenced through IDs and version numbers, the geometries used in this dissertation were created with custom Python scripts that carefully constructed each OSM way (line) using the proper version of each participating node (point or vertex) in the line. The methodology I used for maintaining the nodes and ways in memory during this process would require additional random access memory (RAM) or code optimization in order to handle any more than about 1 GB of uncompressed history data.

An undoubtedly useful supplement to the qualitative information derived from the OSM metadata and user profile pages would be to directly ask OSM contributors to report whether they have ever visited the places they have mapped, and what prompted them to make the edits they did; however, contacting large targeted samples of OSM contributors in order to perform research for this dissertation proved elusive. An attempt I made in 2015 at administering a brief open-ended survey to several hundred contributors through the internal OSM messaging system yielded a better response rate than anticipated and much congenial communication with contributors. Unfortunately, the survey had to be abandoned prematurely because a smaller subset of survey recipients did not believe that research inquiries should be sent through these messaging channels. Working more closely with established discussion forums and working groups in the OSM community may lead to more welcoming outcomes in the future, however the responses might be biased toward the most active contributors to a greater degree than would be achieved through direct messaging subsets of contributors selected using other types of methods. In any case, it appears the project has reached a large and diverse enough user base that the ability to send large volumes of inquiries through the OSM messaging channels may not ever be possible again to the comprehensive degree carried out by Budhathoki (2010) during the earlier days of the project.

The three papers in this dissertation each illuminate in their own particular ways the imbalance of participation in OSM throughout different regions of the world, with lower levels of mapping activity prevailing across rural areas, smaller cities, and many parts of the Global South.

I do not attempt to address this problem in detail here; however, throughout the chapters I have offered a few ideas about how local mappers could be recruited as one strategy to fill the gap. I believe one of the most promising channels of outreach effort through local schools and community colleges, where students are often required to learn something of geography, computers, or both, and instructors often desire to incorporate hands-on, visual activities that have a real world impact. In places where school systems lack the resources to teach digital mapping on their own, support from state or national government institutions may be necessary. For example, in a project sponsored by the Argentine National Institute of Agricultural Technology (INTA), Raffo and Umaña (2014) trained secondary school students to put their rural town of Rio Chico, Rio Negro province into OSM, leading to community discussions about how to formalize the names of local streets. Coming from the side of academia, the TeachOSM initiative described by Cowan and Hinton (2014) provides a toolset for instructors to incorporate OSM into their GIS curriculums. They describe a scenario where students are directed to under-mapped places of the world to perform armchair mapping tasks such as imagery tracing. I have found in my own teaching of introductory and advanced GIS courses that students can often find basic amenities in their own hometowns that have not yet been added to OSM, thus allowing them to experience the pride of sharing local knowledge that so often compels contributors back to the OSM mapping environment for multiple sessions of participation. Students often remark that they continued mapping beyond the parameters of the assignment because they were enjoying themselves so much.

One inevitable challenge of researching a crowdsourced information project is that statistics summarizing the amount and nature of the data will eventually become out of date as the project continues forward. Of course, the research presented in these chapters can be repeated using the same temporal range of OSM history that I used; however, applying the same analyses with later OSM histories is expected to produce different results. The methods and tools

presented in this work, however, are just as applicable for studying OSM and other VGI efforts in the future, and if invoked repeatedly in different areas of the world may shed light on common patterns or stages of data growth across geographies.

Avenues for further study

Although this dissertation and other research efforts have shown that the development of OSM proceeds in an uneven fashion geographically, the OSM project itself seems poised to continue growing and attracting contributions from around the world. The prevalence of local language use among OSM contributors in many areas of South America is evidence of the project's maturity and reach beyond the realms of Europe and US where OSM first gained widespread attention. The increase in businesses and government entities that incorporate OSM into their services is already beginning to boost the number of paid mappers coming to the project, who contribute out of an interest in raising the overall suitability of the data for applications requiring comprehensive coverage rather than being driven by an affinity for any one particular place or theme (Quinn 2016). As more interests take an active role in contributing to OSM, it will be interesting to observe the degree to which conflict ensues (if at all) between institutionally-supported contributors and unpaid recreational mappers. Both can experience strong feelings of ownership over features added or modified in the database, even when the data is supposed to be owned by the community.

The heavy and apparently growing influence of top-down contributors working at broad scales in OSM contrasts with traditional conceptualizations of VGI as a bottom-up activity facilitated by local citizen sensors (Goodchild 2007). Topics of potential dispute between the two camps might include the choice of attribute schemas used, data imports or reverts performed by one of the parties, disproportionate amounts of influence in OSM governance exercised by

powerful contributors, proposed adjustment to the OSM data usage license, and other factors currently unforeseen. Crisis mappers and single-feature "hobbyists" in OSM whose attention is focused on a place for a relatively short time might also find themselves embroiled in conflicts with longer-term local users over feature placement, naming, and classification. On the other hand, the more optimistic possibility is that the additional attention to the map from diverse contributors will be perceived as welcome and beneficial by all parties involved. This is an issue to continue monitoring as OSM matures.

Another opportunity for conflict (and a fundamental challenge with OSM's crowdsourced contribution model) is the ability for one person to add information that would be considered private, sensitive, or proprietary to another person. For example, nothing prohibits me from adding and labeling my professor's house on the map for all to see. Although the geometry and attribute information about the house and its owner may be accurate and even learnable through public records, the personal nature of the information in such an exposed context (especially if no neighbors' houses are labeled) might easily raise ethical concerns about the intent and effect of the edits. This is a fairly benign example when compared with the potential for mapping activities that are locally considered covert, embarrassing, illegal, or historically associated with persecution, thereby exposing participants in those activities to abuse, shame, arrest, or other consequences. The level of sensitivity of many sites and activities such as marijuana dispensaries, places of worship of religious minorities, or social gathering places of gay and lesbian communities could vary across time, space, cultural norms, and political regimes. On the natural landscape, sites such as caves, coral reefs, or endangered species habitats may be environmentally fragile and vulnerable to overuse or destruction if widely publicized on a map. To anticipate these place-specific nuances, OSM contributors must attain a level of geographic literacy beyond the ability to decipher coordinate systems and interpret aerial photographs. They must also be able to anticipate both the positive and negative consequences that could come from mapping an object,

and then weigh those consequences when making decisions about what to contribute. For example, placing a detailed map of an informal "squatter" settlement in OSM could create a persuasive means of lobbying for services and attention in a chronically underserved community, while at the same time exposing that community to possible eviction or policing activity that residents view as oppressive. Studies of ethics in OSM crowdsourcing might benefit from analysis tools such as Crowd Lens to learn the nature and habits of contributors who focus their efforts on mapping sensitive items.

This dissertation has presented several instances where the OSM database contains items of rich local knowledge that could only be offered by someone highly familiar with a place; however, the massive amounts of remotely contributed data pouring into OSM raise the question of whether the unique place-based knowledge is often diluted (at best) or overridden (at worst) by imports, armchair tracing, and other remote influences. Continued development of methods for culling out and quantifying the locally produced information may prove useful in some contexts, such as helping geospatial technology professionals understand how much extra information might be gained if they were to switch from institutionally produced maps to OSM. One conceptual challenge for distilling the local knowledge is that automated data imports tend to have their attributes and geometries corrected or augmented by the OSM user community over time; thus, a US Census TIGER street line may start as imported data but end up as an exhibit of local knowledge if its geometries are refined and attributes enhanced by an OSM contributor familiar with the neighborhood. This local knowledge would prove especially valuable in cases where popular commercial maps had not undergone the corrections seen in OSM.

Studies such as the ones comprising this dissertation repeatedly show that many OSM contributors make only a few edits to the project and then never return. Yet the potential value of the knowledge contributed by a one-time local mapper should not be discounted (Elwood et al. 2013), especially when these begin to add up, as Anthony et al. (2005) demonstrated with

Wikipedia contributions. More study is needed about why people leave OSM, what kinds of influences might be driving them away, and what manner of experiences are most likely to encourage them remain active OSM mappers. Tools such as the Crowd Lens one introduced in this dissertation are helpful at identifying this subgroup of casual short-lived contributors for further study.

Concordant with Haklay's (2014) assertion that OSM is a unique enough project to deserve its own specific vein of study within VGI, this dissertation contains many observations that are unique to OSM and cannot be extended to all VGI. These include the specific geolinguistic distributions of the contributors and their likeliness of local residence, the analysis of institutional influence on the project as people look for a free alternative to commercial maps, and the challenges of recruiting a global contributor base. However, a more generalized examination of some of the issues raised in this dissertation could prove informative for researchers and practitioners who engage VGI and public participation GIS (PPGIS), or even geotagged crowdsourced data such as Wikipedia articles about world locations. These include the effects on volunteer focus and behavior when money or other rewards are offered for participation, the differing influences of volunteers stemming from varying levels of local knowledge, and the challenges of recruiting volunteers in areas affected by digital and participation divides. Thus, the same social influences that affect OSM construction could be made manifest in VGI, PPGIS, and crowdsourced information gathering efforts when considered in more general ways.

Finally, in its use of contributor comments, language analysis, and profile page content, this dissertation has demonstrated the wealth of ancillary qualitative information that researchers and GIS practitioners can use to better understand the OSM phenomenon. The incorporation of these sources with an interactive geovisual analytics tool might be thought of as an instance of "qualitative GIS", bringing in rich contextual details to supplement the raw points, lines, and

attributes (Elwood and Cope 2009). There are a number of other possibilities for useful qualitative data sources that have not been explored here, but could be harnessed in future work, including the archives of OSM discussion forums and transcripts from one-on-one interviews with contributors. One can imagine these sources giving root to a "qualitative OSM" in which the map data is linked to these other sources of evidence, exposing the ways that each object on the map is known and valued by the different people who modified it. Jung and Elwood (2010) and Burns and Skupin (2013) have outlined theory and demonstrated software frameworks that could inform such an approach. A further step on this path might be to spatially join or link geotagged media from other crowdsourced databases such as Mapillary, Wikimapia, and Wikipedia into cartographic representations of OSM, thereby adding new facets of ways that people know and interpret particular places.

Chapter 5 references

Anthony, Denise, Sean W. Smith, and Tim Williamson. 2005. "Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia." *Hanover: Dartmouth College*. http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf.

Budhathoki, Nama Raj. 2010. "Participants' Motivations to Contribute Geographic Information in an Online Community." Ph.D., United States -- Illinois: University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/bitstream/handle/2142/16956/1_Budhathoki_Nama.pdf?sequence=2.

Burns, Ryan, and André Skupin. 2013. "Towards Qualitative Geovisual Analytics: A Case Study Involving Places, People, and Mediated Experience." *Cartographica: The International Journal for Geographic Information and Geovisualization* 48 (3): 157–76.

Cowan, Nuala, and Richard A. Hinton. 2014. "TeachOSM." In *FOSS4G 2014*. Portland, Oregon. https://vimeo.com/106872862.

Elwood, Sarah, and Meghan Cope. 2009. "Qualitative GIS: Forging Mixed Methods Through Representations, Analytical Innovations, and Conceptual Engagements." In *Qualitative GIS: A Mixed Methods Approach,* edited by Meghan Cope and Sarah Elwood. SAGE Publications Ltd. http://dx.doi.org/10.4135/9780857024541.n1.

Elwood, Sarah, Michael F. Goodchild, and Daniel Sui. 2013. “Prospects for VGI Research and the Emerging Fourth Paradigm.” In *Crowdsourcing Geographic Knowledge*, 361–75. Springer. http://link.springer.com/chapter/10.1007/978-94-007-4587-2_20.

Goodchild, Michael F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211–21.

Graham, Mark. 2011. “Time Machines and Virtual Portals: The Spatialities of the Digital Divide.” *Progress in Development Studies* 11 (3): 211–27. doi:10.1177/146499341001100303.

Haklay, Mordechai. 2014. “OpenStreetMap Studies (and Why VGI Not Equal OSM).” *Po Ve Sham - Muki Haklay’s Personal Blog*. August 14. https://povesham.wordpress.com/2014/08/14/openstreetmap-studies-and-why-vgi-not-equal-osm/.

Haklay, Mordechai (Muki). 2013. “Neogeography and the Delusion of Democratisation.” *Environment and Planning A* 45 (1): 55–69. doi:10.1068/a45184.

Harley, John Brian. 1989. “Deconstructing the Map.” *Cartographica: The International Journal for Geographic Information and Geovisualization* 26 (2): 1–20.

Jung, Jin-Kyu, and Sarah Elwood. 2010. “Extending the Qualitative Capabilities of GIS: Computer-Aided Qualitative GIS.” *Transactions in GIS* 14 (1): 63–87. doi:10.1111/j.1467-9671.2009.01182.x.

Lin, Yu-Wei. 2011. “A Qualitative Enquiry into OpenStreetMap Making.” *New Review of Hypermedia and Multimedia* 17 (1): 53–71.

Quinn, Sterling. 2016. “Using Small Cities to Understand the Crowd behind OpenStreetMap.” *GeoJournal*. doi:10.1007/s10708-015-9695-6.

Raffo, Fernando, and Fernando Javier Umaña. 2014. “Posicionando a Río Chico En El Mapa.” In *State of the Map 2014*. Buenos Aires, Argentina. https://vimeo.com/album/3134207/video/112071538.

Shneiderman, Ben. 2002. “Inventing Discovery Tools: Combining Information Visualization with Data Mining1.” *Information Visualization* 1 (1): 5–12. doi:10.1057/palgrave.ivs.9500006.

Appendix A

Data from Chapter 2 analysis

This appendix contains tables summarizing the analyses in Chapter 2, comparing English-commented OSM changesets with measures of economic development, income and basic household needs.

Table A-1. 2014 GDP (PPP) per capita and percent of OSM changesets commented in English at the country level

Country	**Gross domestic product (GDP) for 2014 at purchasing power parity (PPP) per capita, in international dollars**	**Number of OSM changesets in study dataset in this country commented in English**	**Percent of OSM changesets in study dataset in this country commented in English**
Argentina	18749	2325	14.96
Bolivia	5364	1403	59.98
Brazil	12221	12195	22.09
Chile	19067	1713	26.83
Colombia	11189	3161	38.75
Ecuador	10080	1139	38.11
French Guiana	Excluded	Excluded	Excluded
Guyana	Excluded	Excluded	Excluded
Paraguay	6823	132	36.57
Peru	11124	990	25.47
Suriname	Excluded	Excluded	Excluded
Uruguay	16723	413	12.83
Venezuela	13604	571	19.38

Table A-2. 2014 monthly income in Brazilian states and percent of OSM changesets commented in English.

Brazilian state	**Monthly income for 2014 in R$ (Brazilian real)**	**Number of OSM changesets in study dataset in this state commented in English**	**Percent of OSM changesets in study dataset in this state commented in English**
Acre	670	33	73.33
Alagoas	604	126	6.98
Amapá	753	84	61.76
Amazonas	739	125	26.65
Bahia	697	759	23.94
Ceará	616	283	40.9
Distrito Federal	2055	205	20.24
Espírito Santo	1052	444	60.33
Goiás	1031	270	19.85
Maranhão	461	185	55.39
Mato Grosso	1032	372	50.07
Mato Grosso do Sul	1053	357	44.96
Minas Gerais	1049	1989	23.39
Pará	631	209	20.15
Paraíba	682	121	20.54
Paraná	1210	513	19.61
Pernambuco	802	231	45.56
Piauí	659	202	69.66
Rio de Janeiro	1193	1242	20.5
Rio Grande do Norte	695	70	10.04
Rio Grande do Sul	1318	1488	25.39

Rondônia	762	35	26.72
Roraima	871	5	17.86
Santa Catarina	1245	364	15.77
São Paulo	1432	1886	15.77
Sergipe	758	86	27.74
Tocantins	765	104	24.47

Table A-3. Percent of population lacking at least one basic human need in Argentine provinces and percent of OSM changesets commented in English.

Argentine province	**Percent of population lacking at least one basic human need in 2010**	**Number of OSM changesets in study dataset in this province commented in English**	**Percent of OSM changesets in study dataset in this province commented in English**
Buenos Aires	11.2	632	10.22
Catamarca	14.6	23	45.1
Chaco	23.1	83	14.14
Chubut	10.7	22	12.22
Ciudad Autónoma De Buenos Aires	7	279	18.37
Córdoba	8.7	255	21.18
Corrientes	19.7	14	6.42
Entre Rios	11.6	60	18.02
Formosa	25.2	9	30
Jujuy	18.1	19	18.81
La Pampa	5.7	9	9.89
La Rioja	15.5	26	45.61
Mendoza	10.3	192	34.41

Misiones	19.1	39	3.28
Neuquén	12.4	27	6.78
Rio Negro	11.7	96	23.76
Salta	23.7	80	45.71
San Juan	14	51	43.22
San Luis	10.7	14	10
Santa Cruz	9.7	84	56.76
Santa Fe	9.5	153	10.7
Santiago Del Estero	22.7	40	30.53
Tierra Del Fuego, Antártida E Islas Del Atlantico Sur	14.5	26	50
Tucumán	16.4	86	38.05

Appendix B

Online user evaluation of Crowd Lens, focused toward geospatial technology professionals

Below is the full text of the online user evaluation of Crowd Lens, administered to geospatial technology professionals as described in Chapter 4.

Thank you for volunteering to evaluate this tool for visualizing contributions to OpenStreetMap. This tool is called Crowd Lens, and it was developed as part of an academic research project at Penn State University. You will be asked to explore the tool and assess its effectiveness through a series of questions. Please use a laptop or desktop computer (not a smartphone or tablet) to perform this evaluation. You must have Mozilla Firefox or Google Chrome installed in order to evaluate Crowd Lens.

Before using the tool, please ***answer a few background questions*** *about your work with OpenStreetMap.*

How would you describe the work you do with digital mapping and GIS?

How do you currently use (or how have you considered using) OpenStreetMap in your work?

On a scale of 1 to 5, how confident are you in the quality of OpenStreetMap data for your work purposes?

- 5 ("I am HIGHLY CONFIDENT in the quality of OpenStreetMap data for my work purposes")
- 4
- 3
- 2
- 1 ("I have MAJOR CONCERNS with the quality of OpenStreetMap data for my work purposes")

Why did you choose your answer above?

*Now please **take a few minutes to read the following** introduction page that describes the capabilities of Crowd Lens. When you are finished, close the introduction page and return to this step.*

<Link to Crowd Lens introduction page> (See Appendix C for content)

*Now **take a minimum of 10 minutes to explore the Crowd Lens tool** using the link below. Please use Google Chrome or Mozilla Firefox on a laptop or desktop computer.*

As you use the tool you are welcome to explore any feature of interest to you. As you do so, try to form an understanding of how OpenStreetMap has taken shape over time in these towns through the efforts of the individual contributors. Also, please consider how the development of OpenStreetMap in these towns affects your own perception of the quality of OpenStreetMap data that you have used (or considered using) in your projects.

<Link to Crowd Lens>

To finish with the evaluation, please ***answer the following questions***.

What insights (if any) did this tool provide about OpenStreetMap data that you had not known or considered in the past?

How do you think the insights above will affect (if at all) the way you work with OpenStreetMap?

How do you feel this tool could be improved to provide additional insights that would be valuable to you?

On a scale of 1 to 5, to what degree has this tool affected your perception of the quality of OpenStreetMap data for your work purposes?

- 5 ("I feel MORE CONFIDENT in the quality of OpenStreetMap data for my work purposes after using this tool.")
- 4
- 3
- 2
- 1 ("I feel LESS CONFIDENT with the quality of OpenStreetMap data for my work purposes after using this tool.")

Why did you choose your answer above?

In which continents have you lived? (Check all that apply)

- North America
- South America
- Europe
- Africa
- Asia
- Australia or Pacific Islands
- I prefer not to answer

Which best describes your age group?

- 18–29
- 30–39
- 40–49
- 50–59
- 60+
- I prefer not to answer

What is your gender?

- I have entered my gender below
- I prefer not to answer

Appendix C

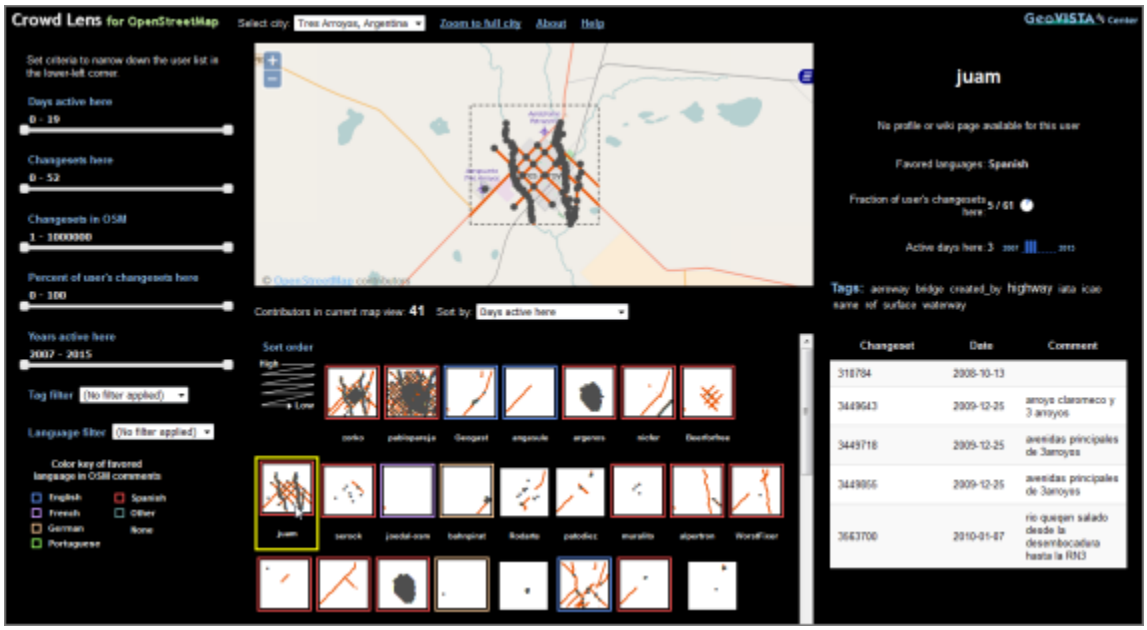
Introductory documentation presented to Crowd Lens testers in Chapter 4[24]

What is Crowd Lens?

OpenStreetMap is a "crowdsourced" geographic database built by volunteers on the Internet. Crowd Lens is a tool for learning who built OpenStreetMap in a particular place. It lets you explore subsets of the crowd, or view detailed information about any one particular contributor. Crowd Lens can help you answer questions such as:

- How many people influenced the map in this city?
- How many people recently modified this information?
- Was most of the work here done by one person, or was it spread out evenly among contributors?
- Has systematic vandalism, import, or bot activity occurred in OpenStreetMap here?
- What can we learn about individual contributors if we use all publicly available metadata?

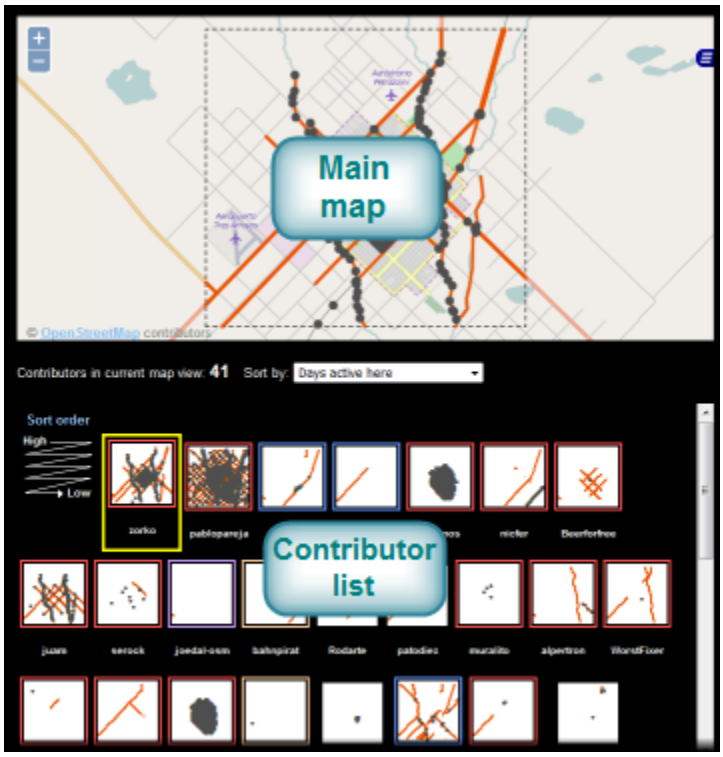
[24] To preserve the flow and integrity of the original documentation page content, I have not introduced figure numbers or captions in this appendix. Note that some of the images in this appendix also appear in the main body of Chapter 4 (with captions).



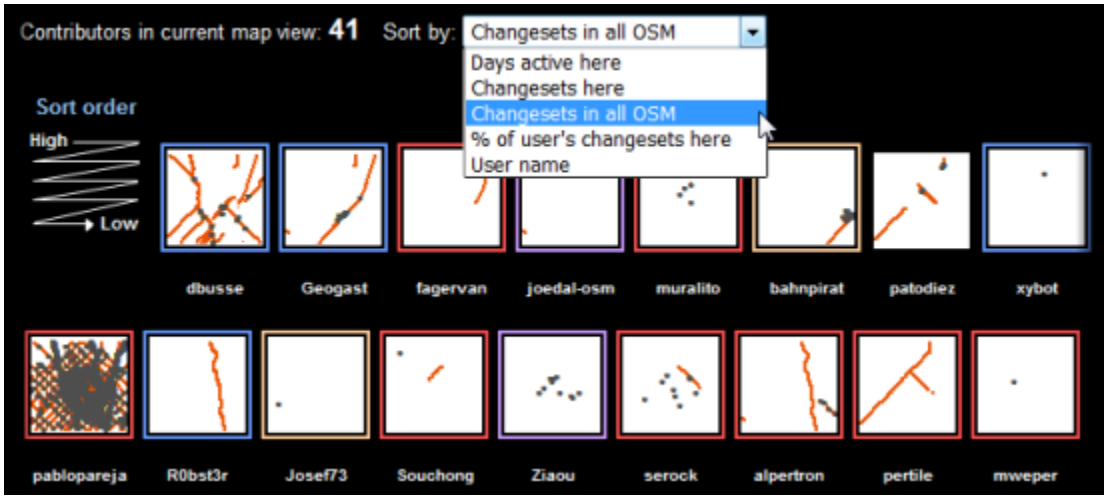
Getting started

To get started with Crowd Lens, select a city from the top dropdown menu.

You'll notice a big map appear and a bunch of small maps below it. These small maps constitute a contributor list, representing the work of each OpenStreetMap contributor that edited this city. You can use this list of maps to distinguish between systematic and casual contributors.

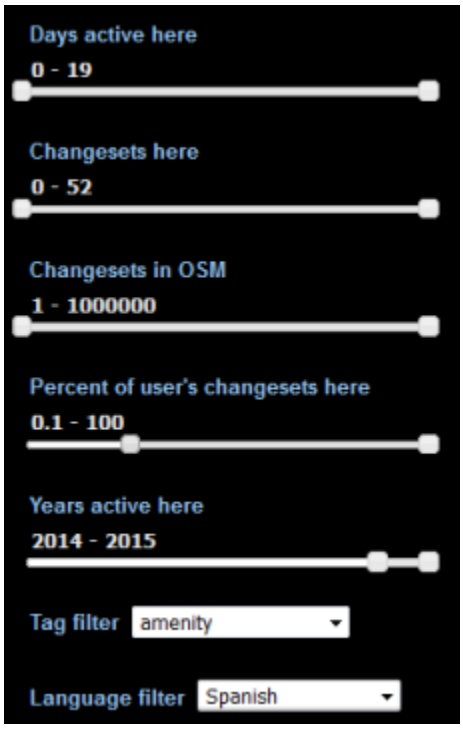


To discover the most active contributors, you can sort this contributor list by certain characteristics such as number of days active in the project, or number of changesets in the entire OpenStreetMap project. The sorting occurs from top to bottom, left to right.

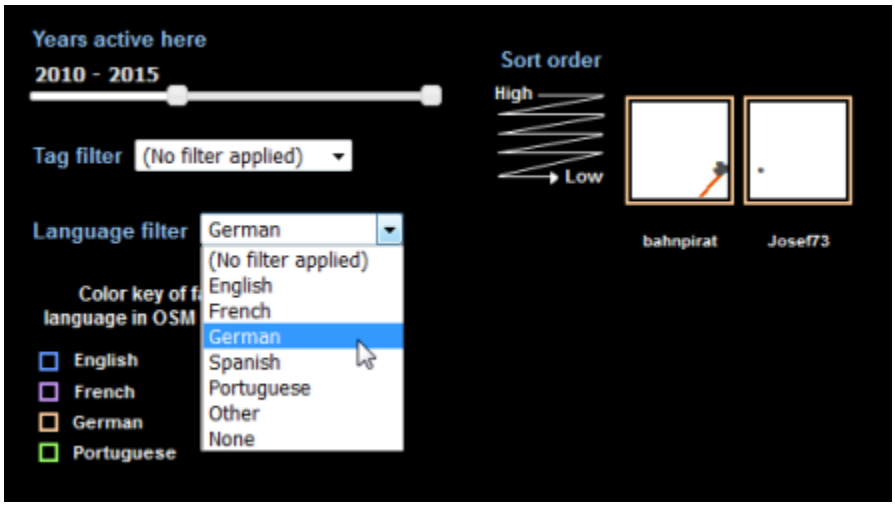


Learning about the crowd

Use the filters on the left side of the screen to learn about the characteristics of the crowd that built OpenStreetMap in this city. The filters consist of range sliders and dropdown lists that narrow down the contributor list to just those contributors who meet certain criteria, such as contributors who used a certain tag or who edited the map in a particular year.



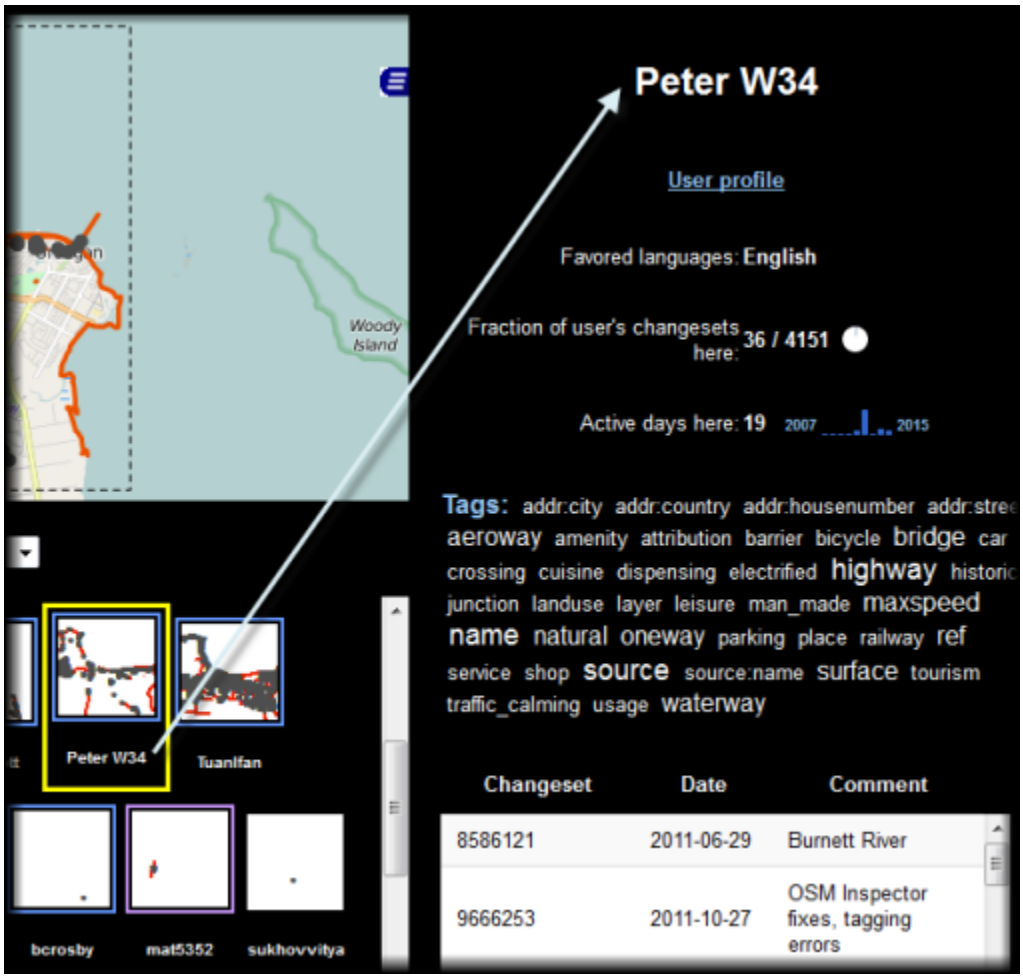
One of the filters allows you to select contributors who favored a particular language in their OpenStreetMap changeset comments. Note that this language is the one most commonly used by the contributor in the OSM project as a whole, not the city alone.



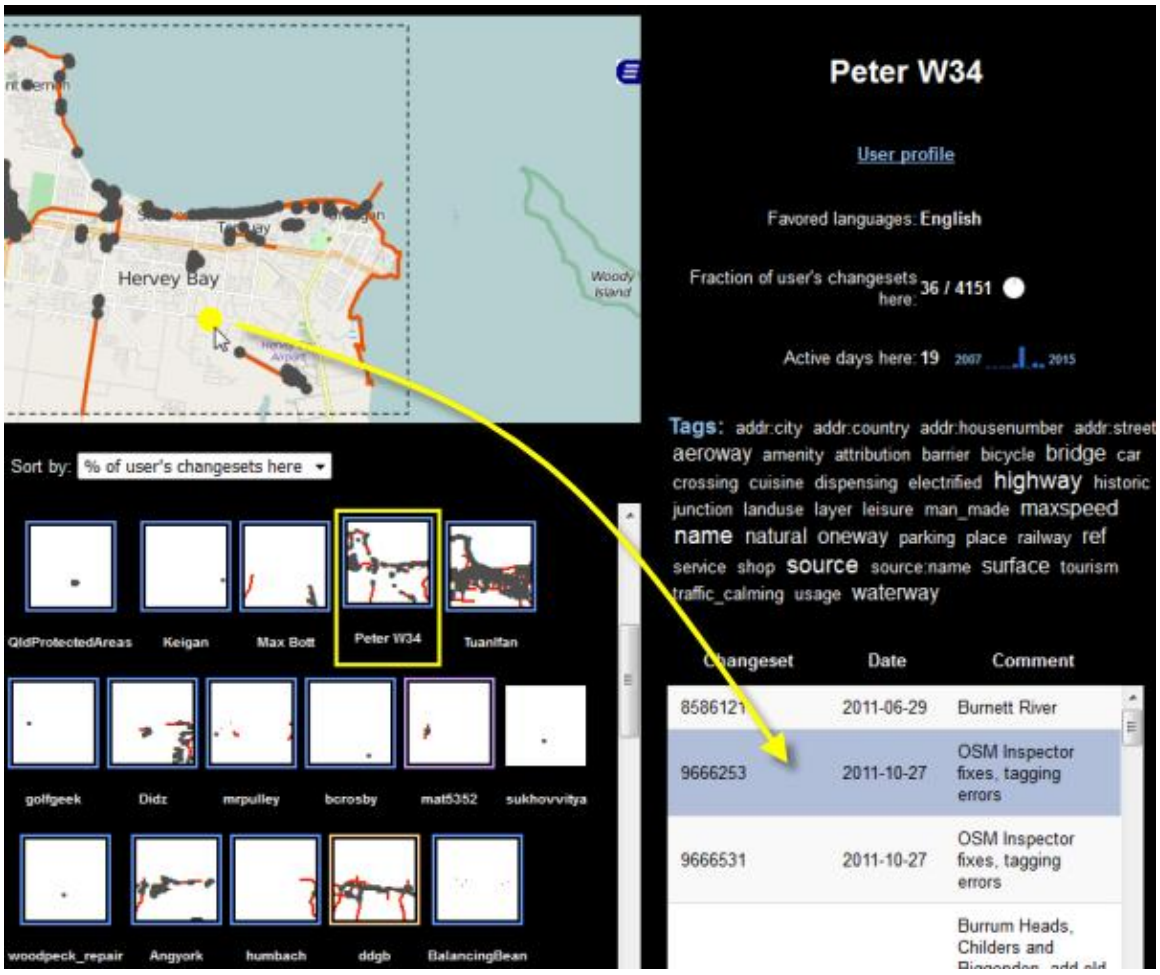
The main map in the center of the screen is also a filter; panning and zooming narrows down the contributor list to show you only the contributors who influenced the current map view. This is a great way to see who built OpenStreetMap in a particular neighborhood.

Learning about individual contributors

Click any contributor's little map to see further details in the right hand panel describing that contributor's work in OpenStreetMap. This includes the tags used by the contributor and the dates and comments associated with each changeset submitted by the contributor.



The map at the top of the screen displays the OpenStreetMap nodes and ways submitted by the contributor. Click any node or way to see its associated changeset highlighted in the table. Click any row in the changeset table to see the changeset highlighted on the map.



More information

For additional details about the Crowd Lens implementation, including lists of supported environments and known issues, please see the About page.

VITA
Sterling Quinn

Sterling Quinn earned a Bachelor of Science degree in Geographic Information Systems from Brigham Young University in 2005. He then went to work as a product engineer at the mapping software company Esri for eight years, where he documented, designed, and tested web mapping technologies. During this time he earned a Master of Geographic Information Systems degree from Penn State University. In 2013 Sterling left Esri and moved to State College to work on his doctoral degree full time.

Sterling has taught courses on introductory mapping, GIS programming and automation, and cloud and server GIS. In 2015 he was awarded the GeoForAll Global Educator of the Year award for his work authoring the Penn State "Open Web Mapping" online course.

Sterling has advised numerous undergraduate students during his time at Penn State and served as co-coordinator of the Undergraduate Research Opportunities Connections (UROC) program. During the 2014–2015 academic year he served the Penn State Department of Geography as an elected graduate student representative to the faculty, and in 2015–2016 he served on the graduate program and curriculum committee.

Sterling will begin work as an assistant professor of geography at Central Washington University in September 2016.